ミクロデータの公有化と利用の技術的課題

渋谷 政昭（高千穂商科大学）

要約

調査データを、調査主体の管理監督を離れた公共のものとするとき、被調査個体の秘密を守りながらデータの情報をできるだけ活用するための、理論と技術を要約する。特にコンピュータ技術の利用を考え、結びに大局的な環境について議論する。

1．まえがきと準備

ミクロデータを公有化するにあたって、どのような制度、機構が必要であるか、日本の現行政の困難がどこにあるか、すでに公有化している国々はどのように実現しているかについては、この報告書、シリーズ、の他の著者たちが議論している。さらに前進して、具体的な制度、機構についての提案が検討されている（永山貞則 1998）。

ここではより技術的な面をとりあげる。大きくわけて二つの部分がある。最初に、ミクロデータの公有化を可能とするための理論と技術。次に、制度、機構ができたとして、その運用、活用に必要な技術。これらについて要約し、議論し、最後に結論として政治的、文化的な側面について考察する。

まず準備として、ミクロデータの公有化について復習する（Willenborg and de Waal, 1996)。

(1)ミクロデータの副次的分析

統計調査のデータを公表するときに、多くの場合は、分割表、配列などの形式にまとめる。あるいは特定のモデルを前提として、そのパラメータの推定値と、モデルの正当性を裏付ける適合の尺度を公表する。しかし調査の原データ（生データ）、つまり分析、解析しやすいように整理してあるが未加工のデータ、には多くの情報が含まれており、調査を計画した人の意図から外れた目的に有用なことがある。調査には多くの人手と費用をかけるのであるから、当初の目的を果たしたからと、原データを死蔵し廃棄してしまうのはむだ使いである。何らかの形でこれを保存し、利用可能として公益に資するのが基本的な目的である。当初の目的以外の分析を副次的分析（secondary analysis）と呼ぶ人もいる。

収集した原データを何らかの形で公有化するとき、氏名、住所、電話番号など本人を同定（識別 identify）するデータは当然除かれ、匿名化される（"anonomized"）が、性別、

年齢、職業、収入などは重要な属性であり、除くことはできない。公有化の最大の困難は調査対象となった個人、企業など個体の秘密の漏洩、プライバシの侵害である。調査にあたっては、公表の限度を述べて調査への協力を依頼しており、それから著しく逸脱することはできない。特に、集計結果だけを公表し、個体に関する個別データは秘匿されることを約束しながら、何らかの経路で個体名が曝露され、個別データがもれることは許されない。調査主体、機関が秘匿についての信頼を失うと、その後の調査が非常に困難になる。以下では個体一般の代わりに、個人と呼ぶこともある。匿名化された生データのことを、ミクロデータ（micro data sets）と呼ぶ。

(2)漏洩制御、ミクロ統計

統計調査データの公有化に際して、調査データに含まれる、被調査者が他人に知られたくないデータ、を秘匿するのが漏洩制御（disclosure control）あるいはデータ保護（data protection）である。つまり、匿名化されていても属性を調べることにより、特異な、特徴ある個体が浮かび上がり、個体名が確認されてしまう（再同定、再識別 reidentification）可能性がある。これを避けるためのデータ処理であある。漏洩制御の方法を一言で言うと、特異な、特徴のあるデータを、一般的な、典型的なデータの中に紛れ込ませることである。いわゆる外れ値を「外れない範囲」に押え込む。

一般に公表されるのは、漏洩制御の処理をしたミクロデータで、これを簡単に「ミクロ統計」と総称しよう。安全で役に立つミクロ統計を作成することが緊急の課題である。

コンピューター技術では、データの安全性（security）という概念がある。これはおもにネットワーク、とくにインターネットにつながったコンピュータの上、あるいはインターネット送信中の、データの窃盗、破壊、改竄に対する安全策で、通信の管理と暗号の利用が主な対策法である。ここの議論でも、この意味の安全策が重要になるが、全体としてはネットワークから離れたオフラインでの安全である。

