日本銀行金融研究所「金融研究」第10巻第4号（平成3年12月）

統計データの個票公開とプライバシーの保護

——推論制御の理論、その紹介と応用

岩村　充
西島裕子

1. はじめに

現在、わが国では非常に多くの「統計」が作成・発表されている。その数は、国勢調査や消費者物価のように政府機関が作成している指定統計と、卸売物価指数のように政府以外の組織や機関が作成している届出統計とを合わせ、約5,500種類にも及ぶ。

ところで、これらの統計調査の結果として公表されるのは、多くの場合、調査対象となった個体（標本）の持つ属性の平均値であり、統計調査によって得られた「生」のデータではない。例えば、総務庁が全国の家計から8,000世帯を無作為抽出し、その所得や消費の内容を調べた結果である「家計調査報告」において公表されるのは、その各収支項目の平均値であって、個別の標本（各々の家計）の収支金額ではない。このように統計調査の結果得られる個別データをそのまま発表せず、その平均値のかたちに加工して公表するのは、その方が全体としての傾向を把握し易いからであるが、同時に調査対象となった家計のプライバシーを保護するためでもある。

しかし、統計の利用者の立場からは、このような公表の仕方で本当に十分なのかどうか、議論があり得るであろう。先の家計調査報告の例でいえば、利用者の目的が、単に「平均的」な家計の所得・消費動向を把握したいというものであったとすれば、公表データは標本全体の平均値で十分だが、仮に所得や各消費項目間の相関や因果関係を追究しようというものであったとすれば、標本全体の平均の動きを知らされただけではほとんど役に立たないからである。

もちろん、項目間の関連性を分析したいという利用者の要求に対しても、その要求が統計の作成者にとって「典型的」な関心に基づくものであると認識された場合には、そうし

本論文の作成に当たっては、加納　悟（横浜国立大学）、新開陽一（大阪大学）、橘木俊詔（京都大学）、中島昌子（日本女子大学）、本多佑三（神戸大学）の各氏から有益なコメントを頂いた。

た関心を充足するために、特に加工された内訳データが提供されることもある。家計調査報告において、標本世帯を所得階層に分けて、各階層の平均的な所得や消費の内容が公表されているのは、この例である。

しかし、このようなかたちで統計作成者が提供する内訳データは、あくまでも統計作成者が「典型的」と認めた関心に応えるためのものであるにとどまり、統計利用者の自由で多様な関心を満足させてくれるものではない。家計調査報告のケースでいえば、年齢と所得・消費行動との関係をクロスセクションでみたいとか、特定の資産（例えば土地や家）を購入する前後の消費行動の変化を時系列で追跡したいといった個別データに対する関心は満足させられない。[1)]

では、統計利用者の個別データに対する関心を満足させることは、本当に無理なのであろうか。

この関心に応える最も単純な方法は、個別の標本に関するデータ（いわゆる「個票データ」）から、氏名や住所等の特定の個人を指す情報を消去したうえで、残りのデータを磁気テープ等の媒体に記録して提供してしまうことである。しかし、この方法は、統計利用者の多様な関心と調査対象のプライバシー保護とを両立させる名案のようでありながら、プライバシー保護という観点からは、不十分なアプローチである。

この方法が不十分である理由は、名前や住所等が消去されたデータを通じてでも、特定の個人のプライバシーに侵入することが、通常あまり難しくないからである。家計調査報告の例でいえば、プライバシー侵入を企てる者が対象に関して若干の予備知識を持っていれば（例えば、最近自動車を買ったか、どの程度の家賃の家に住んでいるか、クレジットカードの会員になっているか、という程度の知識を持っていれば）、名前と住所が消去されたファイルからでも「狙った人」のデータを発見できる可能性はかなり高いし、もう少し詳しく個人情報を知っている立場の者であれば発見の可能性はさらに高くなる。プライバシーへの侵入を企てる者が狙った相手に対して全く予備知識を持っていないということがむしろ希であるとすれば、個票公開の可能性を考えるに当たって、名前や住所の消去が、プライバシー保護のために必要な条件ではあっても、十分な条件ではないことは明らかである。

もっとも、この方法は、プライバシー保護の観点から不十分とされるだけであって、誤った方法であるとはいえない。したがって、条件の良い一部の統計調査については、公表の仕方を工夫することにより、統計利用者の関心とプライバシー保護とを両立させる余地がある。例えば、米国のセンサス局は、国勢調査データについて、その全データのうちから100分の1の世帯を無作為抽出し、名前や住所等の情報を消去したうえで磁気テープに収めて公表している。[2)] この場合、公表された統計データを使って「狙った人」のプライ

1） 統計の利用者の目的が学術研究目的であることが明確な場合は、一定の条件の下で個票の閲覧ができることもある。しかしこのような閲覧は、プライバシー保護との関係で通常極めて限定的にしか許可されていない。

2） センサス局のデータ公開の方式については、U.S. Department of Commerce [1978] を参照。

バシーに侵入しようとしても、成功する確率は100分の1しかないし、たまたま「狙った人」のものらしいデータを公表統計の中に発見したとしても、それが本当に「狙った人」のデータなのかどうか、侵入者としては確信が持てないであろう。

センサス局の公表方式は、このようなかたちで統計利用者の関心と個票対象者のプライバシー保護とを両立させてくれるが、その欠点は、この方式を適用できるのが、国勢調査のように非常に多数の標本を取得する統計に限られていることである。実際、家計調査報告のように最初に取得する標本数が8,000程度の統計だと、そこから100分の1の公表用標本を抽出すれば、その数は80世帯分にしかならず、とても統計データとしての実用性を満足させられない。また、逆に十分な数の公表用標本を確保しようとすれば、最初に取得すべき標本数が膨大なものとなってしまい、統計調査の実施コストが耐えられない程大きくなってしまうであろう。

ところで、近年のデータベース技術の発達は、この統計利用者の多様な関心に対する充足と、調査対象のプライバシー保護との両立とを、従来とは異なる角度から解決してくれる可能性を提供してくれる。これは、統計調査の結果得られる個別データをそのままのかたちで公表するのではなく、個別データはコンピュータが管理するファイルにデータベースとして保管しておき、個別データに対するアクセスの必要が生じた利用者はデータベースにオンラインで「検索（query）」を行い、個人のプライバシーに侵入する惧れのないようなかたちに加工された統計量を取得する、というようなシステムが想定可能になったからである。[3)]

本論文は、このような技術進歩を踏まえて、現在よりも多少とも進んだ統計データの公表の方法がないかを探ろうとするものである。これをいい換えれば、統計データを公表するに際し、統計作成の趣旨に反する情報が知られてしまうのを防ぎながら、統計作成の趣旨に合致する限りは、なるべく多くの情報を公表データから得られるよう、利用者の推論（inference）を制御する手段を検討することでもある。このような検討を行うことは、「推論制御（inference control）」と呼ばれ、一般にデータ保護（data security）に関する理論の1分野であるとされるが、1980年代前半の米国において幾つかの重要な貢献が行われて以降、1980年代の後半に入っては、新しい貢献は少なくなっているのが実情である。しかし、最近におけるデータベース・システムやネットワーク・システムの急速な普及を考えると、この推論制御の理論の応用面における重要性は、むしろ高まっているといって過言

3） このようなデータベースの概念は、一般には、縦軸に個人や家計等を、横軸にその個人の性別、収入、家族構成等のデータをとって表形式に展開されたデータセットと、それに対し検索すべきデータの範囲（男性であること、収入1,000万円以上であること、男性でかつ収入1,000万円以上であること、等）を指定しての自由な検索を実現するアクセス方式との組合せとして与えられる。これが、いわゆるリレーショナル・データベース（relational data base）である。リレーショナル・データベースのアイディアは、1970年代に発表されたものであるが、むしろ1980年代に入って目覚ましい発展を遂げ、現在ではパーソナル・コンピュータのデータベースの大半を占めるに至ったほか、大型コンピュータの分野にも利用が拡大しつつある。

でないと思われる。また、少なくとも、こうした理論の存在すら一部の研究者を除いては知られていなかったわが国の状況は、変わってもよいのではないだろうか。[4)]

本論文の主たる目的は、統計利用者の多様な関心に対する充足と、統計調査対象のプライバシー保護との両立という観点から、これまでわが国であまり知られていなかった推論制御の理論について、そのエッセンスを紹介し、この問題について読者の関心を喚起するところにある。したがって、理論の数学的展開のトレースに時間を割くゆとりのない読者においては、数式部分、特に2.の数式部分をスキップしても差し支えない。ここでの趣旨は、統計データの有効利用と個人のプライバシー保護との両立を図ることは必ずしも簡単なことではないが、検索に対して応答されるデータに一定の手順で算出されたランダムな値を「ノイズ」として加えることにより、この両立を実現できる可能性があることを示すところにある。しかし、推論制御の理論自体よりもその応用面に関心を持つであろう大多数の読者にとっては、ここでの数式的展開のトレースは必ずしも不可欠の作業ではない。そのような読者にとっては、3.を通じて、この理論の意味について実感的なイメージを得る方がむしろ有益だからである。今後の展望については、4.で簡単に考察する。

2. 推論制御の理論

(1) 線形不法侵入

統計データ（個票データ）の公表に当たって、プライバシーの保護とデータとしての有用性を両立させるための最も直感的なアイディアは、対象の数が極端に少なくなるような検索を拒否することであろう。個票データを公表する趣旨は、統計調査の対象となった個体（個人や家計等）の持つ様々な属性（性別、年齢、収入等）の間の統計的な関連性であって、個々のデータではないのだから、含まれる個体数が極端に少ない集団を対象とした検索には回答しなくても差し支えないはずであるし、一方、含まれる個体数が十分に多い集団から得られる統計量（平均値あるいは合計値）であれば、検索者による任意の「括り」に応じて回答しても、プライバシーへの侵入はないと考えるのである。家計調査報告の例でいえば、非常に高い収入を指定して「年収○○円以上の家計の全収支項目の数量および金額の平均値をリストせよ」というような検索が入った場合は、対象世帯数が少なすぎるので、プライバシーを開示してしまう惧れがあるとして回答を拒否するが、検索者が対象世帯数が多くなるよう「年収○○円以上」の基準を引き下げてくれば、その対象世帯数が一定値を超えたところでデータを回答することにする、というようなデータベース・システムを想定するわけである。

しかし、このアイディアはプライバシー保護策としては不十分である。その理由は、個票データを収容したデータベースに対して「繰り返し検索」が可能であれば、1回毎の検索に対しては各個票データを十分にマスクするだけの標本数を合算して、その合計値あ

