参考1

統計数理(2003)
第51巻第2号 241–260

特集「個票開示問題の統計理論」
[総合報告]

個票開示問題の研究の現状と課題

竹村 彰通†

(受付 2003年2月6日；改訂 2003年9月5日)

要　旨

本稿では，個票開示問題の理論に関して基本的事項を解説するとともに，個票開示問題に関する統計数理的な理論研究の現状と課題についてサーベイを与える．国際的な研究の流れを紹介するが，その中でも筆者自身および関連の研究者の研究成果に重点をおいている．個票開示問題には安全性と有用性のトレードオフという難しい側面があり，十分な技術的な理解に基づいて冷静な議論をおこなう必要がある．

キーワード：キー変数，母集団一意，大域的再符号化，局所再符号化．

1. 個票開示問題の背景と統計数理的研究の意義

ここではまず個票開示問題の背景について説明し，この問題の統計数理的研究の位置づけについて述べる．

計算機やネットワークの発達とともに，個人でも大量のデータを扱うことが可能になってきている．このような中で統計表に集計される以前の個票データは，新たな統計データの利用法をうながすものである．しかしながら，個票データに含まれる回答者が識別される危険性を考慮すると，個票データの提供においては安全性と有用性の間のトレードオフに留意する必要がある．ただし，本稿で個票データという場合には，住所や電話番号といった直接の識別情報は削除していることが前提であり，識別といっても変数の組み合わせによる間接的な識別のみを問題としている．また以下では個票データの提供・開示にともなって回答者が識別される危険性を，開示リスク(あるいは識別リスク，漏洩リスク)という．

安全性と有用性のトレードオフの問題は，本稿で論じるように理論的にも十分な研究が必要とされる問題である．個票開示問題を議論する際には，十分な技術的理解に基づいて冷静な議論をおこなうことが重要である．個票データの安全な提供と，提供されたデータの利用方法の問題は統計科学における新たな研究課題として魅力あるものである．個票開示問題の解決には，統計理論，統計制度，官庁統計実務，計算機統計等の分野にまたがる多面的な接近が必要であり，その意味でも統計科学全般にかかわる課題である．

我が国では制度的な対応がなされていないため官庁統計に関する個票データが極めて制限された形でしか提供されていないが，欧米では個票データの開示がかなり進んでおり，実際の開示リスクに関する経験も蓄積されつつある．開示の範囲や形式にもさまざまなものがあり，アメリカ合衆国における一般公開用個票データ “Public Use Microdata Sample” (PUMS)のようにcd-rom等によって一般に公開される形態や，学術研究その他の目的で利用者を特定・限定して提供する形態などがある．このような様々な形態自身が，安全性と有用性のトレードオフ

† 東京大学大学院 情報理工学系研究科：〒113-0033 東京都文京区本郷 7-3-1

のバランスのしかたに対応していると考えることができる．欧米諸国の統計制度と個票開示の現状については松田 他(2000)に詳しい記述がある．また直接各国の統計当局のホームページより，現在の情報を知ることができる．

官庁統計の他にも，欧米では大規模な研究プロジェクト等にともなう，官庁以外の諸組織の行った社会調査の個票データが提供されている．欧米ではこれらのデータはデータアーカイブとよばれる組織によって管理・提供されている．我が国では，1970年代の宍戸駿太郎を研究代表とするデータバンクとその後の三宅プロジェクト等があったが，恒久的な組織としては1996年より東京大学社会科学研究所内に日本社会研究情報センターが設立され，データアーカイブとしての活動を始めた．しかしながら，活動を始めてからの期間も短く，欧米に比較してこの面でもかなり立ち遅れているのが実情である．データアーカイブの活動と意義については佐藤他(2000)を参照されたい．わが国でも，個票レベルでの官庁統計データおよび社会調査データの整備と提供が早急に望まれるところである．

最近では個票開示問題に関する書籍もいくつか出版されている．Willenborg と de Waal による2冊のモノグラフ(Willenborg and de Waal (1996, 2001))はこの分野の様々な概念を紹介しており分野の概観のために有用である．最近出版された2冊の論文集(Domingo-Ferrer (2002), Doyle et al. (2001))は現在の研究の動向を示している．ただしこれらの書籍では，政府の統計作成当局の実務的な観点からの記述も多く，理論的には必ずしも十分とはいえない部分もある．個票データを提供している統計当局にとっては，個票開示問題はすぐれて実務的な問題であり，必ずしもその数理的な構造をあきらかにすることが目的でない．しかしながら，統計数理の観点からみると，個票開示問題はまだ十分に理論的に整備されていない部分も多く，より多くの統計理論家の参加が望まれる．

本稿の目的は統計数理の観点からの個票開示問題研究のサーベイである．しかしながら，数理的側面は個票開示問題の一部にすぎないので，この点についてあらかじめ確認しておこう．理論的な研究も問題の全体像の中の方向性を見すえながらおこなう必要があるからである．個票開示問題は制度的な側面，統計実務的な側面，計算機への実装の側面，などさまざまな側面がある．特に個票データの開示リスクの実際の評価においては，仮に統計数理的なモデルが整備されたとしても，モデルで仮定される多くの母数の値は未知であり，その意味では主観的なリスク評価を避けることができない．例えば，侵入者(攻撃者，個票データに含まれる個人情報を知ろうとする者)がどの程度の外部情報を入手し得るかについての想定によって，開示リスクの見積りは大幅に変化する．また侵入者の攻撃の動機についても，直接的な利益を目的とするのではなく，攻撃そのものを楽しむような場合もあり得るかも知れない．さらにプライバシーや企業秘密に関する社会的な通念などの文化的背景なども数理的な考察ではとらえきれない．このように，開示リスクの評価においては，さまざまな側面を総合的に判断する必要があり，安全性と有用性のバランスをとりつつ回答者と利用者の双方に信頼される形で個票データを提供しなければならない．

以上のように個票開示問題においては数理的なアプローチはその一部にすぎないが，一方で数理的な手法は開示リスクの評価において，数値的かつ客観的な評価(ものさし)を与え得ることが大きな利点である．総合的判断が不可欠な中で，総合的判断が単なる主観的な判断となってしまわないためには，判断の基礎となる客観的かつ数値的なリスク評価の信頼性を高めることが，一層基本的な重要性を持つのである．これは統計的手法の諸分野への応用を含めた統計科学全体の中で，統計数理の果している役割と同様である．筆者の立場は，個票開示問題における統計数理的方法は，限定的ではあるが基本的な重要性を持っているというものである．個票開示問題における数理的方法の意義と役割については第9節においてさらに議論する．

以下，本稿では個票開示問題の理論的側面にしぼって概説をおこなう．邦文でのまとまった

解説はほとんどないので，技術的な細部に立ち入ることをせず，研究の流れを解説することに重点をおく．なお松田 他(2000)の 2.1 節(pp. 145–167)に渋谷による数理的側面の導入的な紹介が与えられているので，それも参考にされたい．また以下では，統計利用者にとって関心のある個人や世帯を対象とする統計調査の場合について個票開示問題の理論を説明する．特に用語の簡便のために個人を念頭におき，標本中の n 人の個体などという．これは事業所・企業を対象とする統計調査に関しては，企業秘密や企業の開示説明義務などの複雑な問題があるからである．

2. 個票開示問題に関する用語と諸概念

ここでは個票開示問題における基本的な用語と概念について整理する．

個票データの中の個体が誰であるかがわかってしまうことを個体の**識別**とよび，個体の識別の起こる危険性を開示リスク(識別リスク，漏洩リスク)とよぶ．統計調査においては回答者はたくさんの項目(変数)について解答する．名前，住所，電話番号など，回答者を直接特定できる項目を**直接識別子**とよぶが，個票データと言った場合には，直接識別子は含めないことが前提である．この意味では，個票データといっても直接識別子を用いた識別が問題となるわけではない．

開示リスクの問題は，間接的な情報の組合せによっても識別が起こり得る点にある．回答者が回答する項目の中には，性別や年齢など個人の基本的な属性を表す変数(フェイスシート項目)と，例えば所得や就業状態など調査の目的となっている変数がある．性別や年齢などの変数はわざわざ調査をしなくてもある程度容易に知ることのできる変数であり，したがって個票の回答結果に接しなくてもこれらを個人を特定するために用いることができる．そのなかでも，性別や年齢のように間接的に個人の識別に用いることのできる変数を**キー変数**とよぶ．これに対して調査の直接の目的をなす属性の中には，回答者のプライバシーに関するものなど，他人にむやみに知られたくない属性も多い．このような変数を，回答者にとってはセンシティブであるとして，**センシティブ変数**とよぶことにする．キー変数値の組合せによって個体が識別された時に，その結果としてその個体のセンシティブ変数の値が知られてしまうことが問題である．