2．ミクロ統計構築の理論

(1)問題の定式化

統計的決定理論は、不確定な現象について意思決定し行動するときの一般理論である。観測実験に費用をかけてデータを増やせば、不確定現象についての情報が増加し、より合理的な決定、行動を行い期待損失を減らすことができる。あるいは、利得損失をより精しく予測できることを前提とする。このとき、観測実験の費用と、決定、行動の期待損失をバランスさせるのが、理論の大筋である。

漏洩制御もこれに類似した課題で、二種類の危険（期待損失）のバランスとみなすことができる。ひとつは原データがもっている情報の減少であり、もうひとつは秘密の漏洩に

よる損害である。後者は破滅的であるので、これを十分小さく、実質的に 0 とした上で、前者をできるだけ小さくしたい。竹村彰道（1998）はこのような枠組みを考え定式化を試みている。定式化はともかくとして、二種類の損失をどのように評価するかが重要な課題となる。

調査データについて一般的な情報量を測るにはどうすればよいか。たとえば分割表にまとめ、ロジスティック線形模型をあてはめ、その情報量を利用する。しかし、漏洩制御の処理をすると、通常は同種のモデルの範囲に入らず、損失を計量できない。利用目的と漏洩制御処理に応じて、必要とする統計量のバイアス、分散にどのような影響があるか、個別に調べることは可能である。

一般的評価をしようとすると、母集団についてのモデル、超母集団、を仮定し、漏洩制御の方式を定めた上で、諸種の統計量の平均二乗誤差などの分布を調べることになる。

(2)秘密漏洩の危険

もうひとつの秘密漏洩の期待損失についての測定は困難である。まず何が秘密であるか、時代とともに人々の感性が変わる。離婚暦、夫婦の年齢差について寛容でない人は減少し，秘密ではなくなっている。しかし社会的な合意があっても、当事者の立場、主観は多様である。調査に対する信頼性がからむため、訴訟を受け損害賠償をする費用では不十分である。識者の判断に委ねるしか方法はない。

病気、離婚、出産などの履歴、高額所得、保有財産、職業あるいは失業状態、などが、機微に触れる（sensitive）調査項目であろう。項目全体が機微に触れるよりは、項目のあるカテゴリーに入ることが秘密の場合が多い。このときにはカテゴリーの合併が簡単で有効である。大項目と小項目が重複してあるときには、小項目全体を除く必要も生じる。たとえば地域、職業などである。

(3)補助情報

漏洩の可能性を評価するとき、ミクロ統計を利用して個人の秘密を探る攻撃者が何を意図しどのような手段を使うか想定することが必要だという議論がある。しかし攻撃者の意図や行動を分類し整理して、いくつかのシナリオを書く試みは有効ではない。ちょうどテロリストの危険を想定するようなもので、下手なフィクションを考えると戯画的になってしまう。

肝心なことはむしろ攻撃者が、公有化するミクロ統計以外に、どのような補助情報をもっているかである。これによってまず攻撃者の行動が変わる。たまたま補助情報を得たことにより犯意を刺激されることもあるだろう。逆に明確な意図をもっていれば、役立つ情報を集めるだろう。諸団体の名簿、企業の人事データ、商業的な顧客名簿、など一部の人にアクセスが容易なデータ、データベースはいろいろある。

(4)マッチング

本質的なのはミクロ統計と補助情報との結合、マッチングである。諸データの結合(join, matching) は、ミクロ統計活用の大きな一部である。一対一の完全なマッチングではない統計的なマッチングも、利用法の一つとしてその有効性が研究されている（椿広計 1997）。皮肉な世の常として、ミクロ統計の活用法が、攻撃者の有力な道具となる。

攻撃者が補助情報をなにも持っていないときには、ミクロ統計からサンプリング対象の母集団全体についての漠然とした知識が、あえて補助情報と呼べるものである。つまりミクロ統計と、サンプリングの母集団とのマッチングである。このときの秘密漏洩の危険評価は後述の孤立個体の分布の議論に帰着する。