4） 推論制御の理論を紹介した日本語文献は、筆者の知る限りでは、データ保護に関するやや専門的な教科書である Denning[1982] の邦訳（上園忠広・小嶋 格・奥島秋子訳、「暗号とセキュリティ」、培風館、1988年の第6章）だけである。

るいは平均値を開示する仕組みであったとしても、繰り返しによって得た検索結果を組合せることによって、個票データの開示(disclosure)を受けたのと同じ効果を得ることができるからである。このことを簡単な例を用いて説明しよう。

第1表は、ある病院に来診したN人の個体（患者）のM種類の属性（身長、体重、性別、血液型、病名、等々）をリストしたものであるとしよう。ここで、便宜上、各個体は第1属性（身長）の降順（大きい順）に並べられているとする。このデータが何らかの医学的貢献を狙って、検索可能なデータベース・システムとして公開されたとしよう。

第1表　病院を訪れた患者とその属性のリスト

患者＼属性	1	2	・	・	・	・	M
1	x_{11}	x_{12}	・	・	・	・	x_{1M}
2	x_{21}						・
・	・						・
・	・						・
・	・						・
・	・						・
N	x_{N1}	・	・	・	・	・	x_{NM}

先に説明した考え方に従えば、このデータベースに対する検索は、システムの管理者が設定した限界値pと等しいか、それよりも多い個体を含む集団の統計量に対してしか許可されないから、[5]検索者がデータの第1属性（例えば身長）に注目して、

$$x_{i1} \geqq q$$

を満たす集団について、そこに含まれる個体の数とその各属性値の平均値を問い合わせた場合、この検索要求は、

$$x_{p1} \geqq q$$

が成立するときのみ、その集団に属する個体数nと各属性のその集団における平均値を並べたベクトル：

$$U_n = \left(\frac{1}{n}\sum_{i=1,n} x_{i1}, \frac{1}{n}\sum_{i=1,n} x_{i2}, \cdots, \frac{1}{n}\sum_{i=1,n} x_{iM}\right)$$

を回答として得る。検索がここで終われば、検索対象となった集団に含まれる各個体のデータは、集団内の他の個体データと混ぜ合わされることにより十分なノイズが加えられるので、各個体のプライバシーは保護される。

しかし、検索者が検索の閾値qの水準を変えて、例えばqを僅かずつ引き下げて、再度の検索を試みたらどうなるであろうか。ここで、qの引下げ幅があまりにも小さければ、得られる答は同じであるから意味がないが、検索者がちょうど検索対象となる個体数が$n+1$となるようにqを変化させて、応答：

$$U_{n+1} = \left(\frac{1}{n+1}\sum_{i=1,n+1} x_{i1}, \frac{1}{n+1}\sum_{i=1,n+1} x_{i2}, \cdots, \frac{1}{n+1}\sum_{i=1,n+1} x_{iM}\right)$$

5）　もし、このデータについて、全個体（全患者）についての合計値または平均値が公表されているとすれば、検索制限は「あまりにも多数の集団」、具体的には、「N－検索対象個体数」がp以下となるような「大きな集団」に対しても、制限される必要がある。その理由は、このような検索を許すと、全患者に関するデータから、検索対象に関するデータを差し引くことによって、含まれる個体数がp以下の集団についての検索を許したのと同じ効果が得られてしまうからである。

を得たとすれば、検索者は前の検索結果と新しく得られた検索結果を組合せて、

$$(n+1)\cdot U_{n+1}-n\cdot U_n$$

を解くことにより、第1属性の大きさが$n+1$番目の個体について、その全ての属性を知ることができる。すなわち、検索者は、複数の検索を組合せることによって、1回限りの検索では不可能だった個別データへのアクセスを実質的に実行することができるのである。

この方法論を拡張することにより、データベースとして管理されている個体別データについて、検索対象となる集団（以下、これを「検索集合」と呼ぶ）を変化させて、その集団の属性の合計値（あるいは平均値）を照会することができさえすれば、検索集合のサイズ（検索集合に含まれる個体数）に関して如何なる制限を設けようとも、最終的には全ての個別データを開示してしまうことができる、という命題を導くことができる。次にこのことを示そう。

検索の対象となるデータベースは、Nの個体に関するM種類の属性をリストしたものとして、N行×M列の行列$P=(x_{ij})$として表現するものとしよう。すなわち、前記の第1表である。

ここで、このデータベースのある一部を検索集合R_1として指定し、各属性データについて、R_1に含まれる個体についての合計値を照会したとすれば、[6] それはi番目の個体がR_1に含まれるとき値1、そうでないとき値0をとるパラメータ$\alpha_{1i}(i=1\cdots N)$を定義し、このα_{1i}を横に並べたN次の行ベクトル：

$$\alpha_1=(\alpha_{11},\alpha_{12},\cdots,\alpha_{1N})$$

を、行列Pに乗じて、M次の行ベクトル：

$$\begin{aligned}\beta_1&=\alpha_1 P\\&=\Big(\sum_{i=1,N}\alpha_{1i}\cdot x_{i1},\sum_{i=1,N}\alpha_{1i}\cdot x_{i2},\cdots,\\&\qquad\sum_{i=1,N}\alpha_{1i}\cdot x_{iM}\Big)\end{aligned}$$

を得ることにほかならない。

次に、検索者がR_1とは異なる検索集合R_2を指定して、同様の手順により、

$$\beta_2=\alpha_2 P$$

を得るという操作を繰り返すとすれば（繰り返しの回数をNとしよう）、このとき指定したN個の検索集合$R_k(k=1\cdots N)$につき、N次のパラメータベクトル$\alpha_k(k=1\cdots N)$と、M次の検索結果ベクトル$\beta_k(k=1\cdots N)$が各々N個ずつ得られることになる。そこで、パラメータベクトルα_kを縦に並べた$N\times N$行列をAとし、また検索結果ベクトルβ_kを縦に並べた$N\times M$行列をBとすれば、AとBは、原データをリストした$N\times M$行列Pとの関係で、

$$AP=B$$

6） 通常、このような検索は、「合計値」ではなく「平均値」に対して行われることが多いが、合計値と平均値は「丸め(rounding)」の問題を別にすれば、同等の情報を有する統計量なので、ここでは合計値に対する検索が行われた場合を例にとって説明する。なお、「丸め」の意味については、後で考察する。

ただし、$A=\begin{pmatrix}\alpha_1\\ \alpha_2\\ \cdot\\ \cdot\\ \cdot\\ \alpha_N\end{pmatrix}$, $B=\begin{pmatrix}\beta_1\\ \beta_2\\ \cdot\\ \cdot\\ \cdot\\ \beta_N\end{pmatrix}$

と表記できる。ここで、A は検索者にとって既知であるから、[7]その逆行列 A^{-1} も検索者にとって既知となるので、検索者は、結局、

$$P=A^{-1}B$$

の演算を行うことにより、全ての個体の全ての属性データを知ることができてしまうことになる。[8]

一般に、統計調査の結果得られたデータの部分集合（検索集合）を取り出して、その線形結合として得られる合計値や平均値等の統計量を応答するような仕組みのデータベースを利用して、データ提供者の趣旨に反するようなかたちで個別データに侵入してしまうことを、「線形不法侵入(linear system attack)」と呼ぶが、以上の説明は、データ提供者が検索集合のサイズに関し、どのような制限を行ったとしても、それだけでは線形不法侵入を防ぐことが、本質的に不可能であることを示している。

もちろん、線形不法侵入が可能であるためには、データベースに対する繰り返し検索が許されていることが条件となるから、例えば繰り返し検索の回数について一定の制限を設ければ、その危険を減少させることはできるが、このようなデータベースについて、そこに含まれる個別データへの侵入を可能とするのに必要な検索回数は、驚くほど小さいことが理論的に明らかにされている。[9]したがって、本論文で取り上げるようなデータ公開の趣旨に応えるためには、繰り返し検索を完全に禁止するか、それができなければ、繰り返し検索によって個別データへの侵入を試みても、必要な繰り返しの回数が非常に大きくなるようにデータを制御して、繰り返し検索による個別データへのアクセスが実質的に不可能になるような仕組みを工夫する必要がある。本論文は、そのようなデータ制御の仕組みを、回答データに一定の「管理された撹乱」を与えることにより得ようとするアイディアについて検討するものである。

7） もちろん、検索集合のサイズが既知でなければ、この種の推論は困難になるが、これを開示しないことは、検索応答としての統計量の信頼性に対する評価の手段を奪うことになるから、データ公開の趣旨として望ましくないであろう。また、検索集合のサイズを開示することに代えて、検索応答に対し検索集合の分散や標準偏差等の情報を開示することも考えられるが、その場合であれば、そうした情報を基にして検索集合のサイズを推定できてしまう可能性が生じることになる。

8） 逆行列の存在等に関する検討はここでは行わない。

9） 先の病院の患者データについての例示では、この回数は2回である。なお、一般に、少ない検索回数で「能率的」にプライバシーへの侵害を果たす方法を考える問題は「追跡者(tracker)」の問題と呼ばれ、推論制御の分野では比較的多くの研究成果が発表されているテーマである。詳しくは、Denning[1982]およびそこで引用されている多数の文献を参照。

回答されるデータに撹乱を与えるというと、統計の信頼性を重視する立場からは異論があり得るであろう。しかし、われわれが通常手にするデータに、あまり意識されない「小さな撹乱」が加えられていることは、案外多いものである。例えば、四捨五入による「丸め」がそれである。もちろん「丸め」は、いわゆる「有効数字」の考え方に基づいて行われるものであるから、これを単に「撹乱」というのは適当でないが、推論制御の理論の観点からみるとすれば、「丸め」は一種の「管理された撹乱」として個別データへの侵入を難しくするという副次的な効果を示す。そこで、次に、この「丸め」の持つデータ撹乱効果をみてみることによって、われわれのテーマについてイメージの具体化を図る努力をしてみよう。

(2) 数値の「丸め」とその効果

ここでは、検索に対して返される応答データに、「丸め」が行われたときの影響を考察しよう。「丸め」の形式としては、「四捨五入」「切り上げ」「切り捨て」のどれでもよいが、差し当たって、先に説明した患者データのケースを例にとって、データベースの任意の部分集合（ただしサイズn以上とする）を検索集合とする平均値検索への応答に、「小数点以下四捨五入」という操作が行われたときの効果を定式化してみよう。

不法侵入はデータベース中の特定の個体I（以下、侵入者が侵入を企てる相手の個体を「標的」と呼ぼう）のデータに対して、2回の平均値検索によって行われるとしよう。この不法侵入における「丸め」の役割は、次のように表現することができる。