例えば，筆者自身の個人情報について考えてみると，年齢 51 歳，大学教員，研究分野統計学，東京在住，身長 163cm である．これらの情報で筆者を特定できることはほぼ明らかであろう．このように，個体を特定できるような属性の組合せを，Willenborg and de Waal (2001)の 2.10.2 節に従って “fingerprint” (指紋)とよぶことにしよう．すなわち(年齢 51 歳，大学教員，研究分野統計学，東京在住，身長 163cm)が筆者の指紋である．キー変数のそれぞれはある程度容易に知ることのできるものであり，その意味ではキー変数による個体の識別自体がそれほど大きな問題というわけではなく，それよりも識別にともなってセンシティブ変数の値が知られてしまうことが問題であると考えられる．また，センシティブ変数といってもプライバシー上社会通念としてそれほど問題がない場合もあるので，その場合には個体の識別自体は実際上それほど大きな問題とならないかもしれない．しかしながら本稿では個体が識別されること自体をリスクと考えて，その前提のもとで議論をおこなう．

なお，キー変数とセンシティブ変数の区別は絶対的なものではなく，例えば年齢は女性にとってはセンシティブ変数であるとみなされることも多いことに注意しよう．以下では，キー変数が指定されたとして，キー変数値の組合せによって個体が識別されるリスクを考察する．

個体の識別に関して基本的な概念は**母集団一意**(population unique)の概念である．上で述べたように個体の識別はその個体の指紋によって可能となるが，それは母集団において同じ指紋

を持つ他の個体がいないためである．このように，キー変数の組合せについて母集団で個体が一意に定まるとき，その個体を母集団一意とよぶ．あるいは母集団で孤立した個体とよぶ．もし同じ組合せを持つ個体が2人いる場合には母集団2意とよぶ．同様に k 人いる場合には母集団 k 意とよぶ．開示リスクの評価においては主に母集団一意のみを問題とするが，母集団2意等についてある程度のリスクを考慮することは必要である．

母集団一意と並行的な概念として**標本一意**(sample unique)の概念がある．通常の統計調査は母集団の一部を抽出する標本調査である．以下では母集団の大きさを N，標本の大きさを n と書く．標本調査の個票データに含まれる個体が，キー変数の組合せについて標本中で一意である時に標本一意とよぶ．標本2意，3意等も同様に定義される．ここで標本一意と母集団一意の関係について考えよう．母集団一意の個体がもし統計調査で抽出されたとすると，その個体は必然的に標本一意となる．しかし逆は成り立たない．すなわち，標本一意であっても母集団一意とは限らない．通常の標本調査では抽出率 n/N は 1/1000 や 1/10000 といった低い率である．この場合，ある個体が例えば母集団で 100 意であるとしても，この個体が抽出された場合には標本一意となる可能性が高い．従って開示リスクの評価において問題となるのは，標本においても母集団においても一意となる個体である．標本でも母集団でも一意となる個体数の推定が開示リスクの基本であるが，この具体的な推定方法については次節で述べる．

一意数の推定等により開示リスクが大きいと判断された個票データについては，**秘匿措置**を施す必要がある．秘匿措置のもっとも基本的な手法は**大域的再符号化**(global recoding)である．re-coding とは，コードを付け直すの意味である．coding は格付けと訳すこともあるが，日本の官庁用語の符号化をそのまま用いておく．大域的再符号化とは，それまで年齢を5歳刻みで表示していたのを，より粗く 10 歳刻みで表示するなど，個票データに含まれる量的な変数の区間を粗くしたり，質的変数のカテゴリーを統合したりする操作をいう．大域的再符号化によって，一意数等は減少し，より安全な個票データを作ることができる．特に，高額所得者のように，極端な値による識別を避けるために，一定値以上を直接表示しない(すなわち無限までの片側区間で表示する)ことを **top coding** とよぶ．またある変数についてカテゴリーをすべて統合してしまえば，その変数についての情報はなくなってしまうから，これはその変数を削除することと同値である．このように top coding や，変数の削除も大域的再符号化の特殊な場合と考えることができる．

大域的再符号化はある変数について個票データ全体で一律にカテゴリーを粗くするが，これに対して特定の個体についてのみカテゴリーを粗くすることを**局所再符号化**(local recoding)とよぶ．その他のより特殊な秘匿方法を含めて，個体ごとに異なる秘匿措置を講じることを**局所秘匿措置**とよんでいる．局所再符号化の中でも，特定の個体の特定の変数の値に×をつけて欠測値とする秘匿措置を**局所欠測化**(local suppression)とよぶ．特にある個体についてすべての変数を欠測としてしまえば，これはその個体を削除することと同値である．個票データの中から一定の割合の個体を抽出して得られたものを**サブサンプル**(リサンプル)とよぶ．サブサンプルは，バイアスが生じない形でサブサンプルがとられる限りにおいて，ユーザの立場から見ると，最初から標本の大きさ n のより小さい調査がおこなわれたことと同じであることに注意する．

局所再符号化や欠測化は変数の値が粗くなるだけで，データを実際とは異なった値で表示しているわけではない．このような局所秘匿措置を非撹乱的措置とよぶ．これに対して，実際とは異なった値を表示することによって秘匿をおこなう方法を**撹乱的**(perturbative)な秘匿措置とよぶ．撹乱的な秘匿措置の代表的なものは，変数の値を異なる個体間で入れ換えるもので，これを**スワッピング**(swapping, Dalenius and Reiss (1982))とよんでいる．スワッピングの利点は，その変数の1変量の分布を変えない点にある．この他にも Gouweleeuw et al. (1998)によって提唱された **PRAM** (Post RAndomization Method)は，標本調査における randomized response

の手法に対応しており，局所秘匿の方法として有用である．これらの方法については第6節で詳しく説明する．また量的な変数の撹乱的秘匿措置としては，誤差を加える方法も考えられる(Fuller (1993))．なお，撹乱的な秘匿措置は，局所欠測化の後に欠側値の補完(imputation)の方法によって欠側値を補っていると解釈することもでき，スワッピングは欠側値の補完における Hot Deck 法に対応するものである．

キー変数のなかでおそらく最も識別に関して問題となるのは地域に関する情報である．地域が狭い範囲に限定されれば，個体を探すことが容易になり，個体識別のリスクが大きくなると考えられる．これに関連して，標本調査の抽出率の問題があげられる．層別抽出においては，層ごとの特性を考慮して抽出率を層ごとに異なったものとすることが多い．また層の設定においては地域情報を考慮することが多い．例えば日本における継続的標本調査の場合には，都道府県で層化して都市は県庁所在地としたり，東京の抽出率を小さくするといった方法をとるために，抽出率を公表することが地域情報を開示することにつながる場合が多い．抽出率が異なるデータのまじった個票データは，いわば母集団をゆがめて反映させたものであり，その分析においては抽出率の違いを反映した解析方法を用いる必要がある．このような場合には，個票データからのサブサンプルを，各個体を抽出率に逆比例する形で抽出すれば，母集団からの無作為標本と見なせる個票データが作成されることに注意する．層ごとの抽出率が細かく異なり，かつ広い範囲にわたる場合には，抽出率をグループ化して考えればよいであろう．いずれにしても抽出率の異なる個体がまざっている個票データの秘匿および利用に際しては，抽出率の扱いを考慮にいれる必要がある．

以上では開示リスクとして，個体の識別リスクのみを考えてきた．これに対して開示リスクをより広くとらえる立場もあり得る．特に**予測開示**とよばれる概念が典型的なものであり，個体が識別されなくても，その個体のセンシティブ変数の値が狭い範囲でわかってしまうことをリスクととらえる立場である．例えば，個票データの中で特定地域の特定職業の人達の所得が一様に高く狭い範囲に分布していたとすると，その地域のその職業の人の所得は狭い範囲で予測されてしまうというものである．ただしこのような開示は，分布の情報からの予測であり，統計情報が本来有している情報であるとも考えられる．従って予測開示は必ずしも個票データに特有のリスクの問題ではない．現状では，予測開示の理論的扱いについては不十分なので，以下では識別開示に限って議論することとする．

3. 個票データの開示リスクの評価

前節でも述べたように，開示リスクの評価における基本的な問題は母集団一意数の推定問題である．ここではまず母集団一意数推定を議論するための基本的な枠組を説明し，その後さまざまなモデルについて概観する．本節の内容に関するより詳しい解説は本特集の渋谷論文に与えられているが，ここでも概略を説明することとする．母集団一意数推定の理論については，最近になって我が国の研究者を含め急速な進展が見られた．これらの進展については本特集のいくつかの論文で紹介されているので，詳しい内容については他の論文にゆずり，ここでは概観のみを与える．また，この節における技術的な内容は，集団遺伝学，統計的生態学，計量言語学，等で用いられてきた方法と重複するところが多いので，それらの分野との関連についてもふれることとする．