(5)標本、母集団での孤立個体

以下の議論ではミクロ統計が多次元分割表 (contingency table)であることを前提とする。定量的な変量は適当な級 (class、区間) に分けておけばよい。データセットとしては、カテゴリーを並べた一枚の表で、データ処理でデータファイル、と呼ぶものである。つまり、データセットは個人についての記録 (record) の集まりで、記録は各人の諸属性 (attributes) の一定数の集まりで任意ではあるが一定の順序に並んでいる。各属性を表す項目 (field) は分類変数 (categorical variable 多値論理変数) で、カテゴリーの種類は項目ごとに定まっている。

攻撃者が何の補助情報ももっておらず、ミクロ統計だけが利用可能であるとする。そのときには、いくつかのきわだった特徴をもつ個人を探すであろう。データセットの中で、誰とも違った、唯一の属性をもつ個人を孤立個体 (unique) と呼ぶ。つまり孤立個体の属性を他の人と比べると、どこかの項目で異なるカテゴリーがある。項目数が多いと、可能カテゴリーの組み合わせの数は非常に多くなり、データ数をはるかに超えることもある。その場合には孤立個体の数も多い。

攻撃者がミクロ統計から、興味ある孤立個体をひとり選び出したとして、それが母集団の中でも孤立していれば、匿名化したはずの個人が再同定される。母集団の中で孤立していなければ、個人攻撃ではなくなるし、同じ属性の個人すべてを探すことは困難となる。

(6)孤立個体数推定の不確定性

そこで「標本での孤立個体が母集団でも孤立している確率を推定する」という統計理論の課題が生じる。簡単のため、サンプリング法を単純非復元確率抽出に限っても良い。いくつかの論文が書かれているが、実はこれだけの前提では原理的に推定できないことを簡単に説明しよう。

$N = 10^4$ 人から $n = 5*10^j$ 人を単純確率抽出する。N 人中の $M = 10^j$ 人のグループからは平均 $n/M = 5$ 人が選ばれる。さて同じ属性をもった j 人の小グループ が S[j] 組あったとし(j=1, 2, ...)、計 j*S[j] 人をまとめてひとつのグループとみなす。S[j]が n/M よりもずっと大きいとすると、S[j]個の小グループからはほとんどひとりしか選ばれない。さ

て標本のなかに孤立個体が k = 12 人いたとして、母集団にだいたい K = k*N/n = 240 人の孤立個体がいることにはならない。j = 2 人の小グループが k/ j= 120 組でも、j = 3 人の小グループが K/j = 80 組、. . . ある場合でも、（あるいはその適当な組み合わせ）標本中の孤立個体数はほとんど変わらない。

つまり、「n 回の独立な試行では、確率が 1/n よりずっと小さな事象について、まともな推測はできない」という、一種の統計的不確定性原理がこの場合にも成り立つ。この不確定性により、母集団において「完全に同じ属性をもつ個人が j 人いる同属性グループの数 S[j], j=1, 2, ...」(これを頻度の頻度 frequency of frequencies と呼ぶ人がいる)について何か仮定をおく、あるいは超母集団モデルを導入しなければならない。

いわば統計的な補外（外挿 extrapolation）をすることになる。頻度の多い部分から少ない部分への補外であるから、成功するとは限らない。多くの事例について経験を積むしかないであろう。超母集団モデルとしては単一のパラメータの Ewens 分布、2パラメータの負の2項分布、あるいは線形ロジスティク模型に基づくものなどの提案がある（渋谷政昭 1997、佐井至道 1998、Hoshino and Takemura 1998）。

(7)ドイツの経験

攻撃者が現実的なファイルを補助情報としてもっていることを前提として、匿名化された被調査者をどの程度再同定できるか、という貴重な実験がドイツで行われた。これは 1996 年の国際シンポジウムで Heike Wirth さんが報告したものであるが、あえて紹介する (Mueller, Blien, and Wirth, 1995, Mueller and Wirth, 1997)。

ミクロ統計は北ライン・ウェストファリア州の１％世帯の全員 168,368 人についての社会経済調査である。攻撃者の利用するのは、ドイツの科学技術研究者についての年鑑 Kuerschners Dentscher Gelehrtenkalender 1987 で、ドイツの大学教授全員と、秀れた学術的出版のある計 7983 人の研究者を含んでいる。その属性の中で、ミクロ統計と対応するものは、地域、性、年齢など 10 項目で、可能なカテゴリー組み合わせ数は約 1313 億種類である。

