統計的開示抑制 (statistical disclosure control) について

竹村 彰通　　東大情報理工

2009 年 7 月 4 日 PPDM 研究会

第 I 部: 分野のサーベイ

1. 導入と背景
2. 開示リスクの評価法
3. 母集団一意数推定のためのモデル
4. 個体ごとの識別リスクの評価
5. 局所秘匿の手法
6. その他の話題

第 II 部: 分解可能モデルに基づく秘匿措置

7. 多元分割表と階層モデルの記法
8. アメリカのセンサス個票データの例
9. センサス個票データでの計算結果
10. スワッピングによる秘匿
11. 周辺頻度を固定したスワッピングのための条件

導入と背景

- 多くの国では官庁統計の個票データが提供されている
- 日本では極めて限定的な形でしか提供されていなかった (いわゆる「目的外使用」)
  - → 旧統計法の解釈
  - → 個票開示の標準的手法が確立されていない
- 現時点で日本の統計制度は大きく変化しつつある:

  新統計法 (平成 21 年 4 月全面施行)

http://www.stat.go.jp/index/seido/1-1n.htm

「公的統計の体系的かつ効率的な整備及びその有用性の確保を図るため、統計法の全部改正が行われ、同法は、平成 21 年 4 月 1 日に全面施行されました。」

- **統計法のポイント**
  - **公的統計の体系的・計画的整備**
  - **統計データの利用促進**
    (その中の項目として) 委託による統計の作成、匿名データの作成・提供

「学術研究目的または大学などの高等教育などのために、オーダーメードで集計された統計の提供を受けたり、匿名データ（統計調査によって集められた情報を個人や企業が特定できない形に加工したもの）の提供を受けて統計の作成に用いることができます。」

- 統計調査の対象者の秘密の保護の強化
- 統計委員会の設置

- 日本でも実証研究における個票データへの (潜在的) 需要はある
- 統計研究者の中では個票データの提供を積極的に要望する声が多い
- 社会調査データとデータアーカイブの整備も遅れている

各国の事例 (個人的印象)

- アメリカでは，センサスの詳しい個票データを PUMS という形で提供してきた．局所的な秘匿処理であり swapping 必須 (後述)．最近のインターネットでの検索の容易さ等を考えると，詳しすぎるという議論も?
- American Fact Finder でオンライン集計表も柔軟に提供している．この安全性の評価も難しい

- **北欧の諸国では，VPN (Virtual Private Network) の技術を導入して，研究者と政府統計当局の間をインターネットの中に仮想的に暗号化した通路を確保して画面のみを研究者のパソコンに表示するしくみをとっている．**

開示リスクの評価法

- **開示リスク**: 個票データが提供されると，データに含まれる個体が識別される危険がある
- 電話番号等の「直接識別子」は当然データから削除する
- つまり，ここで考える識別は「キー変数」の珍しい組み合わせによる間接的なものである
- **キー変数**: 性別，年齢，職業など間接的に個人を特定するために用いることのできる変数

- **識別可能性の二つの意味:**
  - **論理的に識別され得ること**
  - **攻撃者が実際に識別しようとして，成功する可能性**
- **実際に攻撃者が識別しようとするかどうかは，識別のためのコストとその結果の利益の程度による**
- **「論理的識別可能性」と「実際の攻撃」の間には大きな乖離がある．しかし乖離を数量的に評価することは難しい．**

- 攻撃の動機としても様々なものが考えられる
  - 通販，宣伝などの目的で他のデータベースとのマッチング
  - 攻撃を自己目的とした攻撃
  - 調査個体の「関係者」による攻撃
- 「論理的識別可能性」には実際的な意味はない?
  $\rightarrow$ (ほとんど唯一の) 開示リスクの客観的尺度として重要

- 論理的識別可能性が十分低ければ攻撃をあきらめるであろう
- 論理的識別可能性は統計モデルで数値的に評価できる
  → ただし推測の問題としては非正則で難しい面がある

母集団と標本 (統計調査特有の基本概念)

- 一意性
  - 個票データ中の**母集団一意**
    キー変数の組み合わせによって，母集団で一人しかいない個体
  - **標本一意**: 標本中での一意

- **標本で一意であっても，母集団で一意とは限らない**
- **特に抽出率 n/N が小さい時は，母集団で一意になくても標本では一意になる可能性が高い**
- **推定問題としての定式化: 標本一意の中で母集団でも一意なものはどのくらいあるか?**

- **単純無作為抽出のもとでは，母集団一意の個体も同じ抽出率で抽出される**
- **標本中の母集団一意数の推定と，母集団中の母集団一意数の推定はほぼ同値**

母集団一意数推定のためのモデル

- 開示リスク評価においては，すべてのキー変数を離散化して考えてもよい
- 個票データ自身を多元の分割表と同一視することができる
- 用いるモデル: 離散分布のモデルや分割表のモデル
- 統計的生態学や計量言語学でも同様のモデルが用いられる
- 「稀少種」「種の多様性」，　「稀な単語」「語彙」