まず開示リスクの評価に関して，個票データを分割表ととらえる見方について説明する．キー変数のなかには性別のようにそもそも質的な変数もあるが，年齢のように量的な変数もある．しかしながら，開示リスクの問題を考える際には，連続な変数も一定の区間に区切って離散化して考えるのがよい．例えば年齢を用いて他人を識別しようとしても，外見からは40台後半

といった判断はできても，正確な年齢を知ることはそれほどやさしくはない．そもそも，実際の調査票の設計においては連続な変数の数値をそのまま記入させずに区間を選ばせる形のものが多いし，集計された統計表においては連続変数の分布はヒストグラムの形で表されることがほとんどである．このように量的変数も区間に区切って質的な変数として考えれば，キー変数からなる個票データは高次元の分割表であると考えることができる．すなわちキー変数の個数を p として，$I_1,\ldots,I_p$ をそれぞれの変数のカテゴリー数とすれば，個票データは $I_1\times\cdots\times I_p$ 型の p 元の分割表と考えられる．分割表の用語にならって，キー変数のカテゴリーの組合せで得られるものを**セル**(cell)とよぶことにする．以下では分割表の総セル数を

$$K = I_1 \times\cdots\times I_p$$

と表す．ただしここでは5節で述べる構造的ゼロセルの問題を無視して総セル数を定義している．この設定のもとでは，上で個体について定義した母集団一意，標本一意等の概念をセルについて考えることができる．すなわち，セルに番号をつけて $j=1,\ldots,K$ とし，F_j をセル j の母集団での頻度とする時，例えば $F_j=2$ となるセル j を母集団2意セルという．

このように個票データを多元の分割表ととらえるのであるが，本来の多元の分割表としての扱いは第5節で論じることとする．ここではセルを1次元的に並べ，さらにとりあえずセルのラベルを無視することにより，モデルを単純化して考えよう．セルのラベルを無視して考えると，母集団(あるいは標本)において，一意のセルの数，2意のセルの数，等が問題となる．頻度が i の母集団のセルの数を

$$S_i = \sum_{j=1}^{K} I(F_j = i)$$

と書く．ただし

$$I(F_j=i)=\begin{cases}1, \text{ if } F_j=i\\ 0, \text{ otherwise}\end{cases}$$

は $F_j=i$ となることの定義関数である．$(S_0,S_1,\ldots,S_N)$ を母集団の**寸法指標**(size index)とよぶ．頻度の頻度(frequencies of frequencies (Good (1965)))とよぶこともある．容易にわかるように寸法指標は

$$K=\sum_{i=0}^{N} S_i$$

$$N=\sum_{j=0}^{N} iS_i$$

を満たす．同様に標本についても，セル j の標本での頻度を f_j で表し，標本の寸法指標を

$$s_i=\sum_{j=1}^{K} I(f_j=i),\quad i=1,\ldots,n$$

と定義する．

母集団からの標本抽出として簡単のために無作為非復元抽出を考える．$\lambda=n/N$ を抽出率とすると，母集団の各個体は確率 λ で抽出される．無作為抽出であるから，この確率は母集団一意であるか否かと無関係であり，母集団一意である各個体も確率 λ で抽出される．この事から，標本一意かつ母集団一意である個体数を $\underline{S}_1$ と表す時，$\underline{S}_1$ の期待値は

$$E(\underline{S}_1)=\lambda S_1=\frac{n}{N}S_1$$

で与えられることがわかる．従って λS_1 を標本でも母集団でも一意な個体数の推定値とすればよい．この単純化を前提にすれば，標本中の母集団一意数の推定問題は，母集団における母集団一意数 S_1 の推定問題に帰着される．そこで以下では開示リスクの評価の問題を S_1 の推定問題と同値なものとしてとらえる．ただしモデル化によっては，標本一意が母集団でも一意である確率を直接評価できるモデルを作ることもできる．

ここで，総セル数 K と母集団の大きさ N の大小について考える．N/K は母集団におけるセルあたりの平均個体数(平均セルサイズ)であり，N/K が小さい時ほど小頻度のセルが多くなる．逆に，平均セルサイズ N/K が例えば 100 程度であれば，小頻度のセルは少ないと考えられる．すなわち K が N に比べて小さい時には，個票データは安全になるはずである．この考え方は次の単純な事実によって正当化される．まず，自明に

$$S_1 \leq K$$

である事に注意しよう．従って

$$\frac{n}{N}S_1 \leq \frac{n}{N}K$$

であり，もし右辺の $(n/N)K$ が一桁あるいは二桁程度の小さい値であれば，最悪でも標本データ中の母集団一意 $\underline{S}_1$ は二桁程度である．S_1 は K よりはかなり小さい場合が多いと考えられるから，実際には $\underline{S}_1$ はせいぜい一桁程度となり，個票データは全く安全である．すなわち

$$\frac{n}{N}K \text{ が二桁程度ならば個票データは安全}$$

という簡明な事実が成り立つ．従って以下では，K が N に比べて大きい場合を念頭において議論することとする．

標本で 2 意以上の個体は母集団でも 2 意以上であるから，母集団一意数 S_1 の推定には標本一意数 s_1 のみが関連するようにも思われる．ところで，前節でも述べたように通常の標本調査では抽出率は低く，標本一意の s_1 人の個体の中で何人が同時に母集団一意であるかについては s_1 単独にはほとんど情報がない．この意味では母集団一意数の推定問題はかなり非正則な推定問題ということができる．しかしながら，いろいろな個票データから寸法指標を計算してみると，寸法指標の分布にはある程度のパターンが見られる．例えば，総セル数 K が十分大きい場合には，寸法指標は単調減少となる．従って，母集団における母集団一意数 S_1 を単独で考えるのではなく，$(S_0, S_1, \ldots)$ のパターンに何らかの関数形を想定したり，あるいは適当な順序制約等をおくことにより S_1 の推定が可能であると考えられる(本特集の論文，佐井(2003)を参照)．このことを裏返して言えば，S_1 の推定値が想定するモデルに大きく依存するということができる．実際に寸法指標のデータにさまざまなモデルを当てはめてみると，モデルのとり方によって S_1 の推定値に数倍以上のひらきが出ることもある．このような困難を念頭におけば，S_1 に関する推論の客観性を確保するためには，多くの操作的かつ柔軟なモデルを開発して，モデル群の中からのモデル選択を慎重におこなうことが重要である．個票開示問題の文脈では，データの持つ情報が少ないことから，ベイズ法が重要な役割を果たすことになるが，ベイズ法を用いる場合でもハイパーパラメータを導入して経験的ベイズ法にもちこむなど，事前分布の選択における主観性を緩和する工夫が必要である．本特集の他の論文で示されているように，最近になってわが国の研究者を含めてよい性質を持つモデルが開発されてきており，モデル開発については一定の解決のメドがついたと考えられる．

個票開示問題におけるベイズ法を標本調査法の文脈で解釈すると，それは超母集団モデルに対応する．母集団を固定すると，標本の分布のパラメータは，母集団のセル頻度 $(F_1, \ldots, F_K)$，あるいはセルのラベルを無視すれば母集団寸法指標 $(S_0, S_1, \ldots, S_N)$ である．そして，事前分布

を導入してこれらを確率変数であると想定することは，母集団のセル頻度が超母集団からの確率変数として実現するとモデル化することとなり，標本調査法でいう超母集団モデルとなる．未知のハイパーパラメータを含む超母集団モデルを想定すれば，標本のセル頻度 $(f_1,\ldots,f_K)$，あるいは標本寸法指標 $(s_0,s_1,\ldots,s_n)$ の周辺分布に基づいてハイパーパラメータの推定をおこなうことができる．そして S_1 の推定値としては，推定されたハイパーパラメータのもとでの超母集団分布に関して S_1 の期待値を求めればよい．