この年鑑を基本として、

(1) 年鑑全部をミクロ統計と対応させる。

(2) 年鑑中の大学教員だけを用いることに限定し、ミクロ統計中で職業が大学教員であるものと対応させる。

(3) さらに、年鑑中の人々で、社会経済調査のサンプルに入った、53 人のリストをもっている。

という、三種の補助情報利用を仮想した。つまり（A）年鑑中の三つの部分集団を取り出した。

つぎに（B）各部分集団の各項目ごとのカテゴリーの種類を調べ、これらから外れるカテゴリーをもつ個人を除くことにより、ミクロ統計の中の候補者を縮小した。ただし (2) の

場合には、職業が大学教員であるもの全部を選んだ。

(C) 縮小したミクロ統計の中の候補者三種のなかで孤立個体を調べ上げた。

(D) これらの孤立個体をそれぞれ年鑑中の三つの部分集団の属性と比較し、一対一に対応する者を取り出した。

(E) このなかで、実際に同一人物を同定したかどうかを調べた。

これらの段階で、探索した人の数は、第1表のとおりである。(表の形式を原論文と少し変えた。)

第1表 補助情報と個人の同定

補助情報の種類	(1)	(2)	(3)
補助情報中の人数 (A)	7,983	?	53
縮小したミクロ統計中の候補者数 (B)	3,099	197	151
孤立個体数 (C)	2,467	147	121
一対一対応 (D)	14	11	9
実際の対応 (E)	4	4	9

第三の補助情報は、現実には起こりにくいが強力である。ただし一対一対応数が多くないのは、項目とカテゴリーの定義が二つのデーターセットの間で違うためである。第一第二の場合では、かなり手間を要し、しかも最終的には一対一対応の中の半数以下しか再同定できていないという、楽観できる結果であった。実験結果が示すことのひとつは、標本、母集団の両者で孤立している個体の数を、標本から推定した結果ほど、悲観的ではないことで、理由は補助情報とミクロ統計が完全にはマッチしないためであった。

これまでの議論でも、実験の説明でも、再同定の危険を問題にした。しかし一対多のマッチングも含め、所得などの定量的な推定をすると、漏洩がありうる。この可能性は、指摘されてはいるがより小さい、より解決困難な、危険とみなされ、研究されていない。

(8)決断の過程

以上のように漏洩制御の理論と技術は未成熟である。より魅力的なものが現れたとしても、二つの危険のバランスという本質は変わらないし、危険の評価に灰色領域が多いことも変わらない。漏洩制御は日常の政治的決定の過程である。決の孔が小さい当事者では決断が遅く、研究者の不利益と悩みを解消できない。

3．公有データの運営管理

公有化のための制度を作ると同時に運営を組織化しなければならない。目標は、一つの情報システムとして、経済的効率的したがって利用者にとって使いやすいサービスを提供することである。具体的には、技術面でデータベース、ネットワーク通信、そして要員であるデータアーキビストの準備である。

(1)データベース、データウェアハウス

企業、行政組織が保有するデータベースは主として日常業務のためである。営業、生産、経理、人事などすべての業務の処理のためにデータベースが必要である。一方、企業は業務全体を見渡し、大きな取り引きを行い、長期の計画を立てるための情報源として、異なるデータベースを必要とする。日常業務では、日々の、時々刻々の現状を必要とする。取り引き、業務、ビジネスケースが終了し、毎日アクティブ（active）となっていたファイルがアーカイブ（archive）に移される。情報源となるのはアーカイブの中の、長期にわたる、諸業務部門を整理、要約、統括したデータであり、データウェアハウス（data warehouse　データの倉庫、上屋（うえや））と呼ばれている。

データウェアハウスの必要性は 1970 年代から意識されていたが、1980 年代になって導入が始まり、1980 年代末になって Bill Inmon がデータウェアハウスの言葉を導入した。それまでは大容量のデータを蓄えるのにはオープンリールの磁気テープしか利用できなかったが、この頃から磁気カートリッジや大容量のディスクが登場し始めたのが推進力となった。（Inmon, 1992, Commun. ACM 1998）