- **一意: “珍しい個体”．個票開示問題では母集団での珍しい個体の数の推定となる．**

Size index (寸法指標)

K : セル総数

$F_j,\ j = 1, \ldots, K$ は各セルの頻度

$$S_i = \sum_j I(F_j = i),$$

$i = 0, 1, \ldots :$ サイズ i のセル数

まずは，セルのラベルを無視して，寸法指標の分布を考えるモデルが簡便である．

確率分割のモデル

- **ポアソン・ガンマモデル** (Bethlehem et al.(1990))
  ＝ 負の 2 項分布
  各セルが独立に i.i.d. **で負の 2 項分布となる**
- **多項・ディリクレモデル, 対数級数モデル，**Ewens sampling formula (Hoshino and Takemura (1998))
- Pitman sampling formula (Hoshino(2001))
- Engen's extended negative binomial model (Hoshino(2005)) **とより一般の分布族**

- 実際の推定値は仮定するモデルにかなり依存する
  → 基本的に，稀少な事象の確率はデータからでは推定が困難
  → モデルの想定に依存した解となる
- ポアソン・ガンマモデル: (おそらく) 過小推定気味
- Pitman sampling formula: (おそらく) 過大推定気味
- Engen's extended negative binomial model は Pitman sampling formula と似た性質を持つ

個体ごとの識別リスクの評価

- 上記の超母集団モデルは個票データ全体に含まれる母集団一意数の推定に用いられる
  $\rightarrow$ どの個体がより危険なのかという問題が残る
- 個体ごとの識別リスクの評価

いくつかのアプローチ

1. 各変数あるいは少数の変数の組み合わせについて外れ値に注目する (当然の常識的なチェック)

2. **モデルを用いるもの:**
   - **個票データを多元の分割表と見て，セルの生起確率を推定する**
   - **生起確率の小さいセルに観測値があると危険である**
   - **対数線形モデルを用いたアプローチ:** Skinner and Holmes(1998)
   - **分解可能モデルを用いたアプローチ (第 II 部で述べる)**

- 「構造的ゼロ」 の扱いが問題
  構造的ゼロ : 定義上観測値が現れないセル

3. “最小危険集合” (Willenborg and de Waal (1996)), “指紋”.
   - ある標本一意の個体が少ない数の変数ですでに標本一意であればより危険と考えられる
   - 個体が標本一意となる最小数の変数の集合
   - 理論的な性質が Takemura(2002a) で調べられている

局所秘匿の手法

- 個票データがそのまま提供するのには危険だと判断された場合には，秘匿処理を施す必要がある
- 標準的な処理: “大域的再符号化”(個票データ全体にわたってカテゴリーをより粗くする処理)
- 大域的再符号化では必要以上に分布の情報が失われる可能性がある

- その場合，局所的・撹乱的な秘匿処理が有効である

  → 欠測化，ノイズの付加，swapping (観測値の交換)，局所再符号化，ランダム化等

- PRAM (Post RAndomization Method) は有望なランダム化の手法である．randomized response と同様の考え方．

- 局所再符号化とスワッピングは Takemura (2002b) で論じられている

  → 似た個体をペアにする

  → ペア内で観測値を交換したりカテゴリーを統合する

- 局所再符号化の場合には，ユーザ自身が値をランダムに選ぶ．これは swapping をデータの提供者ではなく利用者がおこなうことになる．
- 問題点: 局所秘匿処理を施した後の開示リスクの手法が確立されていない

その他の話題

表形式のデータの秘匿

- 周辺和の情報が与えられていることから，秘匿したセルの情報がわかってしまう場合がある
- 1 次秘匿と 2 次秘匿
- セルの値が非負であることから，秘匿したとしても範囲が狭くなってしまう場合がある

オンライン検索システムにおける秘匿

- 一見安全な検索であっても，複数の検索を組み合わされると個体が識別される可能性がある
- 複数のユーザが結託する可能性もある
- 検索の度に撹乱する方式も有効であるが，繰り返し検索により平均化されることを考慮する必要がある
- 概念的には，安全な個票データを先に作っておいて，広い検索を許す方式がよいと思われる．

第 I 部のまとめ:

- 個票開示問題の研究はかなりの進展をみせているが，問題自体が難しいために，更なる発展が必要である．
- 日本の統計当局は他の国の経験を参考として，積極的に個票データの提供を推進すべきである (組織やガイドラインの整備)．
- 個票データの提供を継続的にモニターし，柔軟に対応していく必要がある．

第 I 部の参考文献 (やや古くなっています)