超母集団モデルの中で最も基本的なモデルは，母集団の N 人の個体がそれぞれ独立に多項のベルヌーイ試行に従って各セルに落ちて来るとするモデルである．第 j セルの生起確率を p_j, $j=1,\ldots,K$ とすると，母集団頻度の分布 $(F_1,\ldots,F_K)$ は多項分布に従うことになる．これを多項モデルという．ここで母集団から n 人の個体を非復元無作為抽出する状況を考えよう．多項分布が個体の順序に関して対称性(exchangeability)を持つことに注意すれば，標本の n 人は母集団の N 人のうち最初の n 人であると考えても一般性を失わないことがわかる．このように考えると，標本一意セル j が同時に母集団一意セルでもある条件つき確率は，標本に抽出されなかった $N-n$ 人の個体がいずれもセル j に落ちない確率に等しい．すなわちこのモデルでは，標本一意なセルが母集団一意でもある条件つき確率を直接に考慮することができる．多項分布でのモデル化の基本的事項は渋谷 (1997) で与えられている．我々が関心を持つ超母集団モデルは，総セル数 K が大きく，その中でも生起確率の低いセルが多く存在して，結果として母集団一意セルが多く実現するモデルである．Khmaladze (1987)はこのような状況を "Large Number of Rare Events" (LNRE)とよび，いくつかの重要な結果を導出している．統計的推定の観点からすると，生起確率 $(p_1,\ldots,p_K)$ に制約がなくすべて未知とすると，生起確率の推定はほとんど不可能である．また一般に $p_1,\ldots,p_K$ が互いに異なる場合には，寸法指標の分布は組合せ的な和の形でしか書けず，寸法指標からの推定も困難となる．セル間の対称性(exchangeability)を仮定して

$$p_1=\cdots=p_K=\frac{1}{K}$$

とすれば簡単となり，寸法指標の確率分布も明示的に書けるが，これでは分布が完全に指定されてしまい，モデルとしては柔軟性がない．これについては，$(p_1,\ldots,p_K)$ がさらにあるハイパーパラメータを含む分布からの実現値であると考える，すなわち多項分布の混合分布(mixture)を考えることによって，セル間の exchangeability を保ったままより柔軟なモデルを構築することができる．

生起確率の小さいセルの頻度はポアソン分布で近似できるので，多項分布をポアソン分布で近似してやると，ポアソン分布モデルとなる．あるいは，各個体がポアソン過程に従って到着し，さらに生起確率に比例してセルに落ちるとモデル化すれば，近似ではなくポアソン過程を直接用いたモデル化となる．ここで，母集団からの無作為非復元抽出を考えれば，最初に到着した n 人が観測されたと考えても一般性を失わない．このモデルでは n 人を観測した時点でセル j が標本一意である時，これが同時に母集団一意でもある条件つき確率は，残りの $N-n$ 人が到着し誰もセル j にはいらない確率に等しい．

多項分布と同様にポアソン分布でも，混合分布を考えることが有用である．ポアソン分布の混合分布として，最も基本的なモデルはポアソン分布の期待値パラメータにガンマ分布を仮定することである．混合の結果として負の2項分布が得られる．このモデルはポアソン・ガンマモデルとよばれ，非常に多くの分野で用いられる．個票開示の分野ではポアソン・ガンマモデルは Bethlehem et al. (1990) によって用いられるようになった．ポアソン・ガンマモデルのもとでの母集団一意数の分布の性質については佐井 (1998)を参考にされたい．

ポアソン・ガンマモデルにおいて，母集団の大きさ N を固定した条件つき分布で考えると，

多項分布をディリクレ分布で混合した多項・ディリクレモデルが得られる．さらにポアソン・ガンマモデルで，0意のセル数が無限に発散するような極限操作をおこなうと，Fisherの対数級数モデルが得られるが，ここでさらに母集団の大きさ N を固定すると，集団遺伝学の分野で提唱されたEwens sampling formula (Ewens (1972))とよばれる重要な確率分布が得られる．これらの事実についてはTakemura (1999) およびHoshino and Takemura (1998)で説明されている．またOmori (1999) は多項・ディリクレモデルの枠組でベイズ法を直接適用して母集団一意の事後確率を評価している．

ポアソン・ガンマモデル，多項・ディリクレモデル，対数級数モデル，Ewens sampling formulaの4つのモデルは，極限操作と条件つけによってお互いに関連しており，これらのモデルの推定結果も本質的に同等なものとなる．最近になって星野(Hoshino (2003, 2002a, 2002b))は，離散無限分解可能モデルを用いて，この4つのモデル間の関係を特殊なケースとして含む一般的な分布族を導くことに成功した．この成果は本特集の星野論文でも紹介されているが，これにより母集団一意数推定問題において柔軟なモデルが利用できることになった．

ポソンン分布の混合とは別のモデルの考え方として，確率分割のモデルの研究の流れがある．確率分割の基本的な考え方はSibuya (1993)で説明されている．特にPitman sampling formulaとよばれる確率分布のモデルはEwens sampling formulaの拡張になっており，寸法指標のデータへの適合度が高い(Hoshino (2001))．Pitman sampling formulaについては本特集の大和論文を参照されたい．

以上で紹介した超母集団モデルは有限個のハイパーパラメータによって指定されるパラメトリックモデルであるが，母集団の寸法指標にノンパラメトリックな順序制約をおく方法が佐井によって研究され，これも有望なアプローチである(本特集の論文，佐井(2003)を参照)．

以上では一意のみに注目して考えて来たが，2意以上についても考えておくほうがよい．母集団2意の個体は，母集団一意ほどのリスクはないが，やはり一定のリスクはあると考えられる．佐井 (2000) では予測個体数の観点から，例えば母集団 k 意の個体のリスクを母集団一意のリスクの $1/k$ と評価することによって，標本2意以上のリスクも評価したより総合的な開示リスクの評価を論じている．

すでに述べたように，この節で紹介したモデルの多くは，他の分野でも共通に用いられるものである．計量生態学で生物の種数の推定等に用いられるモデルについてはEngen (1978) を参照されたい．Baayen (2001) は単語の頻度分布に関する最近のすぐれた成書であり，計量言語学の分野で用いられる統計的手法がサーベイされている．統計の分野ではシェイクスピアの語彙に関するEfron and Thisted (1976), Thisted and Efron (1987) の研究がよく知られているが，Baayenの著書により計量言語学の分野での統計的手法の拡がりを概観できる．この方面の和書として影浦(2000)がある．

4. 大域的再符号化による秘匿処理

母集団一意数の推定に基づいて，開示リスクが大きすぎると判断される個票データについては，大域的再符号化を用いて開示リスクを減少させる必要がある．作業的には，試行錯誤的に大域的再符号化を施しては，前節で述べたリスクの評価をおこない，個票データがある程度安全になるまで大域的再符号化を繰り返すこととなる．

この際，大域的再符号化によっては，安全性と個票データの有用性が両立し得ないこともあり得る．例えば，企業や事業所を対象とした標本データについては，規模に関する変数は極めて歪んだ分布をしているため，たとえ区間表示であっても個体の秘匿は困難であり，公開用のデータとして安全なデータを作ろうとすれば，データの情報をほとんどすべて秘匿せざるを得

ないであろう．このような場合，安全性と有用性を両立させるのは困難である．

ただし Takemura (2002b) で例示されているように，大域的再符号化をあるところであきらめて，局所再符号化を積極的に採用することにより，個票データの有用性を保ちつつより安全な個票データを作成できる可能性もある．従って，大域的再符号化により個票データの情報を一律に粗くしすぎないように注意すべきである．次節で述べるように特に危険と考えられる個体について局所秘匿措置を施せば，安全かつ有用な個票データが得られる可能性があるからである．

大域的再符号化においては，有用な情報をできるだけ残す配慮が必要であり，試行錯誤的な作業となる．この試行錯誤はかなり面倒なものであるから，ある程度機械的に大域的再符号化の作業がおこなえると望ましい．これに関して，佐井・竹村(2000)ではポアソン・ガンマモデルの枠組の中で，カテゴリーを実際に併合しなくても，開示リスクの減少をあらかじめ評価できる方法を与えている．

もし，個票データの安全性と有用性の数値的な尺度が与えられれば，一定の安全性を保証した上で，有用性を最大にする大域的再符号化を，コンピュータを用いて機械的に探索することも考えられる．この目的のためには，個票データの持つ有用性，あるいは情報量を数値的に評価してやればよい．例えば，個票データの持つ情報量の評価としてシャノン流のエントロピーを用いる(Willenborg and de Wall (2001), 3.5 節参照)ことは有用な方法である．しかしながら，情報の意味的な側面を考えると，情報の有用性の数値化はそれほど容易ではない．エントロピーについても，統計的情報の有用性の観点からエントロピーの数値が何を意味するかは明らかではない．ここで情報の意味的な側面と言っているのは，次のようなことである．例えば年齢を区間に分けて表示する場合を考えてみると，分析によっては 18 歳以上と 18 歳未満を区別して解析したい場合もある．このような場合には，年齢を 0 歳から機械的に 10 歳刻みに区間分けしたデータは，当該の分析の目的に照らせば有用性がそこなわれているデータとなる．従って，大域的再符号化の探索手段としてエントロピーなどの情報量の数値的評価基準を用いることは有用であるが，やはり変数の意味や典型的な分析例を念頭においた取捨選択の作業は避けられないと思われる．