２種類のデータベースの特長は第２表の通りである。

終りの 2 行は、2 種類の仕事の違いを表わすための業界用語（jargon)である。 二つの列の違いのために、両者は別のコンピュータシステム（しかも異種のコンピュータシステム）の上で稼働していることが多い。

大規模なデータウェアハウスが多くの企業で構築され、データウェアハウスの設計、構築、運用のためのハードウェア、ソフトウェア、コンサルティングが、コンピュータ業界の中で大きな割合を占めている。

アーカイブに蓄えられるようなデータを統合するためには、大規模なデータモデルの構築、諸フィールド項目の定義、記述や命名の標準化、データの物理的な構造の再編成、ユーザーインタフェイスの開発、などの準備が必要である。

データウェアハウスの内容が更新されることはないが、時間とともに成長し、拡大する。ある時間間隔のデータを日報、月報、年報などの形にまとめ、必要に応じて取り出しやすいように再編成しなければならない。つまり横断的（sectional）に見たデータを経時的(longitudinal）に眺めることになる。また同時点で生じ、個別に記録していた複数個の量を、結合することになる。これを人間の操作で行うことは難しく、定期的に動き、自動

第2表 業務処理用と分析処理用のデータベース

業務用データベース	データウェアハウス
適用業務向き	課題向き
個別的	統合的
絶えず更新	更新されない
現時点のデータ	過去のデータ
利用の時間が限定	利用時間が不定
オンライン業務処理 OLTP	オンライン分析処理 OLAP
on line transaction processing	online analytical processing

的に処理をするソフトウェアが必要となり、開発され、商品化されている。

情報源としてのデータウェアハウスを構築することとそれを活用することとは別の話である。企業内諸部門の計画に利用され、特に上層部の意識決定に利用されて利益をもたらすのでなければ、構築の意味はない。有効に活用するかどうかは、データウェアハウスのデータが網羅的で良質的であるかによるし、意識決定にあたってどれほど客観的に量的に思考できるかにもよる。これは高度の知識の問題であるが、人間の判断を助けるための道具はいろいろ工夫されている。統計学、オペレーションズリサーチの伝統的な方法に加えて、人工知能 (artificial intelligence , AI) 研究の人たちが、大量データから特異な意味のあるデータを取り出す方法を模索している。関係者達は、これをデータマイニング (data mining データの採鉱)、 KDD (Knowledge Discovery and Data Mining) などの言葉で飾っている (Bigns, 1996, Commun. ACM 1996)。

一言にまとめるとデータマイニングは、探索的なデータ解析の新段階である。新しい点はデータ量の飛躍的な増加であり、それを扱うためのアルゴリズムが開発されている。もう一点は、必ずしも集団の統計的特徴ではなく、孤立個体を見出すことに興味をもっていることである。まさしくミクロ統計攻撃者の理論である。

統計調査データは、元々静的で構造が比較的簡単であるが、たとえば長期にわたる国際間比較、他の調査との比較のためには、データハウス構築以上の仕事が必要であろう。

(2)データの記述

データアーカイブ、ミクロ統計の中でも、データウェアハウスの中でも、常に登場するのがデータの意味の説明である。データベースシステムでは、各変量のデータタイプのリストをメタデータ (meta data) と呼んでいる。どちらかというと、人間よりは機械のための記述である。統計データベースでは、コードブック (code book) が人間と機械、双方のための記述である。さらに、調査対象の母集団、サンプリング法、調査用紙、調査時期など全般にわたること、調査項目、カテゴリーごとの定義、解釈、例外、など記述すべきことは多くそれを理解する利用者の負担は大きい。