侵入者は標的Iを含むサイズ$n+1$の集合（$n+1$人の患者のデータの集合）を指定し（以下、この集合をR_0と呼ぼう）、まずこのR_0を検索集合として平均値検索を行う。ここで検索結果のベクトルをU^0とすると、その第j成分（検索集合R_0に属する患者の第j属性の平均値）は、

$$U^0{}_j=\frac{1}{n+1}\sum_{i\in R_0}x_{ij}+e^0{}_j$$

として定式化できる。ここで$e^0{}_j$は、四捨五入によって生じる真の値からのズレ（撹乱）であり、ほかの何の情報もなければ、$(-0.5, 0.5)$の区間に一様分布する確率変数である。

次に侵入者は、R_0からIを除いた集合をR_1として、このR_1に対し、R_0に対する検索と同様の検索を行う。その結果は、

$$U^1{}_j=\frac{1}{n}\sum_{i\in R_1}x_{ij}+e^1{}_j$$

となる。侵入者は、このようにして得た2つの値$U^0{}_j$と$U^1{}_j$を用いて、患者Iの第j属性についての推論値を、

$$\begin{aligned}w_{Ij}&=(n+1)U^0{}_j-nU^1{}_j\\&=x_{Ij}+(n+1)e^0{}_j-ne^1{}_j\end{aligned}$$

として得ることができる。いうまでもなくw_{Ij}の期待値はx_{Ij}であるから、w_{Ij}は真の値x_{Ij}の不偏推定量であるが、興味深いのは、その撹乱項：

$$e_j=(n+1)e^0{}_j-ne^1{}_j$$

の拡がりであろう。この拡がりが十分大きければ、侵入者が得られる推論値w_{Ij}はそれが不偏推定量を与えるものであっても、推論が曖昧過ぎて実質的に意味があるとはいえないはずだからである。そこで、$e^0{}_j$と$e^1{}_j$とが完全に独立の場合を例にとって、w_{Ij}の撹乱項e_jの「大きさ」が検索集合のサイズnと

の関係でどのように変化するかを第2表で示しておこう。[10)]

第2表 $|e_j|$ の値が一定値εを超える確率

n \ ε	1	10	100	1,000
1	0.1250	0	0	0
10	0.8205	0.0432	0	0
100	0.9802	0.8916	0.0049	0
1,000	0.9980	0.9890	0.8992	0.0005

第2表によって明らかなとおり、「丸め」による撹乱は、検索集合のサイズ n が大きくなると、相当に大きくなる。もし、侵入者が得ようとしているデータが3～4桁程度（1,000程度）のオーダーのものであるとすると、侵入者が手にする「推論値」には、n が100のときであれば、確率90%で当該データの約1%の誤差が含まれていることになるし、[11)] n が1,000のときであれば、確率90%で当該データの約10%もの誤差を生じてしまうことになる。このような撹乱の大きさが、侵入側あるいは防御側にとってどれ程の意味があるかは、データの性格や侵入の目的にもよるが、いずれにしても n を一定値以上に大きくとらざるを得ないようなルールが作られていれば、不法侵入への防止策として「有効」であるかのようにみえる。

しかし、このような「丸め」による撹乱効果に頼って個別データを保護しようとするアイディアには、いくつかの問題がある。

「丸め」によるデータ撹乱の問題の1つは、データの「丸め」により個別情報がマスクされる程度が、当該データの「大きさ」に依存してしまうことである。具体的にいえば、「丸め」の対象となるデータは大きければ大きいほど、「丸め」によって生じる撹乱の影響は相対的に小さくなるので、そのデータはより少なくしか保護されていない。しかし、データベースの管理者としては、検索の対象となる全てのデータについて「何らかの基準」に従って、できる限り一定の強さで個別情報をマスクしたいはずである。そうした観点からは、単なる「丸め」に代えて、もっと管理し易い撹乱の与え方を考える方がよいのではないだろうか。

「丸め」によるデータ撹乱について注意すべきもう1つの問題は、繰り返し検索への対応である。よく知られているように、同一の分布（分散を σ^2 としよう）を持つ多数（L 個）

---

10) ただし、このような見方をするについては、撹乱項 e^0_j と e^1_j との間の独立性に関し、検討しておく必要がある。検索集合 R_0 と R_1 との差は I を含むか否かだけであるから、e^0_j と e^1_j との間の独立性は、x_{Ij} と x_{ij} $(i \in R_1)$ との相対的な大小関係と「丸め」のオーダーとに依存するが、もし e^0_j と e^1_j が常に同じ値をとる場合には、「丸め」には「撹乱」としての効果はほとんどないことになる。なお、e^0_j と e^1_j が独立の場合の e_j の密度関数は、

$$f(z)=\begin{cases}(2z+2n+1)/2n(n+1) & (-n-0.5\leqq z\leqq -0.5)\\ 1/(n+1) & (-0.5\leqq z\leqq 0.5)\\ (-2z+2n+1)/2n(n+1) & (0.5\leqq z\leqq n+0.5)\end{cases}$$

となる。第2表の数値はこの式によって求めたものである。

11) 表中の $n=100$、$\varepsilon=10$ の項は、誤差の大きさが x_{ij} の値の約1%に相当する $\varepsilon=10$ を超える確率が、約90%(89.16%)であることを示す。

の独立な確率変数の和の分布は、標準偏差 $\sigma/\sqrt{L}$ の正規分布に収束するから、前に説明した、サイズ $n+1$ の検索集合 R_0 への平均値検索とサイズ n の検索集合 R_1 への平均値検索を組合せて個体 I のデータへと侵入しようとするケースにおいても、この組合せ検索を L 回繰り返せば、x_{Ij} に関する L 個の不偏推定量 w^k_{Ij} $(k=1\cdots L)$ が得られるので、これらを係数和1の条件を満足させながら線形結合した、

$$w^*_{Ij}=\frac{1}{L}\sum_{k=1,L} w^k_{Ij}$$

も x_{Ij} の不偏推定量を与え、かつ、その標準偏差は個々の w^k_{Ij} の標準偏差の $1/\sqrt{L}$ となる。[12] つまり、侵入者は、個々の検索ではそれ程「精度の高い推論値」が得られない場合であっても、検索を多数回繰り返していくことにより、いずれは十分「精度の高い推論値」に到達することができることになる。[13] このような「繰り返し検索による撹乱のホワイトノイズ化」ともいうべき侵入方法に対しても、より防御力の強いデータ撹乱の方法は考えられないのであろうか。

次に述べる「人為的なデータ撹乱」のアイディアは、上で述べたような問題に応えて、検索応答に際し一定の人為的なデータ撹乱を加えることの具体的な方式およびその効果を考察するものである。

(3) 人為的なデータ撹乱 ―その1―

データベースに収容されている個別データの持つ情報の有効利用の観点から、その任意の部分集合に対する検索を許しつつ、プライバシー保護の観点から、検索応答に一定の人為的な撹乱を与えるというアイディアは、Beck [1980] によって最初に示された。ここでは、Beck のアイディアに沿いながら、このような人為的な撹乱の効果と問題について説明するが、撹乱項の与え方やその効果に関する評価については、本論文としての独自の変更を加えてある。Beck のアイディアそのものについて関心を持つ読者は、オリジナルの文献をも参照されたい。

ここで検討する撹乱付与の考え方とは、データベース中の第 i 個体のデータ x_i に関し、[14] そのデータにアクセスする第 k 番目の検索に対して、ある人為的な撹乱 e_{ik} を定義し、x_i $(i=1\cdots N)$ の統計量に関する検索が

---

12) 厳密にいえば、このことが成立するためには、異なる k における w^k_{Ij} の撹乱項につき、①互いに独立であること、②等しい標準偏差を持つこと、の条件が満足される必要がある。もっとも、①の条件は、先に論じた e^0_j と e^1_j の独立性に比較すれば、はるかに満足され易い条件であるし（あるいは、この条件が満足されるように検索集合を選ぶことが可能であるし）、②の条件については、仮にこれが満足されなかったとしても、そのときは w^*_{Ij} の推論式を $w^*_I=\sum_{k=1,L}\lambda_k w_{Ij}$ として、その係数 λ_k を $\Sigma\lambda_k=1$ の条件で探索することにより、本文で述べたのと同等の効果を期待することができるので、いずれもここでの議論に不可欠の条件でない。

13) データベースのサイズが N、検索集合のサイズを n とすれば、R_0 は ${}_{N-1}C_n$ 通り存在するから、L の最大値も ${}_{N-1}C_n$ である。これは、ある程度大きな N の値の下では、ほとんど上限がないといってよいほど多数回の繰り返し検索が可能であることを示すものである。

14) 本来なら、第 i 個体の第 j 属性という意味で x_{ij} と表記すべきであるが、表現の単純化のため以下の説明では添字 j を省略する。

行われた場合に、真の値x_iに代えてx_i+e_{ik}を応答する仕組みを考え、この下でe_{ik}の与え方を操作して所期の目的を達成しようというものである。具体的にいえば、検索集合Rの合計値$\sum_{i\in R}x_i$への検索に対して、

$$T(R)=\sum_{i\in R}(x_i+e_{ik})$$

を応答するとの想定の下での、[15]e_{ik}のデザインについて論ずることになる。なおこの項の議論では差し当たり、Rのサイズについては制限がないものとする。

人為的なデータ撹乱のアイディアは、$\bar{x}$をデータベース全体を通じるx_iの平均値とし、$\bar{x}_r$を検索集合Rを通じたx_iの平均値とし、またnを検索集合Rのサイズとしたとき、

$$E(Y_{ik})=0 \quad V(Y_{ik})=d^2$$
$$E(Z_{ik})=0 \quad V(Z_{ik})=\frac{d^2}{n}(\bar{x}_r-\bar{x})^2 \tag{1}$$

という条件でランダムな値をとる確率変数Y_{ik}とZ_{ik}とを考え、撹乱e_{ik}を、

$$e_{ik}=(x_i-\bar{x}_r)Y_{ik}+Z_{ik} \tag{2}$$

という形式で構成するものである。ここで、Y_{ik}およびZ_{ik}は異なるiおよびkについて全て独立であるとする。また、dはデータベースの管理者が決定する実数パラメータであるが、その意味については後で明らかになる。

e_{ik}の期待値は0であるから、検索者が得る応答$T(R)$は、真の値の不偏推定量であるが、その分散は、

$$V(T(R))=d^2\sum_{i\in R}(x_i-\bar{x}_r)^2+d^2(\bar{x}_r-\bar{x})^2$$

として与えられる。したがって、検索者はRについてその平均値の不偏推定量を$\frac{1}{n}T(R)$として得ることができ、その分散は、

$$V\left(\frac{1}{n}T(R)\right)=\frac{d^2}{n}S_r^2+\frac{d^2}{n^2}(\bar{x}_r-\bar{x})^2 \tag{3}$$