5. 個体ごとの識別リスクの評価

ここまでは個票データセット全体の開示リスクの評価と，大域的再符号化による秘匿措置を考えてきた．これらはデータセット全体に共通一律のリスク評価と秘匿措置である．方法的には，セルのラベルを無視した寸法指標に基づく超母集団モデルによるリスク評価と，それに基づく秘匿措置であった．しかしながら，個票データセットをより詳しく見れば，キー変数の値が単独あるいは組合せとして外れ値であって明らかに危険と見られる個体と，そうでもない個体が見出されるであろう．つまりキー変数が単独あるいは組合せとして「珍しい」個体ほど識別リスクが高いと考えられる．そして特に識別リクスの高い個体には，個別の秘匿措置，すなわち局所秘匿措置を施す必要がある．この節では個体ごとの識別リスクの評価方法について説明し，次節において局所秘匿のさまざまな方法を説明する．

個体ごとの識別リスクの評価としては，まずは常識的かつ当然の手続きとして，個別の変数ごとの外れ値に注目しなければならない．例えば高額所得者は明らかに識別リスクが高い．次に，変数値の組合せに関する外れ値にも注意する．例えば(男性，看護士)という組合せは現在ではまだ珍しいために識別リスクが高いと考えられる．以下では，これらの当然の手続きは前提とした上で，より進んだ個体ごとの識別リスク評価方法を紹介する．

個体の識別が，その個体の fingerprint によっておこることを第 2 節で述べた．ところで，一

つの個体は複数の fingerprint を持ち得る．例えば(年齢，職業，分野，地域，身長)の他に，(性別，年齢，身長，体重，職業)も fingerprint となるかも知れない．ある一変数の値が極端な外れ値であるために識別可能な個体は識別リスクが高いし，2つの変数の組合せで識別可能な個体もある程度識別リスクが高いと考えられる．このような観点から，その個体を識別するのに必要な最小数のキー変数の組合せを **最小危険変数群**(minimum unsafe combination of variables)とよぶ．そこで，最小危険変数群をなす変数の数が小さい個体ほど危険であると考えることができる．最小危険変数群は母集団に関しても考えることができるが，ここでは標本に含まれる標本一意個体の相対的な識別リスクを評価することを念頭においているので，標本内での最小危険変数群について考えることとする．最小危険変数群の概念は Willenborg and de Wall ((1996), 5.4 節)で導入された．Takemura (2002a)では，最小危険変数群の概念とその裏返しとしての最大安全変数群の概念の理論的性質について論じている．最大安全変数群とは，標本一意の個体について，標本一意とならない最大個数のキー変数の組合せをいう．Takemura (2002a)では，標本一意な各個体について最小危険変数群と最大安全変数群を求めるアルゴリズムも与えている．またこれらの概念は標本2意，3意等にも拡張できる．

最小危険変数群の手法はいわば記述統計的な手法であるが，確率モデルを用いたアプローチとしては，多元の分割表をセルのラベルを無視せずに分割表のままで扱い，セルの生起確率を推定する方法が考えられる．このアプローチの基本的な考え方は次のようなものである．まずは簡単のため最も基本的な多項分布モデルを考えよう．標本一意となっているセルが，母集団でも一意となる条件つき確率は，そのセルの生起確率が小さいほど高い．多項モデルにおいてセル j の生起確率を p_j と表す．n 人の中でセル j の頻度が1であったという条件のもとで，このセルが母集団でも一意，すなわち標本に含まれなかった $N-n$ 人の誰もがセル j に落ちない条件つき確率は

$$(1-p_j)^{N-n}$$

と表される．この条件つき確率は p_j の単調減少関数で，p_j が0に近付くときに1に収束する．従って，標本一意のセル j が母集団でも一意となる条件つき確率は，生起確率 p_j が小さいほど高い．また，p_j の推定値 $\hat{p}_j$ が得られれば，$(1-\hat{p}_j)^{N-n}$ によって，標本一意セル j が母集団一意でもある条件つき確率を推定できる．セル j を一意と仮定したから単なる相対頻度は $\hat{p}_j = 1/n$ であるが，我々の問題においてはこの推定量はもちろん無意味であって，生起確率が小さいセルの確率に対して何らかのスムージングの操作が必要である．スムージングはセルの生起確率を少ないパラメータでモデル化することによって実現することができる．

分割表の確率モデルとして標準的なモデルは対数線形モデルである．そして，対数線形モデルにおいて，例えば2変数の交互作用項までをモデルにとり入れることが考えられる．開示リスクの評価の道具として多元の分割表に対数線形モデルを用いた文献としては Skinner and Holmes (1998), Fienberg and Makov (1998) があげられる．対数線形モデルの問題点としては，分割表の総セル数 K が大きい時には，対数線形モデルのあてはめが計算量的に困難になるという問題が指摘できる．この問題は統計物理的な確率モデルでも，分配関数(基準化定数)の評価の困難さの問題としてよく知られている．そこで Takemura (2002c) は計算の簡便のために Lancaster 形の線形モデルを用いることを提案している．ただし生起確率に線形モデルを当てはめると，セルの生起確率の推定値として負の推定値の得られる場合のあることが問題となる．

セルごとの開示リスクの評価として分割表の確率モデルを利用することの目的を相対的なリスクの評価ととらえるならば，必ずしもモデルの全体的な適合度にこだわる必要はないかも知れない．すなわちモデルを相対的に開示リスクの高いセルの候補を抜き出す道具ととらえるわけである．モデルによって抜き出されたセルのリスクの評価については，分割表の確率モデル

以外の方法を併用することも考えられる．

分割表の確率モデルをあてはめる際に重要な問題となるのは**構造的ゼロセル**(structural zero cell)の問題である．これは次節で述べる局所再符号化においてより深刻な問題となる．構造的ゼロセルとは，定義上観測値のあり得ないセルをいう．例えば(運転免許証保有，18歳未満)というような組合せはあり得ないから，このようなセルの頻度は定義的にゼロである．実は社会経済事業を扱った個票データにおいては，構造的ゼロセルは非常に多く出現する．従ってモデルをあてはめる際にも，構造的ゼロセルの生起確率を0と制約してモデルを推定すべきである．しかしながら，モデル推定の際に，構造的ゼロセルを事前にすべて指定するのは，変数の組合せについて一々その意味を考察する必要が生じ，実際的に不可能である．一つの便法としては，例えば2次元の周辺分割表において頻度が0となっているセルは，すべて構造的ゼロセルと見なすということも考えられる．すなわち2次元周辺表にまとめて考えると，各セルに十分な頻度が観測されるはずであるから，構造的ゼロセルではない場合には正の頻度が観測されると予想できるからである．対数線形モデルは構造的ゼロセルが存在しても推定における理論的な困難は生じないが，線形モデルでは構造的ゼロセルの扱いは面倒であり，構造的ゼロセルの生起確率が負と推定される問題が生じる．この場合は負の推定値をゼロに切り上げるなどの操作が必要である．

以上では，多元分割表の多項分布によるモデル化を論じたが，生起確率のパラメータに事前分布を導入して，ベイズ法あるいは経験ベイズ法を用いることも考えられる．これは今後の研究課題の一つである．

6. 局所再符号化による秘匿処理

ここでは，前節の方法等によって，局所秘匿が必要な個体が指定されたとして，局所秘匿の具体的な方法について述べる．局所秘匿法の概観としては，Willenborg and de Wall (2001) の3–5章および Fienberg et al. (1998)が参考になる．

局所秘匿として最もわかりやすいのは局所欠測化の方法であろう．個別の個体については，最大安全変数群を求め，最大安全変数群のみを残して他の変数を欠測化すれば，標本一意でなくなり，最小の個数の欠側値で秘匿ができる．しかしながら，標本一意は必ずしも母集団一意ではないから，欠測化によって標本一意をすべて避けるのは秘匿が過大となる可能性が大きい．一定の安全性を確保した上で，局所秘匿の数を最小化する形での問題の定式化は Willenborg and de Wall (2001) の4.2節で議論されているが，欠測化がもたらす個体間の影響までも考慮すると局所秘匿の最適化は難しい問題である．これについては次節でより詳しく述べる．