負担を軽減するために、また国際間の比較などのために、データの記述を標準化しようとする提案は古くからある。たとえば1996年国際シンポジウムで紹介されたData Document Initiative、ICSSD (International Committee for Social Science Information and Documentation) などの活動がある (de Vries , 1997) 。しかしなかなか収束せず、標準の提案があまりに多い状態である。標準化が困難であれば、データごとに、データといっしょに、記述を貯えるしかない。柴田里程 (1994) は応用統計データについてのデータとその記述を形式化する (D&D, data and description) を提案した。現在のように、一ヶ所に情報を集中するのではなく、分散したものをネットワークでつなぐ方式では、当然メタデータなどの標準化方式では対応できない。

(3)ネットワーク通信

研究者がミクロ統計を利用するとき、どのような手続きでどのように作業するか。現在日本のシステムは、多くの人の多くの時間を要することになっている (松井博 1998) 。効率化のためには、当然インターネットを通じたコミュニケーションを採用することになる。

多量のデータを公有して多くの研究者にたいして提供している活発な機関がいくつかあり、その情報システムが参考になる。

ここでは代表的な機関のひとつであるルクセンブルグ所得研究組織 (Luxembourg Income Study, LIS) について紹介する。詳細は、同研究所の Timothy M. Smeeding さんが 1996 年国際シンポジウムで発表している (Smeeding, 1997) 。また LIS のホームページ http://lissy.ceps.lu が参考となる。

(4)ルクセンブルグ所得研究組織

運営を完全にコンピュータ化している。それによって毎日 24 時間、週 7 日のサービスをし、毎日平均全世界の数十人の人の 200 件以上の仕事を処理している。利用に際しては事前に登録し、許可を得なければならないが、申請、許可の手続きもインターネットにより電子メールで手続きをし、利用法についての文書も電子メールで受け取る。データベースにアクセスし解析を行うには SPSS または SAS でプログラムを書く。 LIS システム用に若干の追加、変更を行うが、大部分はコメントカードの形式の記述である。これを電子メールで送ると、パーソナルコンピュータのネットワークがインターフェイスとして受けつけ、処理をし、結果の統計量、表、などを電子メールで送り返す。すべては自動的である。

漏洩制御のために、SPSS、SAS および LIS 開発ソフトを使用することに制限し、その入出力ステートメントを自動的にチェックする。他の使用法をいっさい許さない。もちろんすべてのデータは LIS に導入する前に匿名にする。所得額はマッチングを防ぐために数値をある桁に四捨五入する。結果の出力は登録利用者の電子メールアドレスにだけ送られる。すべてのジョブの請求と、計算結果をレビューする。登録していない人のアクセスを禁止するため、ジョブの要請を仮想コンピュータが受付ける。ミクロデータの入っているデータベースには、LISの要員が仮想コンピュータを使ってアクセスすることだけができる。

(5)他の方式

このような方式は、ミクロ統計の内容を限定し、全世界に公開するために優れた方法である。しかしユーザの探索的なデータ解析のためには、対話型のシステムを用い、グラフィカルな出力を見ながら、より柔軟にモデルを選択できることが望ましい。

対話型利用のために、種々の自動チェック機構、暗号システムを導入することは重要な課題である。実用的なものが利用可能となるまでは、物理的に閉鎖された環境で仕事を行うこともやむを得ない。それでも中心設備と専用回線で継がれた場所で、テレビ電話で連絡しながら作業できる設備を作ることは難しいことではない。

表形式の統計公表、および簡単な構造のミクロデータのにたいする漏洩制御ソフトウェア開発の努力がオランダで進められている (Hundepool and Willenborg, 1997)。ミクロ統計の内容が多岐に渡れば、いろいろな手続きを自動化することは困難となり、別の方式の処理が経済的かもしれない。すべての分析結果について、たとえ見込み違いで新しい知見が得られなくても報告書を提出してもらい、査読、出版することが望ましい。そのような分析そのものを業績として評価する習慣を作らねばならない。

(6)アーキビストとデータ文化

1997 年国際シンポジウムでもミクロデータ公有化のために活動している各国人々の報告があった。特に、二人のイギリス女性教授、Denise Lievesley, U. Essex と Angela Dale, U. Manchester の講演は刺激的であった (参照、森博美 1997)。理由は、二人は「データアーキビスト (data archivist)」 という職務、専門に誇りを持ち、30 年近くの伝統を持つ組織の代表者として、その政策を説明し困難を示し、伝道者のように 愛他心 (altruism) について語ったからである。