となる。ここで、

$$S_r^2=\frac{1}{n}\sum_{i\in R}(x_i-\bar{x}_r)^2$$

は、検索集合Rにおける標本分散である。したがって、検索集合Rの平均値に関して検索者が得られる応答は、

① 検索集合に含まれるデータの分散が大きくなるに従って増加し、

② 検索集合の平均$\bar{x}_r$がデータベース全体の平均$\bar{x}$から隔たるに従って増加するが、

③ 検索集合のサイズnを大きくとれば減少する、

という強さ（＝拡がり＝分散）を持つノイズを含むことになる。①および②は、検索集合内のデータが「隙間」を大きくとって並んでいる程、応答に強いノイズが加わることを意味し、[16]③は、検索者がノイズが強すぎると感じた場合には、検索集合のサイズを大きくとることにより問題を解決できることを意味する。

次に、このような撹乱が加えられたデータを組合せて標的Iのデータx_Iを知ろうとする侵入者が、どのような問題に出会うかをみてみるが、その前に、$I\in R$であるとき、

---

15) この項の議論は、平均値検索でも合計値検索でも妥当するが、ここでは数式的な表現の簡単な合計値検索を例にとって検討を進めよう。

16) ①については自明であろう。②については、$\bar{x}$の近くで頻度が高くなるような「普通」の分布をx_iについて想定する場合のみいえることである。この問題については、もう一度後で触れる。

$$V(T(R)) = d^2(x_I - \overline{x_r})^2 + d^2 \sum_{(i \neq I) \cap (i \in R)} (x_i - \overline{x_r})^2 + d^2(\overline{x_r} - \overline{x})^2$$
$$\geqq d^2(\overline{x_I} - \overline{x_r})^2 + d^2(\overline{x_r} - \overline{x})^2$$
$$\geqq \frac{d^2(x_I - \overline{x})^2}{2} \quad (4)$$

が成立することを確認しておく。[17]

さて、侵入者が試みるのは、多数（L 回）の検索結果を線形結合して x_I を推定しようとするタイプの不法侵入であるから、これは、侵入者が使用する検索集合を R_k $(k = 1 \cdots L)$ と表すとすれば、[18] x_I に関してその推論値を、

$$w_I = \sum_{k=1,L} b_k \cdot T(R_k) \quad (5)$$

というかたちで得ようとするものとして定式化できる。ここで b_k は R_k $(k = 1 \cdots L)$ の選び方に応じて決まる係数列であり、形式的にいえば、先に線形不法侵入 $P = A^{-1}B$ として説明した行列 A の第 i 行ベクトルの要素列、または、そのようにして求めた要素列をさらに線形結合したものである。[19]

ところで、侵入者の目的は x_I のできるだけ良い（ノイズの小さい）推論値を w_I として得ることであるから、これを推論値 w_I の真の値 x_I からの誤差の2乗で計ることとし、その期待値（ノイズの期待値）を $r(w_I)$ と表記することとすれば、[20] $r(w_I)$ は、

$$r(w_I) = \sum_{k=1,L} b_k^2 \cdot V(T(R_k)) \quad (6)$$

とすることができるので、[21] 侵入者にとっての問題は、この $r(w_I)$ の値を最小化するような検索集合 R_k $(k = 1 \cdots L)$ の選び方の問題を解くことにほかならない。

ここで議論の単純化のために、検索対象となるデータベース中に値0となるデータが多数存在し、検索者はそのようなデータを多数（$n-1$ 個）選んできて、これと x_I を組合せることにより検索集合を作るとした場合について考えることとしよう。このような検索集合を S と書くこととすれば、

$$E(T(S)) = x_I$$

であるから、侵入者は、同一の S に対する検索を繰り返すことにより、

$$w_I = \sum_{k=1,L} \frac{1}{L} T(S) \quad (7)$$

を推論値として得ることができる。w_I に含まれるノイズは、検索集合のサイズ n と x_I と $\overline{x}$ の大きさ（位置関係）に依存するが、ここで x_I と $\overline{x}$ が原点0を挟んで反対の位置にある、すなわち、

$$x_I = -\overline{x}$$

であるときについて考えることとすれば、侵入者は n を大きくとることによって、w_I の

---

17) 第1段目の展開は $i \neq I$ の x_i を無視しただけである。また第2段目の展開は、$y_1 = x_1 - \overline{x_r}$、$y_2 = \overline{x_r} - \overline{x}$ とおいて、$(y_1)^2 + (y_2)^2 - (y_1+y_2)^2/2 = (y_1-y_2)^2/2 \geqq 0$ によって得られる。

18) ただし、R_k は異なる k について同一の集合が選ばれることを禁じられていない。

19) 異なる k について同一の集合が R_k として選ばれると、x_I の推定値 w_I は、複数の線形不法侵入として定義されるベクトルをさらに線形結合したものとして与えられる。

20) 「w_I の分散」という簡潔な呼び方を避けた理由は、場合により、この $r(w_I)$ の演算手順が通常の「分散」のそれと異なることがあるからである（2.(4)参照）。

21) 攪乱項は異なる k について独立であることに注意しなければならない。

ノイズを小さくすることができ、n の無限大極限としては、

$$r^*(w_I)=\frac{d^2(x_I-\bar{x})^2}{2L} \tag{8}$$

にまでノイズを抑えることができる。(4)式の不等号条件を考慮すれば、これは x_I の値を知ろうとする侵入者が到達できる最小のノイズと等しい。[22] したがって、本論文ではこのような検索集合 S を「侵入者にとって最適な検索集合」とし、同一の S についての検索を繰り返す侵入方法を同じく「最適な侵入方法」と呼ぼう。

いうまでもなく、ここでの「最適」とは、侵入者にとって上で示したような意味で都合のよい条件が与えられたケースについてのみいえることである。そのような条件が満足されなければ、別の侵入方法が「最適」であることは十分あり得るし、また、もっと単純な侵入方法の方が、それが本当に「最適」か否かは別として、「実用的」であるかもしれない。[23] しかし、以下の議論では、差し当たってこのような侵入方法が推論値のノイズを最小にする方法の少なくとも1つであることに注目し、このケースに絞って検討を進める。

さて、侵入者にとって最適の侵入方法が上のような形式で与えられるとすれば、データベースの管理者にとっては、(8)式をデータベースの「保護基準」として設定することができる。すなわち、ある定数 c を用いて、

$$r(w_I)>c^2(x_I-\bar{x})^2 \tag{9}$$

という形式で下限が画される「ノイズ」を侵入者に与えることをデータベース管理者の目的とし、その定数 c を、「常識的に考慮すべきデータベースの検索回数の上限値」を C としたとき（$L\leqq C$ としたとき）、

$$c=\frac{d}{\sqrt{2C}} \tag{10}$$

として与え、これをデータベース保護の「基準」とするのである。

(4) 人為的なデータ撹乱　—その2—

ところで、前項で考えたような撹乱の与え方は、実用的にみてまだ十分なものではない。その理由は、あらゆる角度からみて十分といい得る水準まで検索実行回数 L を大きく想

---

22) 前記(4)式を検討すれば、x_I と $\bar{x}$ が0を挟んで反対の位置にあるとき、n を大きくすることにより $V(T(S))$ が $d^2(x_I-\bar{x})^2/2$ に近付くことは明らかである。また、検索集合を k にかかわらず $R_k=S$ とすれば $V(T(S))$ は同一の値をとるから、ここで w_I が x_I の不偏推定量であることより $\Sigma b_k=1$ となる必要があることを考慮すれば、$\Sigma(b_k)^2$ の最小化条件（これが $r(w_I)$ の最小化条件となる）として $b_k=1/L$ が求められる。

23) 例えば、他に何の情報もない侵入者にとっては、x_I を唯一の要素とする検索集合を取り出して、これに L 回の繰り返し検索を行うことは、十分「実用的」な方法であろう。この場合、推論値のノイズは、$r(w_I)=d^2(x_I-\bar{x})^2/L$ となるので、上で示した「最適」の侵入方法の場合よりは大きなノイズを被ることとなるが、その程度は「最適」のケースのたかだか2倍であるし、何よりも、侵入者にとってデータの分布や事前情報にかかわりなく実行可能な侵入方法であるうえ、侵入者が事前情報に頼らずにノイズの大きさを自ら評価できるから、これは十分に「実用的」といえるからである。なお、このような侵入の仕方が行われた場合のリスク評価は2.(4)の展開に従えば、容易に実行可能であり、しかもその結果は、侵入者のノイズを大きくするという意味で本論文で検討するよりも侵入者にとって不利（データベース管理者にとって有利）なので、以下ではこのケースを明示的には扱わないこととする。

定しながら、データベースの「安全度」を示す基準値 c を切り下げないようにするためには、d を十分に大きくとっておく必要があるが、これは「正当」な理由に基づくデータベース検索をも無意味にしてしまうまでに、個々の検索応答に対するノイズを大きくしてしまう可能性があるからである。

この問題に対する解決策として、Beck [1980]は、検索応答に加える2つの撹乱項を多数の(小)撹乱項の和として構成しておき、検索応答の度に(小)撹乱項を操作する、という方式を提案している。以下では、このBeckのアイディアを2.(3)で示した「最適」な侵入のケースに適用して、どのようなことがいえるかを考えてみる。

われわれの課題は、データベースの第 i 番目のデータに関する第 k 番目の検索に対して、真の値 x_i に、

$$e_{ik}=(x_i-\overline{x_r})Y_{ik}+Z_{ik}$$

で定義される撹乱項を加えたものを応答するという枠組みの下で、そのような検索を多数回（C 回）繰り返したときに得られる x_i の推論値 w_i に含まれるノイズ $r(w_i)$ について、

$$r(w_i)\geqq c^2(x_i-\overline{x})^2 \qquad (11)$$

;c はデータ保護基準として外から与えられる正の定数

という形式の安全基準をクリアーしながら、他方、「正当」な検索（特定データの開示のための繰り返し検索でない検索、任意の検索集合に対する1回限りの検索）について、その真の値からの誤差（分散）をできるだけ小さくしたい、というものである。ここで、前項の問題設定と異なるのは、撹乱項 Y_{ik} および Z_{ik} を単独の確率変数とするのではなく、多数（m 個）の確率変数の和として構成するところにある。具体的には、m をデータベース管理者が決める正の整数として、Y_{ik} および Z_{ik} を以下のように定義する。

$$Y_{ik}=\sum_{h=1,m} Y_{ik}^{(h)} \qquad Z_{ik}=\sum_{h=1,m} Z_{ik}^{(h)} \qquad (12)$$