最近提案された方法で，実用的でもありまた簡明な方法として PRAM (Post RAndomization Method，マルコフ連鎖を用いた秘匿，Gouweleeuw et al. (1998))があげられる．PRAM は，$K \times K$ の適当なマルコフ推移確率行列 $P = (p_{ij})$ を設定して，セル i に属する個体を確率 p_{ij} で セル j に移動させて秘匿するものである．同じセルに留まる確率 p_{ii} を1に近くしておけば，実際に移動する個体は少なくなる．従って $1 - p_{ii}$ の値を調整することによって秘匿の程度を制御することができる．ユーザの立場から見ると，PRAM は randomized response の手続きによって得られたデータと同様であり，概念的には簡明である．Randomized response とは，例えばエイズへの感染など正直に答えにくい質問をする際に，「コインを自分でふって表ならば正直に答え，裏ならば逆を答えてください」といった形で，答えが質問者に直接わからない形でデータを収集する方法である．もちろん，ゆがみのないコインを用いてしまうと，答えは Yes と No が等確率となってしまい無意味であるが，コインの表の確率が1/2以外であれば，真の値について統計的推測をおこなうことができる．Randomized response データの解

析の方法としては，回答者の真の答を潜在変数あるいは欠測値として，EM アルゴリズム等の欠側値が存在する場合の統計的推測の方法を用いればよい．PRAM は個票データの提供者が randomization をおこなうものであるが，個票データのユーザから見ると randomized response と同等である．ただしこの同等性は推移確率行列 $P = (p_{ij})$ が公表されていることが前提である．秘匿方法自体を秘匿するかどうかに関する問題点については第 9 節でも論じる．推移確率行列 $P = (p_{ij})$ が公表されている限り p_{ii} が必ずしも 1 に近くなくても統計的推測は可能であるから，この場合 PRAM によって秘匿されたデータは必ずしも原データに非常に近い必要はない．Randomized response モデルについては Chaudhuri and Mukerjee (1988), van den Hout and van der Heijden (2002) を参照されたい．

次にスワッピングと局所再符号化について述べる．スワッピングは似た個体間で観測値を交換してしまう秘匿方法であるが，スワッピングと局所再符号化には密接な関連がある (Takemura (2002b))．例えば，個体 i の年齢が 40 歳，個体 j の年齢が 45 歳の時，これらの年齢の値を入れ換えるのがスワッピングであるが，これらの年齢をプールして，両個体について年齢を 40–45 のように区間表示することが考えられる．このように，局所再符号化の一つの方法として，似た個体間で観測値をプールして区間表示する方法が考えられる．Takemura (2002b) では重みつき最適マッチングのアルゴリズムを応用することにより，似た個体同士のペアを最適に構成する方法を提案している．ユーザの立場からすると，局所的に区間表示されたデータは，統計パッケージへの読み込みなどで扱いがやや面倒である．しかしながら，区間から無作為に値をとり出すことを考えると，その結果はスワッピングされたデータと同様のデータとなる．従って，局所再符号化されたデータは，スワッピングの操作をデータ提供側からユーザ側へシフトしたものと考えることもできる．またユーザは以上の作業を繰り返すことにより，結果の変動を確認できるから，解析の結果に対する局所再符号化による情報の損失を確認することもできる．

似た個体同士を集めるには，統計的クラスタリングの標準的な方法を用いることもできる．クラスタを作った後で，各個体の観測値をクラスタ内の平均値でおきかえる方法は Micro-aggregation とよばれている (Defays and Anwar (1998))．Micro-aggregation の手法でもクラスタ内の平均値におきかえず，区間表示にすれば局所再符号化となるし，クラスタ内の個体で観測値を交換すればスワッピングとなる．各個体の観測値をクラスタ内の平均値でおきかえる方法は欠側値の imputation で実際に行われることが多く，その限りではこれ迄も利用されてきた方法である．

スワッピング等の局所秘匿処理において留意すべき点は，構造的ゼロセルの問題である．例えばスワッピングを単純におこなって年齢を入れ換えると，(運転免許証保有，18 歳未満) のような論理的な不可能な組合せが生じてしまう可能性がある．構造的ゼロセルに個体を移動した場合には，これが秘匿された個体であることがわかってしまい，秘匿を解除されてしまう可能性もある．従って，構造的ゼロセルにはいらない形で局所秘匿をおこなうことが望ましいが，この点はまだ実用的な解決法は提示されていないと思われる．

前節で述べたように，それぞれの変数の意味を考えた上で構造的ゼロセルのリストを作るのは困難なので，2 次元の周辺表で頻度が 0 のセルを構造的ゼロセルと見なして考えるのが簡便である．このことから，個票データを多元の分割表として考え，すべての 2 次元の周辺度数を変えないように他の分割表に移動するということが考えられる．実はこの問題は，分割表の対数線形モデルにおいて，マルコフ連鎖法を用いて正確検定をおこなう問題と密接に関連している．そして，低次元の周辺度数を固定した分割表間の移動の問題は，技術的には非常に難しい問題である (Diaconis and Sturmfels (1998), Dobra (2001))．周辺度数のある程度の変動を許すことにすれば問題は容易になる可能性もあるが，今後の研究課題である．

7. 局所秘匿後のデータの開示リスクの評価

個体ごとの開示リスクの評価に基づいて局所秘匿処理をおこなったときには，秘匿後の個票データについての開示リスクをあらためて評価する必要がある．しかしながら，局所秘匿後の個票データの開示リスクの評価については，現状ではほとんど理論的な結果が得られていない状況である．局所秘匿後のデータの開示リスクの評価方法の確立は今後の重要な研究課題である．

まずはこの問題の難しさを説明しよう．いま個票データのある個体 i の観測値にすべて×をつけて欠測化したとする．いま他の個体 j が残りの個体の中で標本一意であるとしても，個体 i の観測値はすべて欠測となったから，個体 i の観測値が個体 j の観測値とすべて一致し個体 j が標本一意でない可能性もある得る．このように考えると，個体 i を欠測とすることにより個体 j の安全性が増したかのように思われる．しかしながら，この議論を認めてしまうと，i 以外の個体はすべて安全となってしまう．一方で，個体 i の観測値にすべて×をつけて欠測化することは，個体 i を個票データから削除することに他ならず，標本の大きさ n を $n-1$ にしただけである．従って，他の観測値が安全になったと考えることは不合理であると思われる．すでに述べたように，サブサンプルは単に n を減少させているだけだから，サブサンプルをとれば個票データが非常に安全になるとは言えないことに注意しよう．

さらに論理的に困難な問題は，局所秘匿後に「なぜその場所を秘匿したか」ということがわかってしまい，秘匿が解除されてしまう可能性があるという点にある．これを簡単な例で例示しよう．いまある大学のゼミが学生4名から成るとして，性別と学年が表1のようであったとする．女性 (F) は一人であるのでFを秘匿すると表2となる．しかしながら，秘匿した学生がもし男性ならばそもそも秘匿しなかったはずであるから，事後的には秘匿した箇所が女性である事がわかってしまう．すなわち局所秘匿の理由を考えることにより，秘匿が破られる論理的な可能性があるのである．

実はこれは次節で述べる表形式のデータの2次秘匿と同様の問題であることがわかる．すなわち秘匿した場所以外の情報から，秘匿した場所の情報が解除される可能性がある．このために，危険と思われるセルのみならず，一見安全と思われるセルにも秘匿を施す必要がある．

以上のように局所秘匿における個体間の相互作用を考慮すると，局所秘匿後の個票データのリスク評価には基本的な困難があることがわかる．一方で，ある個体の秘匿措置により他の個体もある程度安全になると考えられるが，他方で，秘匿措置が解除されてしまう可能性もある．ところで，上の簡単な例は非常に小さい表であるために，秘匿の解除の可能性が浮き彫りになっているが，通常のサイズのデータセットでは，個体間の相互作用の影響を正確に把握することは計算量的に困難な問題であると思われる．この場合，個体間の相互作用の影響を利用することは攻撃者にとっても困難であり，開示リスクの評価においては個体間の相互作用を無視しても大きな問題ではないとも考えられる．

個体間の相互作用を無視すれば，局所秘匿後の個票データの開示リスクの評価は，個体ごと

表1.

学生	性別	学年
A	M	4
B	M	3
C	M	3
D	F	4

表2.