アーカイブという言葉は、コンピュター技術で多分 1970 年代から使われていた。Washington D. C. の the National Archives のことは、いろいろな機会に、紹介されてきた。しかしアーキビストという役割について議論され始めたのは、情報公開制度が地方自治体にでき始め、古文書館が創立されてからである。稗田阿礼、藤原定家、の仕事である。海外のデータアーカイブの話が紹介されながら、アーキビストの役割について十分に強調されていないようである。A. Dale さんは講演の中で「データ文化」について語った。なるほど、新しい文化を作り出し維持するという気構えで、はじめて成り立つ専門職業である。

「データ科学」という言葉を最近、林知己夫さんと柴田里程さんが、異なった視点から唱えている。理工学部とくに実験観測系の研究室では相手にされないスローガンであるが、文系の世界では量的思考の重要性を強調しなければならない。データを利用し始めれば直ちに、データの質の良否が問われ、データを収集管理する公共機関の重要性が痛感される。データ文化はデータ科学よりも規模が大きく、データを作る人たちと、それを利用したい人との緊張した間柄、その媒介となる第三者機関の愛他的努力、から作り出された構造と

人たちである。

4．データ公有化の環境

(1)統計の真実性

統計法は「統計の真実性を確保する」ことを第一の目的としている。第二次世界大戦中に統計が軍事機密として隠され、歪曲され、国民に正しく伝えられなかったことが、戦局の無謀な拡大、破局の一原因であった。この法律を制定し、統計機構を組織された人々は、過去にたいする強い反省に基いてこれを強調した、と了解している。これはまた、暗黙に「個体データの真実性」と対比し、必要前提としている。別に明記した、調査対象である個体の秘密の保護、調査票を統計上の目的以外に使用することの禁止、はその具体化である。

しかし真実性の概念を「真、善、美」のような理念よりも具体的な水準で使うとすれば、意味することを議論しておくべきであろう。たとえば、統計法の下で継続し公表されてきたこれまでの統計が真実、などという解釈があっては困る。公表されているのは調査されたデータを集計した一側面であって、収集されたデータの中には多くの統計的事実が記録され、潜んでいる。大規模調査のデータは公表されている型の要約以外の要約、集約が可能である。

調査票を真実の個体データであるかのように語るのも困る。非回答の偏りを正し、記入誤りを調整する仕事を正当に評価すべきである。調査対象の活動・生活様式、調査組織の人員構成、記録・通信・計数・機器の能力、などの環境変化によって非抽出誤差は変化する。それを測定するには、詳細なデータ—の慎重な分析を必要とする。行政記録との比較なども双方の品質管理の手段として役立つであろう。その結果は調査の再設計に生かされる。日本の統計の質が高いという国際評価があるようだが、コストと比較した品質であろうか。真実性が無謬性にまで到って、現実の泥臭さにふたをしてはならない。

イギリス国勢調査のミクロ統計は、職業病の疫学調査から始まった。生活環境、生産、経済の急速な変化に対応する統計の必要は絶えず指摘されている。統治するための国勢学の伝統の、調査するもの、されるもの、の二分法では、互いに規制を課すだけとなる。すべての人が豊かになるために統計を作り利用することが当然の目標である。集団に対する事実であれば、どのような統計量でも公表する努力が必要であろう。それによって、統計の真実性、信頼性も改善される。

(2)研究者の倫理

集団を分析しているとき特異な個体が研究者の目に入ることは可能である。たとえば多重回帰における外れ値である。研究者は極端な外れ値の特性を調べた上で、分析のために

残すか、除去するかを決定する。外れ値を調べることにより新しい発見もあるだあろう。もちろん、そのときのプロジェクトから外れた解析を直ちに実行、発表することはできないだろうが、中間結果を探索する自由まで制限すると、新しい研究が妨げられる。