ここで、$Y_{ik}^{(h)}$ および $Z_{ik}^{(h)}$ は、1回の検索においては（k を固定した場合においては）、異なる i および h について全て互いに独立に与えられる確率変数であり、その平均および分散は、

$$E(Y_{ik}^{(h)})=0 \quad V(Y_{ik}^{(h)})=d^2$$

$$E(Z_{ik}^{(h)})=0 \quad V(Z_{ik}^{(h)})=\frac{d^2}{n}(\overline{x_r}-\overline{x})^2$$

とする。d は、前項と同様、データベース管理者が決める正の定数である。

Beck [1980]のアイディアの基本は、k を固定する、すなわち、1回限りの検索においては、$Y_{ik}^{(h)}$ と $Z_{ik}^{(h)}$ を互いに全て独立に与えるが、k を変化させる、すなわち、検索を繰り返した場合においては、$Y_{ik}^{(h)}$ と $Z_{ik}^{(h)}$ の一部だけを変化させることにより、データベースへの侵入者がノイズを除去するときの「効率」を悪化させ、問題を解決しようというものである。具体的には、次のような手順を考える。

データ x_i に関する第1回目の検索に応答するときには、全ての $Y_{ik}^{(h)}$ と $Z_{ik}^{(h)}$ を上の条件に従って互いに独立かつランダムに生成する。しかし、データ x_i に関する第2回目以降、第 ℓ 回までの検索については、$Y_{ik}^{(m)}$ と $Z_{ik}^{(m)}$ のみをランダムに生成し、$1\leqq h\leqq m-1$ の $Y_{ik}^{(h)}$ については第1回目に得られた値をそのまま、一方、$Z_{ik}^{(h)}$ については第1回目に得られた値に検索集合のサイズと平均からの隔たりに応じたスケール調整を行ったものを使用

する。[24] $\ell+1$ 回目の検索については、$Y_{ik}^{(m)}$ および $Z_{ik}^{(m)}$ のほか、$Y_{ik}^{(m-1)}$ および $Z_{ik}^{(m-1)}$ の値もランダムに再生成し、その後 2ℓ 回目までは、$Y_{ik}^{(m-1)}$ については直近の再生成値を、また、$Z_{ik}^{(m-1)}$ については直近の再生成値にスケール調整を行ったものを使用する。$2\ell+1$ 回目では、再び $Y_{ik}^{(m-1)}$ および $Z_{ik}^{(m-1)}$ も再生成し、同様の操作を ℓ^2 回目まで繰り返す。ℓ^2+1 回目では $Y_{ik}^{(m-2)}$ および $Z_{ik}^{(m-2)}$ も再生成する。以上の手順を図解すると第1図のようになる。

第1図

1	回目	$Y_{i1}^{(1)}$	$Y_{i1}^{(2)}$	$Y_{i1}^{(3)}$	……	$Y_{i1}^{(m-2)}$	$Y_{i1}^{(m-1)}$	$Y_{i1}^{(m)}$
2	〃	↓	↓	↓	……	↓	↓	$Y_{i2}^{(m)}$
⋮								⋮
ℓ	〃	↓	↓	↓	……	↓	↓	$Y_{i\ell}^{(m)}$
$\ell+1$	〃	↓	↓	↓	……	↓	$Y_{i\ell+1}^{(m-1)}$	$Y_{i\ell+1}^{(m)}$
$\ell+2$	〃	↓	↓	↓	……	↓	↓	$Y_{i\ell+2}^{(m)}$
⋮								⋮
2ℓ	〃	↓	↓	↓	……	↓	↓	$Y_{i2\ell}^{(m)}$
$2\ell+1$	〃	↓	↓	↓	……	↓	$Y_{i2\ell+1}^{(m-1)}$	$Y_{i2\ell+1}^{(m)}$
$2\ell+2$	〃	↓	↓	↓	……	↓	↓	$Y_{i2\ell+2}^{(m)}$
⋮								⋮
ℓ^2	〃	↓	↓	↓	……	↓	↓	$Y_{i\ell^2}^{(m)}$
ℓ^2+1	〃	↓	↓	↓	……	$Y_{i\ell^2+1}^{(m-2)}$	$Y_{i\ell^2+1}^{(m-1)}$	$Y_{i\ell^2+1}^{(m)}$
ℓ^2+2	〃	↓	↓	↓	……	↓	↓	$Y_{i\ell^2+2}^{(m)}$
⋮								⋮
ℓ^m	〃	↓	↓	↓	……	↓	↓	$Y_{i\ell^m}^{(m)}$

(注) $Y_{ik}^{(h)}$ の記述があるところはランダムに発生させた変数、'↓'とあるところは、前の回の値を使用するもの。$Z_{ik}^{(h)}$ についても同様。

このように、ℓ^j 回 (j は $1 \leqq j \leqq m$ の整数) の検索に応答する毎に、再生成する撹乱項の範囲を拡大していく手順を考えたとき、この手順がデータベース中のサイズ n の検索集合に対する平均値検索と、特定のデータに対する侵入を目的とした繰り返し検索によって得られる推論値とに、各々どのような影響(ノイズ) を与えるかを検討してみよう。

最初にサイズ n の検索集合 R に対する平均値検索への応答に加えられる撹乱項の分散を計算してみよう。1回の検索について考える場合 (k を固定して考える場合)、第 i 番のデータに加えられる撹乱項を構成する $2m$ 個の確率変数、$Y_{ik}^{(h)}$ と $Z_{ik}^{(h)}$ $(h=1\cdots m)$ は互いに独立であるから、

$$V(Y_{ik})=V\left(\sum_{h=1,m} Y_{ik}^{(h)}\right)=md^2$$

$$V(Z_{ik})=V\left(\sum_{h=1,m} Y_{ik}^{(h)}\right)=\frac{md^2}{n}(\overline{x_r}-\overline{x})^2$$

である。また、検索集合 R を構成する n 個の x_i に加えられる撹乱項 ($2m\times n$ 個の撹乱項) も、上記の撹乱項の与え方に関する説明で明らかなとおり、全て互いに独立であるから、

$$V\left(\frac{1}{n}T(R)\right)=V\left(\frac{1}{n}\sum_{i\in R}(x_i-\overline{x_r})^2\ Y_{ik}+\frac{1}{n}\sum_{i\in R}Z_{ik}\right)$$

$$=\frac{md^2}{n}S_r^2+\frac{md^2}{n^2}(\overline{x_r}-\overline{x})^2 \quad (13)$$

を得る。すなわち、「正当」な平均値検索に

24) 具体的には、第1回目の検索集合の平均およびサイズを、$\overline{x}_r^*$ および n^* としたとき、

$$\theta_r=\sqrt{\frac{(\overline{x_r}-\overline{x})/(\overline{x}_r^*-\overline{x})}{n/n^*}}$$

という形式のスケール・パラメータを考え、これを第1回で生成した $Y_{ik}^{(h)}$ と $Z_{ik}^{(h)}$ の値に乗じて、第2回目以降の撹乱項の値とする。

対して加えられるノイズは、md^2に比例して大きくなることがわかる。

次に、侵入者側の問題を考えてみよう。検討すべきは、侵入者が最適の方法で標的データx_Iを開示することを試みたとき、その推論値w_Iに含まれるノイズ$r(w_I)$の大きさであるが、問題を複雑にしているのは、検索集合に加えられる撹乱項：

$$e_{ik}=(x_i-\overline{x_r})\sum_{h=1,m}Y_{ik}^{(h)}+\sum_{h=1,m}Y_{ik}^{(h)}$$

が異なるkについて独立の確率変数ではなくなっていることである（前述の$Y_{ik}^{(h)}$および$Z_{ik}^{(h)}$の変化のさせ方を参照）。以下では、この点に注意して2.(3)で示したような意味で「最適」な方法をとった侵入者が被るノイズの大きさを評価する。

① 表現の簡単化のために、

$$e_{ik}^{(h)}=(x_i-\overline{x_r})Y_{ik}^{(h)}+Z_{ik}^{(h)}$$

の書き換えを行う。そうすると、$1\leqq L\leqq\ell$のとき、(7)式により、

$$\begin{aligned}w_I&=x_I+\frac{1}{L}\cdot\sum_{k=1,L}\cdot\sum_{h=1,m}\cdot\sum_{i=1,n}e_{i1}^{(h)}\\&=x_I+\sum_{h=1,m-1}\Big(\sum_{i=1,n}\frac{1}{L}\sum_{k=1,L}e_{ik}^{(h)}\Big)\\&\quad+\sum_{i=1,n}\frac{1}{L}\sum_{k=1,L}e_{ik}^{(m)}\\&=x_I+\sum_{h=1,m-1}\cdot\sum_{i=1,n}e_{i1}^{(h)}+\frac{1}{L}\sum_{k=1,L}\cdot\sum_{i=1,n}e_{ik}^{(m)}\end{aligned}$$

となるから、[25]

$$\begin{aligned}r(w_I)&=\sum_{h=1,m-1}V\Big(\sum_{i=1,n}e_{i1}^{(h)}\Big)\\&\quad+\frac{1}{L^2}\sum_{k=1,L}V\Big(\sum_{i=1,n}e_{ik}^{(m)}\Big)\\&\geqq\frac{d^2(x_I-\overline{x})^2}{2}(m-1+\frac{1}{L})\end{aligned}$$

を得る。[26]

② sを$1\leqq s\leqq\ell-1$の整数として、$s\ell+1\leqq L\leqq(s+1)\ell$のとき、同様の考え方により、

$$\begin{aligned}r(w_I)=\frac{d^2(x_I-\overline{x})^2}{2}\{m-2\\+\frac{\ell^2s+(L-s\ell)^2}{L^2}+\frac{1}{L}\}\end{aligned}$$

であり、特に$L=\ell^2$（$s=\ell-1$）のとき、

$$\begin{aligned}r(w_I)=\frac{d^2(x_I-\overline{x})^2}{2}(m-2+\frac{1}{\ell}\\+\frac{1}{\ell^2})\end{aligned}$$

である。

③ tを$1\leqq t\leqq\ell-1$の整数として、$s\ell^2+t\ell+1\leqq L\leqq s\ell^2+(t+1)\ell$のとき、やはり同様の考え方により、

---

25) 第2段目の変形は和の順序変更である。第3段目第1項の変形は$1\leqq L\leqq\ell$のとき、$e_{ik}^{(h)}$の値は$e_{i1}^{(h)}$を使用するので、この値で代表させた。