学生	性別	学年
A	M	4
B	M	3
C	M	3
D	×	4

に属するカテゴリーの大きさが可変であるような個票データについての母集団一意数の推定問題ととらえることができる．この立場に立てば母集団一意数推定の理論が応用でき，局所秘匿後の個票データの開示リスクも評価できる可能性がある．これも今後の課題である．

8. 表形式データの秘匿，オンライン秘匿

本稿では個票データの秘匿について述べて来たが，ここでは集計表データの秘匿およびオンライン秘匿について簡単にふれておく．

通常の集計表においては秘匿はあまり問題とならないが，日本の官庁統計でも，事業所を対象とした全数調査である「工業統計調査」「商業統計調査」のように，小さな地域毎の詳しい集計表を公表しようとすると，集計表のセルにおいても1意や2意が現れ，その場合そのセルの値を欠測とする秘匿措置がおこなわれている．集計表の秘匿においては，直接秘匿の対象となるセルの秘匿を**1次秘匿**という．集計表では，通常は周辺和や周辺頻度の情報が記載されているために，1次秘匿だけでは周辺和等の情報から，引き算で秘匿箇所の値がわかってしまう場合がある．このために，直接秘匿対象となるセル以外のセルを同時に秘匿する必要が生じる．この補助的な秘匿を**2次秘匿**という．2次秘匿は1回で済むとは限らず，さらに追加的な秘匿を必要とする場合もあるが，これを含めて2次秘匿という．2次秘匿によって，秘匿されたどのセルも周辺和等の情報からは回復できないようにする．

さらに注意すべきことは，2次秘匿によって各セルの正確な値は秘匿されたとしても，集計表の各セルの値が非負であるという制約を用いると，線形不等式を変形することによって，1次秘匿の対象のセルの値が狭い範囲に限定されてしまう可能性があることである．ユーザの立場からの，秘匿されたセルを含む集計表のセルの補完や，秘匿されたセルの値の範囲の計算については稲葉・岩崎による一連の研究成果(Inaba (1997)，稲葉・岩崎(1996, 1997))があり，特に商業統計表を実例としてユーザの立場から秘匿されたセルを含む集計表の分析法を示している．

2次秘匿はできるだけ行わないことが望ましい．2次秘匿の箇所をできるだけ少なくするなどの2次秘匿のパターンの最適化は，表の構造が複雑になると厳密に行うことが難しくなる．また同一のセルが複数の表に現れる場合には，組合せによってそのセルの秘匿が解除される可能性もある．表形式データの秘匿は，実務的には一定の手続きに基づいておこなわれているが，以上のような問題点をふまえ，理論的にはさらに研究をおこなう必要がある．

次にオンライン秘匿について述べる．いまある個票データをデータベース化し，個票データからのさまざまな集計表を，要求に応じて自動的に提供するようなオンラインデータベースを考えよう．つまり当初から公表される集計表様式が指定されているのではなく，利用者からの要求に応じてのオーダーメイド集計のオンライン版である．この場合，特定の個体の情報を引き出そうという検索要求には答えず，集計量のみを答えるデータベースをここでは統計的データベース (statistical database)とよぶこととする．統計的データベースの性質に関してはDenning (1982)の第6章に解説されているように1980年代から計算機科学の分野でのかなりの研究の蓄積がある．その中で明らかとなって来たことは，集計量のみを答えることにしても，さまざまな検索要求を繰り返すことにより，特定の個体の情報が引き出されてしまう可能性を排除できないということである．例えば，ある統計的データベースに50歳の個体が1人のみ含まれているとしよう．このデータベースに「51歳以上の人の平均年収はいくらか」という検索要求と，「50歳以上の人の平均年収はいくらか」という検索要求を出して，その双方に答が帰ってきた場合，引き算によって50歳の個体の収入が特定されてしまう．

このように，柔軟な検索要求を受け付けることとすれば，検索の組合せによって個体の情報が引き出される可能性がある．これを排除するために，特定の質問しか受け付けないデータベー

スを作るのであれば，それはあらかじめ提供可能な集計表のリストを作ることと同等となり，厳密にはオーダーメイド集計とは言えないものとなる．しかしながら，検索要求の組合せによる開示の可能性を論理的に考察すると，あらかじめ定められた範囲の質問のみを受け付けるデータベースによってしか，論理的安全性は保証されないのではないかと思われる．ところで，個票データを多元の分割表と見るならば，安全な個票データそのものをあらかじめ多元に集計された集計表と考えることができる．この意味では，あらかじめ定められた特定の質問しか受け付けないデータベースとは，安全な個票データをまず作成し，この個票データをデータベースとして，それに対するあらゆる検索要求に答えるデータベースと実質的には同様のものと考えることもできる．実際のオンラインデータベースの実装においては，より柔軟なオーダーメイド集計システムを設計することが重要であるが，論理的には安全なオンライン統計データベースと安全な個票データはほぼ同等のものと理解してよいであろう．

統計的データベースのオンライン秘匿に関しては，検索要求があるごとに結果にノイズを付加することによって，複数の検索を組み合わせられても個体の情報が引き出されないようにすることも考えられている．しかしながら，これも同じ検索要求を繰り返して，帰ってきた答えの平均を計算することにより，秘匿の効果が薄められてしまうという難点がある．

オンライン統計データベースの秘匿について以上のような論理的な困難はあるものの，ネットワークの普及とともにオンライン統計データベースに対する需要も高まってくると考えられる．分厚い統計表を購入しなくても，インターネットを通じて必要な集計データが得られれば，統計のユーザにとっては大変便利である．ただし単なる集計データの提供のみならず，統計調査の背景情報などのメタデータの提供の重要性にも注意する必要がある．オンライン統計データベースの運用における，検索の柔軟性と安全性の確保の問題も今後の重要な研究課題である．実務的には，ユーザにとって使い勝手がよく，かつ実際上十分安全なシステムの設計・実装が課題である．

9. その他の問題点

ここでは，これまでで論じきれなかったいくつかの問題点について議論する．また第1節に引続き，個票開示問題に関する統計数理的研究の意義と限界についても議論する．

個票データの開示リスクに関して，実際上最も重要なファクターは，個票データを実際に攻撃する者がいるかどうかである．仮にもし母集団一意をたくさん含む個票データがあったとしても，個票データには直接識別子は含まれないから，キー変数を組み合わせた意図的な攻撃がなければ識別はおきない．従って攻撃者の動機をどのように想定するかがリスクの総合的な判断においては重要である．おそらく最悪の想定は，攻撃者の動機が統計当局の信用を失墜させること自体にあり，攻撃者が特定の個体を識別したと意図的に宣伝するような場合である．この場合でも，識別は間接的であるから，特定の個体が母集団一意であることを攻撃者が証明することは困難であり，統計当局としてはこのような宣伝を無視することも考えられる．しかしながら，やはり識別自体が問題となるような事態は避けるべきであり，攻撃者に対して攻撃自体が無駄であると思わせる程度の秘匿措置をほどこしておくことが重要であると思われる．攻撃者の動機や攻撃の確率まで考えて開示リスクを総合的に論じた文献として Marsh et al. (1991) および Dale and Elliot (2001) をあげておく．

個票データの開示においてしばしば議論される点は，時点の古いデータであれば公表可能ではないかということである．イギリスなどでは，国勢調査データが長期間保管され，100年以上前のデータは秘匿なしに公開されて，詳しい歴史的研究のために活用されている．これに対して我が国では，これまで統計データを保存せず廃棄して秘密保持を守る制度であり，個票デー

タは破棄されている現状である．直観的には時間の経過とともに，開示リスクが減少して行くことは自明に思われる．問題は，リスクの減少をどのように数量的に評価するかである．これについては次のように考えればよいであろう．まず第一に，時間が経つにつれて人々の記憶は薄れ，キー変数の識別の精度が落ちると考えられる．すなわち，自然に大域的再符号化が起きているわけである．従って，時点の古いデータであれば，開示リスクの評価において，人々の記憶の確かさに見合う程度に大域的再符号化をして評価すればよい．もう一つには，時間とともに人々はある程度移動して行くから，その意味で一定のスワッピングが自然に起こると考えることができる．

個票データの開示においては，秘匿方法自体を秘匿するかどうかという問題がある．秘匿方法自体を秘匿すれば，秘匿はより強くなるから，秘匿方法自体の秘匿・公開は，開示リスクの制御の一部としてとらえなければならない．すなわち，秘匿方法自体は秘匿するがデータ自体の秘匿は緩くする場合と，秘匿方法は公開するがデータ自体には強い秘匿をかける場合とを比較する必要がある．ただし秘匿方法自体の秘匿の効果を数値的に評価するのは難しい．具体例として PRAM を用いて秘匿した場合に，推移確率行列 P を公表するかどうかという点があげられる．P を公開するならば，P は単位行列からかなり離れてもよいであろう．またスワッピングと局所再符号化の関係についても，単なるスワッピングでは秘匿箇所がわからないが，局所再符号化をした場合には秘匿箇所がわかってしまうことに注意する．従って少数のスワッピングをおこなうか，より多数の局所再符号化をおこなうかを比較して判断する必要がある．なお，PRAM において推移行列 P を公表すると，推移行列 P の定常分布を計算される可能性があるために，P を公表する場合には定常分布が原データの経験分布と一致するような P を用いるべきではないことに注意する必要がある．

個票データの開示リスクの評価においては，利用者の範囲の設定も重要な要素となる．cd-rom 等を介して販売する一般公開用のデータであれば強い秘匿措置を施す必要があるが，学術研究その他の目的で利用者を特定・限定して提供する場合にはより緩い秘匿措置でよい．特に企業・事業所に関するデータは秘匿が困難であるために，後者の形で提供せざるを得ない．ただし，一般公開か利用者限定かという提供形態の区別は，必ずしも 0 か 1 かというものではない．例えば利用者を限定したと言っても，研究室や自宅へのデータの持出しを許す場合と，特定の場所から持出しを許さない場合では，秘匿措置は異なってしかるべきである．利用者の立場から言うと，まずは公開用のデータを分析して，より詳細な分析が必要となった場合には統計当局に申請して身元を明らかにし，利用者を特定した形で秘匿措置の緩いデータが利用できると都合がよい．このように，個票データの開示においては，公開か限定かの区別を二者択一的にとらえず，秘匿の強さの異なる個票データを用意して，利用者の範囲に応じて提供することが望ましい．

参 考 文 献

Baayen, R. H. (2001). *Word Frequency Distributions*, Kluwer, Dordrecht.