研究者が好奇心をもち過ぎて漏洩に到るシナリオを持ち出す人もいるだろうが、研究者は自ら倫理綱領を作り、研究の自由を広げなければならない。データウェアハウスの機能が高く整備されていれば、個体データの探索はより容易である。だからといって、個人が誓約した上でデータに触れ、しかも対話的に分析する機会をはばむ理由にはならない。

参考

佐井至道（1998）個票データにおける個体数とセル数との関係について、応用統計学（印刷中）

柴田里程（1994）“データ解析の電子ジャーナル（EJDA）”の実働化、統計数理研究所、共同研究リポート 54

渋谷政昭（1997）多項分布における度数 0,1 のセルの数―漏洩管理のための基礎事実―、応用統計学、161-170

竹村彰道（1997）個票データ開示の理論、松田芳郎、平成 8 年度科学研究費補助金（重点領域）報告書（課題番号 08209102）1-25

椿広計（1998）統計的マッチングの意義と問題点、美添泰人、補助金（重点領域）報告書（課題番号 08209102）27-44

永山貞則（1998）ミクロデータの提供への戦略、松田芳郎、平成 9 年度科学研究費補助金（重点領域）報告書（課題番号 08209102）6-9

松井博（1998）官庁統計ミクロデータの利用の経験と今後の課題、松田芳郎、平成 9 年度科学研究費補助金（重点領域）報告書（課題番号 08209102）10-25

森博美（1997）国際シンポジウムから：イギリスにおけるサーベイミクロデータの提供 －報告と討論より－、Newsletter: Facts from Data No.6. (altruism, data culture, fight などの言葉は、日本での講演用に用意されたスライドで表示され、話されたもので、論文にはない。)

Bigns, J. P. (1996) Data Mining with Neural Networks, McGraw Hill, （社会調査研究所、日本アイ・ビー・エム 共訳、データマイニング、日経 BP 社）

Communications of the ACM (1996 Nov.) vol.39 No.11, U. M. Fayyad and R. Uthurusamy, Data Mining and Knowledge Discovery in Databases （特集記事）

Communications of the ACM (1998 Sept.) vol.41 No.9, A Sen and V. S. Jacob (eds.) Industrial-Strength Data Warehousing （特集記事）

de Vries, R. (1997) Standardization of data documentation for the transaction of data

files, in Matsuda, Y. (ed.) Exploring New Frontiers in Statistical Analysis Using Microdata Sets, First Summary Report, 343-354

Hoshino and Takemura (1998) On the relation between logarithmic series model and other super population models useful for microdata disclosure risk assessment, J. Japan Statist Soc. (to appear)

Hundepool, A. J., and Willenborg, L. C. R. J. (1997) The ARGUS twins: Software for statistical disclosure control, Bulletin of the International Statistical Institute, 51st session, Istanbul, IP 48, Book 2, 21-25.

Inmon, W. (1992) Building the Data Warehouse. Wiley, New York. (藤本康秀、小畑喜一 監訳、データウェアハウス構築編 オーム社 (1997) 活用編、運用編もある。)

Smeeding, T.M., Coder, J., and Vleminckx, K. (1998) Practical solution for public use microdata sets outside government statistical offices: The experience of the Luxenbourg Income Study (LIS) , in Matsuda, Y. (ed.) Exploring New Frontiers in Statistical Analysis Using Microdata Sets, Second Summary Report.

Mueller, W. Blien, U., and Wirth, H. (1995) Identification risks of microdata: Evidence from experimental studies, Sociological Methods and Research, 24, 131-157.

Mueller, W. and Wirth, H. (1997) Confidentiability and disclosure of micro data sets obtained from statistical survey, in Matsuda, Y. (ed.) Exploring New Frontiers in Statistical Analysis Using Microdata Sets, First Summary Report, 355-373. (松田芳郎、平成 9 年度科学研究費補助金 (重点領域) 資料 A02 NO.4,日本統計研究所 (課題番号 08209102) に伊藤伸介の翻訳がある。なお上記の論文を参照のこと)

Willenborg, L. and de Waal, T. (1996) Statistical Disclosure Control in Practice, Lecture Notes in Statistics, vol. 111, Springer, New York (渋谷政昭 (1997) 文献紹介、Newsletter: Facts from Data, No.4)