26) この不等式は、(4)式と同様の不等号条件の検討により得られる。すなわち、次のようにする。

$$\begin{aligned}V\Big(\sum_{i=1,n}e_{ik}^{(h)}\Big)&=V\Big[\sum_{i=1,n}\{(x_i-\overline{x_r})Y_{ik}^{(h)}+Z_{ik}^{(h)}\}\Big]\\&=d^2\sum_{i=1,n}(x_i-\overline{x_r})^2+d^2(\overline{x_r}-\overline{x})^2\\&\geqq d^2(x_I-\overline{x_r})^2+d^2(\overline{x_r}-\overline{x})^2\\&\geqq d^2(x_I-\overline{x})^2/2\end{aligned}$$

$$r(w_I)=\frac{d^2(x_I-\overline{x})^2}{2}\{m-3+\frac{\ell^4 s+(L-s\ell^2)^2}{L^2}+\frac{\ell^2(s\ell+t)+(L-s\ell^2-t\ell)^2}{L^2}+\frac{1}{L}\}$$

であり、特に$L=\ell^3(s=t=\ell-1)$のとき、

$$r(w_I)=\frac{d^2(x_I-\overline{x})^2}{2}(m-3+\frac{1}{\ell}+\frac{1}{\ell^2}+\frac{1}{\ell^3})$$

である。

④ 以上の手順を繰り返せば、一般に$L=\ell^j$のとき、

$$r(w_I)=\frac{d^2(x_I-\overline{x})^2}{2}(m-j+\frac{1}{\ell}+\frac{1}{\ell^2}+\cdots+\frac{1}{\ell^j}) \quad (14)$$

となる。

ところで、x_Iに加える撹乱項はm個の(小)撹乱項の和として構成され、ℓ回の検索に応答する毎に再生成する(小)撹乱項の範囲を拡大するのであるから、この方式で対応できる検索回数はℓ^m回までである。したがって、われわれは、ℓ^m回の検索が、全て特定データx_Iの開示のための「最適」検索の繰り返しとして実行されたときの問題を考えればよいことになり、そのときの$r(w_I)$の値は、(14)式に$j=m$を代入することによって与えられるから、

$$r(w_I)=\frac{d^2(x_I-\overline{x})^2}{2}(\frac{1}{\ell}+\frac{1}{\ell^2}+\cdots+\frac{1}{\ell^m}) \quad (15)$$

となる。

以上の結果により、われわれの課題は先に(11)式として示した不等式に(15)式で得た$r(w_I)$の値を代入して得た、

$$d^2\geqq\frac{2c^2}{\frac{1}{\ell}+\frac{1}{\ell^2}+\cdots+\frac{1}{\ell^m}} \quad (16)$$

;cはデータ保護基準として外から与えられる正の定数

という形式の安全基準を、

$$\ell^m=C$$

;Cは繰り返し検索許容上限回数として外から与えられる正の整数

回の検索の範囲内で満足させながら、「正当」な検索（1回限りの検索）に含まれる誤差のスケール：

$$md^2$$

を最小化しようという問題に帰着することがわかる。

ここで、(16)式右辺の分母部分は、次のように変形することができる。

$$\frac{1}{\ell}+\frac{1}{\ell^2}+\cdots+\frac{1}{\ell^m}=\frac{1}{\ell}\cdot\frac{1-1/\ell^m}{1-1/\ell}=\frac{1-C}{\ell-1}$$

また、$\ell^m=C$により、

$$m=\log C/\log\ell$$

を得ることができるので、結局、最小化すべき誤差のスケールmd^2には、

$$md^2 \geqq \frac{c^2 \cdot \log C}{1-C} \cdot \frac{1-\ell}{\log \ell} \qquad (17)$$

という形式での下限値が存在することがわかる。これは、データベース管理者に対し、想定すべき最大検索回数Cと、データ保護基準cとを与えれば、md^2について実現できる最小値はパラメータℓの与え方に依存することを示す。具体的には、md^2に下限値を与えるℓの値は、関数：

$$y=\frac{1-\ell}{\log \ell}$$

に最小値を与える$\ell \geqq 2$の整数として与えられ、すなわち、

$$\ell = 2$$

となる。[27] このことから直ちに、Cおよびcが与えられたときの最適なm、dの与え方も明らかとなって、それは、

$$2^m = C$$

$$d^2 = \frac{2c^2}{\frac{1}{2}+\frac{1}{2^2}+\cdots+\frac{1}{2^m}}$$

であると結論される。そのようなℓ、m、dの与え方が行われたときに、最大限C回の検索に対しても個別データへの侵入を許さずに、「正当」な検索者に与えるノイズを最小限に抑えることができるのである。

3. 人為的なデータ撹乱の実用性

(1) 数値例

2.では、自由検索可能な統計的データベース中の個別データへの「不法」なアクセスを防止するために、検索応答に一定の人為的な「撹乱」を加えるという方法が考えられることを示した。しかし、そこでの説明は、あくまでもそのような方法の存在可能性を理論的に示しただけであって、その実用性について具体的な評価を示したものではない。ここでは、この実用性の問題につき、若干の数値例を使って考えてみることとしよう。

数値例に現実のデータを使用する訳にはいかないので、擬似データで議論することとしよう。対象とするのは、1989年の全国勤労者世帯（標準世帯）の年間収入である。もっとも、家計調査報告は平均収入を公表しているが、その分布を公表していないから、擬似データを得るためには、分布の状況を推定しなければならない。ここでは、分布を対数正規分布と想定し、その標準偏差については、家計調査報告の17階級別所得から推定することにより、擬似母集団を想定した。[28]

さて、この対数正規分布に従う母集団から、N個の標本を抽出したものが、他の属性データ（例えば、消費データ）と一緒に、自由検索可能なデータベースとして公開されたと考えてみよう。問題にしている状況を最も単純化するとすれば、それは「標的とする家計に

---

27) $dy/d\ell = \{\log \ell - (\ell-1)/\ell\} / (\log \ell)^2$であるが、この微分は、$\ell \geqq 2$のとき$dy/d\ell > 0$なので、条件を満たす$\ell$は$\ell = 2$である。

28) 所得対数値の平均を15.502108とし、標準偏差を0.3599502とした。因みに、所得金額の平均は5,762,550円、標準偏差は2,143,270円である。

ついての所得情報を握っている者（侵入者）が、このデータベースを利用することにより、その標的のデータがデータベース中のどのデータであるのかを知ることができるか」という形式で与えられる。このような問題設定を行う理由は、もし、標的とする所得データがどれであるかを特定することができれば、侵入者はそのデータの欄（行）を横に辿ることによって、その家計の全ての消費行動を「不法」に知ることができてしまうからである。[29)]

2.で説明した方法に従えば、データベース管理者は、データベースのサイズNが与件とされている状況の下で、個別データへの侵入を企てる者が、多数回（2^m回）の検索結果を複合したとしても、そこから得られる第i個体のデータx_iの推論値w_iに関し、

$$r(w_i) \geqq c^2(x_i - \bar{x})^2$$

あるいは、これを標準偏差のタームに直した、[30)]

$$\sqrt{r(w_i)} \geqq c \mid x_i - \bar{x} \mid$$

以上のノイズが残るよう、パラメータcとmを操作することになる。ここで、cやmを大きく設定すれば、その分だけ個別データは強く保護されることになるが、「正当」な検索に対しても、すなわちデータベース公開の趣旨に合致した検索に対しても、より大きなノイズを含んだ応答を返すこととなってしまうので、データベースの「価値」はそれだけ損われる。したがって、ここで紹介した方法が実用性ありといえるかどうかは、このようなトレードオフ関係をどう評価するかにかかることになる。以下では、そうした観点から、若干の数値例を作ってこの問題を考えてみることとしよう。

－数値例１－

まず$N=10^4$、すなわち10,000個体分のデータを収容したデータベースを、パラメータ$m=34$、$c=0.1$で定義されるような撹乱を加えて公開するとしよう。このパラメータ設定は、$2^{34}=1.717\times10^{10}$（100億回以上）の検索結果を複合して$x_i$を推定しようとしても、[31)]それから得られる個別データx_iの推論値w_iには、少なくとも$\sqrt{r(w_i)}=0.1 \mid x_i - \bar{x} \mid$のノイズが保証されることを意味する。ここで$\sqrt{r(w_i)}$の大きさは、標的となっているデータ$x_i$の平均からの距離$\mid x_i - \bar{x} \mid$に依存することに注意して、かなり分布の「縁」に近いデータとして、所得ランク10^3位（所得が大きい方から1,000番目）のデータx_1と、分布の「中心」に近いデータとして、所得ランク6×10^3位（所得が大きい方から

29) もちろん、ここで侵入者が知ることができる消費に関するデータには、所得に関するデータと同様、一定の撹乱が加えられている。しかし、もし侵入者が所得に関するデータについて「十分」に撹乱の影響を除去することができるとすれば、消費に関するデータについても同じように「十分」に撹乱の影響を除去することができてしまうであろう。

30) $r(w_i)$は、推論値w_iの真の値x_iからの誤差の２乗の期待値であるから、$r(w_i)$がw_iのノイズ（w_iのx_iからのズレ）を分散のタームで計ったものとすれば、その平方根をとった$\sqrt{r(w_i)}$は、ノイズを標準偏差のタームで計ったものに相当する。

31) この100億回という検索回数は、通常「高速オンライン」といわれている銀行の勘定系オンラインの最高処理速度（数100件／秒）をもってしても、全検索の終了に数100日を要する量である。

6,000番目）のデータx_2を選んで、その周りの検索集合が含む誤差の大きさを評価してみよう。

x_1 ＝ 8,566,782、 x_2 ＝ 4,930,326、$\bar{x}$＝5,762,547であるから、$\sqrt{r(w_1)}$＝280423.5、$\sqrt{r(w_2)}$＝83222.1となる。一方、x_1、x_2付近でのデータの平均間隔をD_1、D_2とすると、D_1＝1757.1、D_2＝459.4であるから、[32]これと標準偏差で計ったノイズの大きさとの比率を求めると、$\sqrt{r(w_1)}/D_1$＝159.6、$\sqrt{r(w_2)}/D_2$＝181.2となる。この比率をどう評価するかは基本的にはデータベース管理者の問題だが、比率がこれくらい大きければ、差し当たっての評価としては、データx_1およびx_2はいずれも周囲のデータと混ざりあって、十分マスクされると考えてよいであろう。[33]