Bethlehem, J. G., Keller, W. J. and Pannekoek, J. (1990). Disclosure control of microdata, *J. Amer. Statist. Assoc.*, **85**, 38–45.

Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Response: Theory and Techniques*, Marcel Dekker, New York.

Dale, A. and Elliot, M. (2001). Proposals for 2001 samples of anonymized records: An assessment of disclosure risk, *J. Roy. Statist. Soc. Ser. A*, **164**, 427–447.

Dalenius, T. and Reiss, S. P. (1982). Data-swapping: A technique for disclosure control, *J. Statist. Plann. Inference*, **6**, 73–85.

Defays, D. and Anwar, M. N. (1998). Masking microdata using micro-aggregation, *Journal of Official Statistics*, **14**, 449–461.

Denning, D. E. R. (1982). 『暗号とデータセキュリティ』(上園忠弘 他 訳), 培風館, 東京.

Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions, *Ann. Statist.*, **26**, 363–397.

Dobra, A. (2001). Statistical tools for disclosure limitation in multi-way contingency tables, Ph.D. Thesis, Department of Statistics, Carnegie Mellon University (available from http://www.niss.org/adobra.html).

Domingo-Ferrer, J. (ed.) (2002). *Inference Control in Statistical Databases. From Theory to Practice*, Lecture Notes in Comput. Sci., **2316**, Springer, Berlin.

Doyle, P., Lane, J. I., Theeuwes, J. J. M. and Zayatz, L. V. (2001). *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, Elsevier, Amsterdam.

Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know?, *Biometrika*, **63**, 435–447.

Engen, S. (1978). *Stochastic Abundance Models*, Chapman and Hall, London.

Ewens, W. J. (1972). The sampling theory of selective neutral alleles, *Theoretical Population Biology*, **3**, 87–112.

Fienberg, S. E. and Makov, U. E. (1998). Confidentiality, uniqueness, and disclosure limitation for categorical data, *Journal of Official Statistics*, **14**, 385–397.

Fienberg, S. E., Makov, U. E. and Steele R. J. (1998). Disclosure limitation using perturbation and related methods for categorical data, *Journal of Official Statistics*, **14**, 485–502.

Fuller, W. A. (1993). Masking procedures for microdata disclosure limitation, *Journal of Official Statistics*, **9**, 383–406.

Good, I. J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*, MIT Press, Cambridge, Massachusetts.

Gouweleeuw, J. M., Kooiman, P., Willenborg, L. C. R. J. and de Wolf, P. P. (1998). Post randomisation for statistical disclosure control: Theory and implementation, *Journal of Official Statistics*, **14**, 463–478.

Hoshino, N. (2001). Applying Pitman's sampling formula to microdata disclosure risk assessment, *Journal of Official Statistics*, **17**, 499–520.

Hoshino, N. (2002a). On limiting random partition structure derived from the conditional inverse Gaussian-Poisson distribution, Tech. Report, CMU-CALD-02-100, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania.

Hoshino, N. (2002b). Engen's extended negative binomial model revisited, Discussion Paper No. 2002-1, Faculty of Economics, Kanazawa University.

Hoshino, N. (2003). Random clustering based on the conditional inverse Gaussian-Poisson distribution, *J. Japan Statist. Soc.*, **33**, 105–117.

Hoshino, N. and Takemura, A. (1998). Relationship between logarithmic series model and other superpopulation model useful for microdata disclosure risk assessment, *J. Japan Statist. Soc.*, **28**, 125–134.

Inaba, Y. (1997). Statistical inference for statistical tables with nondisclosure cells, *J. Japanese Soc. Comput. Statist.*, **10**, 59–72.

稲葉由之, 岩崎 学 (1996). クロス集計表における秘匿の影響に関する数値的評価, 応用統計学, **25**, 61–72.

稲葉由之, 岩崎 学 (1997). 統計表における秘匿の補完法, 日本統計学会誌, **27**, 263–280.

影浦 峡 (2000). 『計量情報学——図書館／言語研究への応用——』, 丸善, 東京.

Khmaladze, E. V. (1987). The statistical analysis of large number of rare events, Tech. Report, MS-R8804, Center for Mathematics and Computer Science, CWI, Amsterdam.

Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D. and Walford, N. (1991). The case for samples of anonymized records from the 1991 census, *J. Roy. Statist. Soc. Ser. A*, **154**, 305–340.

松田芳郎, 濱砂敬郎, 森 博美 編著 (2000). 『講座ミクロ統計分析 1 統計調査制度とミクロ統計の開示』, 日本評論社, 東京.

Omori, Y. (1999). Measuring identification disclosure risk for categorical microdata by posterior population uniqueness, *Statistical Data Protection—— Proceedings of the Conference, Lisbon, 25 to 27 March 1998-1999 Edition*, 59–76, Office for Official Publications of the European Communities, Luxembourg.

佐井至道 (1998). 個票データにおける個体数とセル数の関係, 応用統計学, **27**, 127–145.

佐井至道 (2000). 予測個体数の期待値に基づく個票データのリスク評価, 統計数理, **48**, 229–251.

佐井至道(2003). 母集団寸法指標のノンパラメトリック推定, 統計数理, **51**, 183–197.

佐井至道, 竹村彰通 (2000). 個票データにおける分類の併合モデル, 応用統計学, **29**, 64–82.

佐藤博樹, 石田 浩, 池田謙一 編 (2000). 『社会調査の公開データ——2 次分析への招待』, 東京大学出版会, 東京.

Sibuya, M. (1993). A random clustering process, *Ann. Inst. Statist. Math.*, **45**, 459–465.

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, **26**, 161–170.

Skinner, C. J. and Holmes, D. J. (1998). Estimating the re-identification risk per record in microdata, *Journal of Official Statistics*, **14**, 361–372.

Takemura, A. (1999). Some superpopulation models for estimating the number of population uniques, *Statistical Data Protection—— Proceedings of the Conference, Lisbon, 25 to 27 March 1998–1999 Edition*, 45–58, Office for Official Publications of the European Communities, Luxembourg.

Takemura, A. (2002a). Minimum unsafe and maximum safe sets of variables for disclosure risk assessment of individual records in a microdata set, *J. Japan Statist. Soc.*, **32**, 107–117.

Takemura, A. (2002b). Local recoding and record swapping by maximum weight matching for disclosure control of microdata sets, *Journal of Official Statistics*, **18**, 275–289.

Takemura, A. (2002c). Evaluation of per-record identification risk by additive modeling of interaction for contingency table cell probabilities, *IASS Proceedings—— SEOUL 2001*, 220–235, International Association of Survey Statisticians, The International Statistical Institute.

Thisted, R. and Efron, B. (1987). Did Shakespeare write a newly discovered poem?, *Biometrika*, **74**, 445–455.

van den Hout, A. and van der Heijden, P. G. M. (2002). Randomized response, statistical disclosure control and misclassification: A review, *International Statistical Review*, **70**, 269–288.

Willenborg, L. and de Waal, T. (1996). *Statistical Disclosure Control in Practice*, Lecture Notes in Statistics, No. 111, Springer, New York.

Willenborg L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*, Springer, New York.

Current Trends in Theoretical Research of Statistical Disclosure Control Problem

Akimichi Takemura

(Graduate School of Information Science and Technology, University of Tokyo)

We introduce the theory of statistical disclosure control and survey the current status and perspectives of theoretical research on statistical disclosure control. In addition to an overview of international research trends, we give some detailed treatments of the works of a group of Japanese researchers including the author.

Key words: Key variable, population unique, global recoding, local recoding.