ところで、次の問題は、このような撹乱が加えられたことにより、「正当」なデータベース検索者がどの程度の「迷惑」を被るかである。この点をみるために、全データベースの$1/10^3$、$1/10^2$、1/10のサイズの検索集合における平均値検索に対して加えられるノイズの大きさを計算してみたのが第3表である。

－数値例2－

第3表で示したノイズの大きさをどう判断するかは、検索者すなわち利用者が考えるべき問題である。あまりよく知られていない事柄について、おおよその傾向を判断したいというのが目的の利用者であれば、かなり大きなノイズをも許容するであろうし、逆に、既に様々な角度からの分析が尽くされているような事柄についてさらに詳細な観察を試みようとする利用者にとっては、同じ大きさのノイズが耐えられない程大きいと感じられるかもしれないからである。

問題は、データベース利用者の大半が、ノイズが大きすぎて実用に耐えないとの受け取り方をしたときの対応である。これまでの説明から明らかなとおり、データベースのサイズNを与件とした下では、パラメータmやcを切り下げればノイズは小さくなるが、それは個別データをより大きな侵入の危険にさらすことを意味する。したがって、個別データの安全性についての判断が切り下げの余地がなければ、そのようなデータは、ここで示したような方法で公表しても、結局はノイズの大きさが嫌わ

---

32) この値は、密度関数を$f(x)$、標本数Nの場合の累積度数を

$$F(x)=N\cdot\int_{-\infty}^{x} f(u)\,du$$

とすれば、x近傍のデータの平均間隔は、

$$\lim_{h\to 0} h/(F(x+h)-F(x))=1/(F(x)')=1/(N\cdot f(x))$$

で与えられることにより求めたものである。

33) この比率は、例えば侵入者が推論値に含まれるノイズ（w_iのx_iからのズレ）に関し、平均0標準偏差$c|x_i-\bar{x}|$の正規分布を想定したうえで、99%信頼区間でx_1あるいはx_2を特定しようとしても、他の個体データが、x_1については822個体分、x_2については933個体分もその信頼区間内に同居してしまうことを意味する。

第3表　$N=10^4$、$m=34$、$c=0.1$のデータベースに加えられるノイズの大きさ

	x_1付近の検索集合			x_2付近の検索集合		
	$n=10$	$n=10^2$	$n=10^3$	$n=10$	$n=10^2$	$n=10^3$
撹乱項の標準偏差	231,320	23,517	14,261	68,609	6,948	3,533
90%信頼区間	± 296,449	± 30,138	± 18,276	± 87,926	± 8,904	± 4,527
99%信頼区間	± 538,132	± 54,708	± 33,175	± 159,608	± 16,164	± 8,218
99.9%信頼区間	± 714,834	± 72,672	± 44,069	± 212,017	± 21,472	± 10,917
真の値	8,567,667	8,568,346	8,642,126	4,930,556	4,930,561	4,931,083

（小数点以下四捨五入）

(注) 1)「x_1付近の検索集合;$n=10$」とは、所得ランク1,000位のx_1の周辺からx_1を含めて$n=10$個のデータを、所得ランク995位から1,004位の個体抽出という方法により得ることを示す。

2) 撹乱項の標準偏差とは、検索者が得る「撹乱付平均値」の「真の値の平均値」からの誤差の標準偏差のことで、2.(4)で示した(13)式により算出する。

3) 90%信頼区間とは、確率90%で「撹乱付平均値」が収まる「真の平均値」の周りの区間の境界値であって、「撹乱項の標準偏差」から正規分布の仮定を用いて算出する。99%、99.9%の信頼区間についても同様である。

れて利用されないであろうという意味で、「公表にはなじまない」ということにならざるを得ない。

しかし、ここで統計調査の実施方法も自由に選ぶことができて、データベースのサイズNを変化させられるとすれば、事情は変わってくる。[34] このことを、Nを10倍または10^2倍にして、$N=10^5$または$N=10^6$としたときの問題として考えてみよう。

Nを大きくすることの効果の第1は、パラメータcを切り下げることができる点にある。数値例1で説明したように、不法侵入に対する安全度を$\sqrt{r(w_1)}/D_1$や$\sqrt{r(w_2)}/D_2$という基準で判断するとすれば、Nが10倍あるいは10^2倍になれば、概ねcを1/10あるいは$1/10^2$にしても、安全度という観点からはほぼ同じ効果が得られ、しかもcが小さくなることにより、「正当」な検索に加えられるノイズの影響は小さく抑えられるからである。

Nを大きくすることの効果の第2は、以前と同様の「精度」での観察を求める利用者が検索集合のサイズnを大きく指定できるようになる点である。仮に利用者の目的がデータベースを10^2個の部分集合に仕切って、その傾向をみたいというものであるとすれば、$N=10^4$であれば$n=10^2$と指定することとなるが、$N=10^5$であれば

34) 現在行われている統計調査のサンプル数（例えば家計調査報告なら8,000世帯）は、基本的に全調査対象を通じた平均値に大きな誤差を生じないようにするという観点から決定されたものである。したがって、統計利用者の関心が全調査対象を通じた平均値よりも個々のデータへと移ってくるとしたら、調査するサンプル数も、統計の利用目的に応じて変わってきてもよいであろう。

第4表 $N=10^5$、$m=34$、$c=0.01$のデータベースに加えられるノイズの大きさ

	x_1付近の検索集合			x_2付近の検索集合		
	$n=10^2$	$n=10^3$	$n=10^4$	$n=10^2$	$n=10^3$	$n=10^4$
撹乱項の標準偏差	5,031	1,425	445	1,298	353	110
90%信頼区間	± 6,449	± 1,826	± 570	± 1,663	± 453	± 141
99%信頼区間	± 11,704	± 3,314	± 1,035	± 3,019	± 822	± 255
99.9%信頼区間	± 15,548	± 4,402	± 1,375	± 4,011	± 1,092	± 339
真 の 値	8,631,583	8,640,206	8,641,261	4,928,550	4,930,622	4,930,876

(小数点以下四捨五入)

第5表 $N=10^6$、$m=34$、$c=0.001$のデータベースに加えられるノイズの大きさ

	x_1付近の検索集合			x_2付近の検索集合		
	$n=10^3$	$n=10^4$	$n=10^5$	$n=10^3$	$n=10^4$	$n=10^5$
撹乱項の標準偏差	143	44	14	35	11	3
90%信頼区間	± 183	± 57	± 18	± 45	± 14	± 4
99%信頼区間	± 332	± 103	± 33	± 82	± 26	± 8
99.9%信頼区間	± 440	± 137	± 43	± 110	± 34	± 11
真 の 値	8,640,206	8,641,070	8,641,176	4,930,622	4,930,830	4,930,855

(小数点以下四捨五入)

$n=10^3$、$N=10^6$であれば$n=10^4$、と指定できる。このようにnを大きく指定できれば、それは2.(4)で示した(12)式からも明らかなように、「正当」な検索に加えられるノイズの影響を小さくするはずである。

このような観点から、$N=10^5$の場合と$N=10^6$の場合とについて、数値例1で示したのと概ね同様の安全度を確保したデータベースについて、その「正当」な検索応答に加えられるノイズの大きさを評価したのが、第4表および第5表である。ここでx_1とx_2は、数値例1と相対的な位置が同じになるように、x_1はデータベースの上から1/10に位置するデータ、x_2は同じく6/10に位置するデータとしてある。第4表でノイズはかなり小さくなり、第5表であれば、通常考え得る検索目的ならほとんど問題にならないレベルまでその影響が抑えられていることが読み取れるであろう。

(2) 考慮すべき問題点

最後に、本論文で説明した人為的なデータ撹乱による個別データの保護法に関し、差し当たって考えられる問題点を何点か指摘して

おこう。

第1の問題は、$\sqrt{r(w_i)} \geqq c \mid x_i - \bar{x} \mid$ という形式の安全基準そのものの適切性である。この基準は、正規分布のように、平均値 $\bar{x}$ の近くにデータが多く存在する「一山」型の分布を持つデータに対しては合理的であるが、仮にデータが両端に近いところに多く存在し、平均値 $\bar{x}$ の近くにはほとんど存在しないような「二山」型の分布を持つデータがあったとすれば、この基準では必ずしも適切とはいえなくなる。[35] とり得る値の数が極端に少ない離散的分布に従うデータ、例えば0と1の2とおりの値しかとらないデータ等についても同様であろう。もちろん、一般にわれわれが得る統計データの大半が、正規分布あるいはそれに近い分布を持つことはよく知られた事実であるから、この安全基準を適用することが望ましくないデータがあり得るということは、それだけでは、ここで紹介したような議論の意義を損うものではない。しかし、この方法の現実データへの適用を考えるのであれば、適用対象とするデータが、そもそもこの手法に向いている分布を持つデータかどうかを、実態に即して検証しておく必要はあるというべきであろう。

第2の問題は、「極端な値をとるデータ」の取扱いである。この手法は、平均値からの距離 $\mid x_i - \bar{x} \mid$ に比例した撹乱を与えることによって、中心から隔たったデータについても個別データをマスクする効果が損われないように配慮はしているが、それでも「極端」に中心から離れたデータについては、[36] 十分なマスクが効かなくなる可能性がある。この点を重視するのであれば、極端な値をとるデータを公表の対象から外す等の特別な配慮が必要となるかもしれない。

また、「極端な値をとるデータ」とは別の意味で、平均に極めて近いデータについては、データ撹乱がほとんど加えられなくなる点について注意しておく必要がある。例えば、x_i が平均値と完全に等しいデータであるとした場合、もし、検索者が x_i を唯一の要素とする検索集合に対して検索を行ったとしたら、撹乱項は完全に0となってしまうので、ここで検索者がその検索集合に対して同じ検索を繰り返したとすれば、「常に同じ答が得られる」ということを通じて、撹乱が0であること（検索応答値が x_i の真の値であること）を推論できてしまうかもしれない。このような問題を回避したければ、検索集合のサイズの下限値を画するようなアクセス規制が必要となろう。

第3の問題は、パラメータ c の選び方である。前の項の議論では、推論値のノイズ $\sqrt{r(w_i)}$ とその付近でのデータ間隔 D_i との比率 $\sqrt{r(w_i)}/D_i$ を指標として、c の値を選ぶことを考えたが、このような配慮だけで c を決めてよいとは限らない。その理由は、侵

35) また、「分布の山」の位置と平均値の位置とがずれている場合も、この方法は「最適」とはいえなくなる。対数正規分布もこのタイプであるから、前の項で挙げた例も、実は、この方法が「最適」とはいえなくなる例の1つということになる。もっとも、対数正規分布程度の「ずれ」であれば、この方法は「最適」とはいえないものの、「不適切」という程の問題を生じるわけでもないであろう。

36) 例えば非常に有名な高額所得者のデータのようなものを考えた場合、ここで定義した撹乱ではマスクとして不十分というケースもあり得るであろう。

入者が利用する標的に関する知識とは、「単一の属性に関する正確な知識」であるとは限らないからである。もし、侵入者が利用する標的に関する知識が、「多数の属性に関する、曖昧だが広範な知識」である場合には、[37] c をある程度大きく設定しておかないと、比較的容易に個別データへの侵入が可能となってしまうことになる。このような侵入の可能性を重視するのであれば、一般的にいって、多数の属性をリストしたデータベースについては、少数の属性しか扱わないデータベースよりも c を相対的に大きく設定することが望ましいということになる。c の具体的な水準について論ずるのが本論文の目的ではないので、この問題には深入りしないが、現実的な利用を考えた場合には避けて通れない論点である。

さらに、本論文の議論が必ずしも熟したものとはいえないことから生じる様々な問題にも目を向けておく必要がある。例えば、本論文で取り扱ったデータ撹乱の手法は、個票データの利用とプライバシー保護の問題について興味ある解決策を提示するものではあるが、それが最善の解決策であるかどうかについては、何も答えられていない。応用を意識するのであれば、他の代替案との比較の観点からもこの手法に対する評価が必要となるはずである。[38] また、本論文で扱ったのは、検索者が任意に指定する検索集合について、その平均値を利用しようとした場合に生じる問題であって、より高次のモーメント等、他の統計量を利用しようとした場合に生じる問題についての議論も行っていない。[39] もちろん、これらの問題が残されていることは、推論制御の手法に関する理論的欠陥の存在を直ちに示唆するものではないが、こうした問題の存在自体、この理論の現実への応用に際し今後の検討を待つべき点が多いことを示すものなのである。

4. おわりに

本論文は、「推論制御」といわれるデータ保護の理論の1分野におけるこれまでの成果を踏まえて、それに若干の新しい解釈を加えるかたちで、統計的なデータベースを利用して標的のプライバシーに侵入する方法と、それを防御する方法に関する理論を紹介したものである。ところが、最初にも述べたとおり、この分野における重要な成果は、Beck [1980] をはじめ、主として1980年代の前半に得られたものであって、必ずしも「最新」のものではない。にもかかわらず、この理論を現実の情報公開で利用しようという動きは、米国においてもほとんどみられていない。以下、そ

37) 例えば、「○○家の収入は○円ぐらいだが、最近○円ぐらいの自動車を買ったらしく、また払っている家賃は、…」というような知識を標的に関して侵入者が持っている場合である。

38) 例えば、統計的データベース中のレコードを互いに永久に交換してしまうことによりプライバシーを保護しようという手法が存在する。この方法論は、一般にデータスワップと呼ばれるが、この方法論と本論文で紹介したデータ撹乱の方法論のどちらが優れているかについては、今後の議論に待たなければならない。なお、データスワップの具体的手法については、例えばReiss [1984] 参照。

39) Beck [1980] は、中央値および度数を統計量として利用しようとするときに生じる問題について、一定の整理を行っている。しかし高次のモーメントについて一般的にどのような現象が生じるかについての定式化は示していない。

の理由を探りながら、この理論の今後の可能性について考察して、本論文の結びとしよう。

推論制御の理論が提起されて以来かなりの年月を経過しながらも、現在までにこれといった応用の試みがなされなかったのは、すでに述べたように、この理論が想定する環境である「自由検索可能なデータベース・システム」が理論としては存在しても、実用的には不十分な能力しか示し得なかったという事情が大きいと思われる。実際、本論文の議論が多くを負っているBeck[1980]の論文が発表された当時では、「自由検索可能なデータベース・システム」の最も標準的な概念であるいわゆるリレーショナル・データベースは、極めて貧弱な性能しか示し得なかったので、これを用いて誰からでもアクセスできるような統計データの個票公開を行うなどということは、コンピュータ・システムの設定・運営費用の問題を度外視したとしても、あまり現実的な話ではなかったはずである。その当時に個票公開を考えるとすれば、本論文の最初に紹介したセンサス局のアプローチ、すなわち、非常に多数の標本から極めて少数の標本を抽出して、磁気テープのようなオフライン媒体に記録して配布してしまうことしか方法はなかったであろうし、そのような環境の下では、ここで紹介したような理論は「画に描いた餅」にすぎなかったのであろう。

しかし、そのような技術環境は最近になって大きく変化した。いわゆるコンピュータ・システムの性能は、最近の10年間で10～20倍になったといわれていることを考えれば、家計調査報告のような無作為抽出型の統計（標本数数千件～数万件）はもとより、国勢調査のような悉皆調査型の統計（標本数数百万件～数億件）についてすら、その個票をデータベースとしてアクセスさせることも可能な状況になってきている。そうした状況の変化を踏まえれば、これまで実用性の見地からはあまり省みられることのなかった推論制御の理論について、その現実データへの応用を考えるべき時がきているといってもよいのではないだろうか。

もちろん、応用を考えるべきということは、応用を開始することと同じではない。応用を開始する前に考えておくべきことは、非常に多いからである。しかし、その中でも重要なことは、この理論を応用するにふさわしいデータとは、われわれが通常「統計」として扱っているデータばかりとは限らないという点である。

これまでの説明からも明らかなとおり、プライバシー保護の観点から一定のデータ撹乱を加えながら個票を公開することの意義が大きい統計とは、一般には、標本数が非常に大きい統計である。そうした統計であれば、個別データへの侵入を有効に防止しながら、十分に有用な（ノイズの少ない）情報を利用者に提供できるからである。これは、差し当たっては、本論文でしばしば例に挙げた家計調査報告のような統計よりは、[40]国勢調査のような統計の方がこの手法を適用するのに向いているということを意味するものである。しかし、さらに踏み込んで考えれば、そのような

40) 逆に、家計調査報告のような無作為抽出型の統計について、ここでの理論を応用して個票公開を行おうとするのであれば、むしろそれに合わせて、抽出標本数を増やす等の対応を考えるべきであろう。

既存の統計の中から適用し易いものを選ぶというアプローチよりも、これまではプライバシー保護への配慮から「統計」としては利用できなかった「生」のデータを、一定の撹乱を加えることにより新たに「統計」として活用しようという観点からのアプローチの方が、より興味深いともいえるかもしれない。具体的には、税務署の課税データを経済学的な分析に利用するとか、病院のカルテ・データを集中管理して疫学的な分析に利用するといったアイディアである。こうしたアイディアが実現可能になれば、単に社会全体として情報処理コストの節約が図れるというだけでなく、情報の有効利用促進を通じ新しい有益な発見を手助けすることも期待できよう。[41)]

いうまでもなく本論文は、データ公開とプライバシー保護という相反し易い2つの目標について、その両立の可能性を論じたものであり、実用性に関する具体的な検討を狙ったものではない。実用に至る過程が本論文のようなペーパーで議論し尽くせる程単純なものではないであろうことは、断るまでもない。さらに、理論的にも本論文で「侵入者にとって最適な方法」としたものが、本当に「最適」なのかどうか、必ずしも明らかでない。この点についても今後の課題というべきであろう。しかし、本論文の狙いは、これまでわが国であまりにも知られていなかったと思われる推論制御の理論の一端を紹介することを通じて、データ公開とプライバシーの保護の問題について、改めて議論を喚起しようとするところにある。そうした観点からの今後の議論の発展を期待して、本論文の結びとしたい。

以 上

(岩村) 日本銀行金融研究所研究第1課調査役
(西島) 日本銀行金融研究所研究第1課

【参考文献】

Beck, L.L., "A Security Mechanism for Statistical Databases," *ACM Transaction on Database Systems*, Vol.5, No.3, 1980.

Cox, L.H., "Suppression Methodology and Statistical Disclosure Control," *Journal of American Statistical Association*, Vol.75, No.370, 1980.

Denning, E.D., P.J. Denning, and M.D. Schwartz, "The Tracker: A Treat to Statistical Database Security," *ACM Transaction on Database Systems*, Vol.4, No.1, 1979.

————, and J. Schlörer, "A Fast Procedure for Finding a Tracker in a Statistical Database," *ACM Transaction on Database Systems*, Vol.5, No.1, 1980.

————, "Secure Statistical Databases with Random Sample Queries," *ACM Transaction on Database Systems*, Vol.5, No.3, 1980.

41) 例えば、全国の病院のカルテのデータが集中的に管理され、ランダムな検索が可能であれば、特定の物質と疾患との間に、発生頻度そのものは非常に低いため、なかなか気付かれないが、結果は重大であるというような因果関係を、今までよりはるかに早く発見することができるようになるかもしれない。

———, "Cryptography and Data Security," Addison-Wesley Publishing Compony, 1982. (上園忠弘・小嶋　格・奥島晶子訳、『暗号とデータセキュリティ』、培風館、1988年)

Jonge, W., "Compromising Statistical Databases Responding to Queries about Means," *ACM Transaction on Database Systems*, Vol.8, No.1, 1983.

McLeish, M., "Further Results on the Security of Partitioned Dynamic Statistical Databases," *ACM Transaction on Database Systems*, Vol.14, No.1, 1989.

Reiss, S.P., "Practical Data-Swapping: The First Steps," *ACM Transaction on Database Systems*, Vol.9, No.1, 1984.

Schlörer, J., "Identification and Retrieval of Personal Records from a Statistical Data Bank," *Methods of Information Medical*, Vol.14, No.1, 1975.

———, "Disclosure from Statistical Databases: Quantitative Aspects of Trackers," *ACM Transaction on Database Systems*, Vol.5, No.4, 1980.

Schwartz, M.D., E.D. Denning, and P.J. Denning, "Linear Queries in Statistical Databases," *ACM Transaction on Database Systems*, Vol.4, No.2, 1979.

Sicherman, G.L., W. Jonge, and R.P.V. Riet, "Answering Queries without Revealing Secrets," *ACM Transaction on Database Systems*, Vol.8, No.1, 1983.

Traub, J.F., Y. Yemini, and H. Woźniakowski, "The Statistical Security of a Statistical Database," *ACM Transaction on Database Systems*, Vol.9, No.4, 1984.

Trueblood, R.P., H.R. Hartson, and J.J. Martin, "MULTISAFE-A Modular Multiprocessing Approach to Secure Database Management," *ACM Transaction on Database Systems*, Vol.8, No.3, 1983.

U.S. Department of Commerce, "Report on Statistical Disclosure and Disclosure-Avoidance Techniques," Washington,D.C.: *U.S. Government Printing Office*, 1978.