- **Baayen, R. H. (2001).** *Word Frequency Distributions.* **Kluwer, Dordrecht.**
- **Bethlehem, J. G., Keller, W. J. and Pannekoek, J. (1990). Disclosure control of microdata.** *Journal of the American Statistical Association*, **85, 38–45.**
- **Defays, D. and Anwar M. N. (1998). Masking Microdata Using Micro-Aggregation.***Journal of Official Statistics*, **14, 449–461.**
- **Domingo-Ferrer, J. (ed.) (2002).** *Inference Control in Statistical Databases. From Theory to Practice.* **Springer LNCS 2316, Berlin.**
- **Doyle, P., Lane, J. I., Theeuwes, J. J. M. and Zayatz, L.V. (2001).** *Confidentiality, Disclosure and Data Acess: Theory and Practical Applications for Statistical Agencies.* **Elseview, 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.**
- **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 assessment.** *Journal of Official Statistics*, **17, 499–520.**
- **Hoshino, N. (2005). Engen's extended negative binomial model revisited.** *Annals of the Institute of Statistical Mathematics*, **57, 369–387.**
- **Hoshino, N. and Takemura, A. (1998). Relationship between logarithmic series model and other superpopulation model useful for microdata disclosure risk assessment.** *Journal of the Japan Statistical Society*, **28, 125–134.**
- **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.**
- **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. (2002a). Minimum unsafe and maximum safe sets of variables for disclosure risk assessment of individual records in a mocridata set,** *Journal of the Japan Statistical Society*, **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.**
- **竹村彰通 (編) (2003). 特集「個票開示問題の統計理論」.『統計数理』第 51 巻第 2 号. 統計数理研究所.**
- **竹村彰通 (2003). 個票開示問題の研究の現状と課題. 『統計数理』第 51 巻第 2 号. 241–260. 統計数理研究所.**
- **Willenborg, L. and de Waal, T. (1996).** *Statistical Disclosure Control in Practice.* **Lecture Notes in Statistics 111, Springer, New York.**
- **Willenborg L. and de Waal, T. (2001).** *Elements of Statistical Disclosure Control.* **Springer, New York.**

第II部: 分解可能モデルに基づく秘匿措置

1. 多元分割表と階層モデルの記法
2. アメリカのセンサス個票データの例
3. センサス個票データでの計算結果
4. スワッピングによる秘匿
5. 周辺頻度を固定したスワッピングのための条件

多元分割表と階層モデルの記法

- m 元の分割表: $I_1 \times \cdots \times I_m$ 分割表
  - セル頻度: $f(i_1, \ldots, i_m)$
  - 多重添字 $i = (i_1, \ldots, i_m) \in \mathcal{I}$
  - セル頻度 $f(i)$
  - $I_1 \times \cdots \times I_m$: 総セル数は m とともに指数的に大きくなる $\Rightarrow$ 頻度ゼロのセルが出てくる

- 確率の導入
  - $p(i_1, \ldots, i_m) = p(i)$: セル i の確率 (生起確率)
  - 多項分布のモデル
  - (無条件の) 生起確率の推定値 $=$ 相対頻度

$$\hat{p}(i) = f(i)/n$$

  - 観測頻度がゼロのセルは確率が 0 か?
    “ゼロカウント問題” ← 正のはず
  - 観測頻度が 1 のセルの確率は $1/n$ か?
    個票開示問題の関心 ← おそらくもっと小さい

- 階層モデルの記法
  - m 個の変数の集合: $\Delta = \{1, \ldots, m\}$
  - D-周辺セル i_D とその周辺頻度 $f(i_D)$
  - $D \subset \Delta$ に対して $\mu_D : \mathcal{I} \to R$ を D に含まれる変数のみに依存する関数とする
  - 生成集合族: $\mathcal{D} = \{D_1, \ldots, D_r\}$ 互いに包含関係にない変数の部分集合の族
    「周辺表を公開する変数群の族」
  - 階層モデル

$$\log p(i) = \sum_{D \in \mathcal{D}} \mu_D(i)$$

- **分解可能モデル**
  - 階層モデルの中で特に都合のよいサブモデルとして分解可能モデルがある
  - 弦グラフ (chordal graph, triangulated graph) が長さが 4 以上の閉路を持たないグラフ
  - クリーク: G の (極大な) 完全部分グラフ
  - 分解可能モデル: $\mathcal{D} =$ 弦グラフ G のクリークの族 とするモデル

– 分解可能モデルの最尤推定値は，G が連結の場合には

$$\hat{p}(i) = \frac{1}{n} \frac{\prod_{C \in \mathcal{C}} f(i_C)}{\prod_{S \in \mathcal{S}} f(i_S)^{\nu(S)}}$$

と有理式で明示的に書ける．ただし $\mathcal{S}$ は G の minimal vertex separator の集合

– モデルの自由度も明示的に書ける

$$\sum_{C \in \mathcal{C}} \prod_{\delta \in C} I_\delta - \sum_{S \in \mathcal{S}} \nu(S) \prod_{\delta \in S} I_\delta$$

– 対数尤度から自由度を引くことによって AIC も簡単

- **問題点**
  - $m = 8$ 程度でも分解可能モデルはたくさんある．m とともに弦グラフの数は急速に増える
  - 多くのモデルの中でどのモデルを選ぶか，モデル選択の基準が難しい．観測ゼロが多い状況では AIC の理論的正当化も難しい．

Table 1: 階層モデル，グラフィカルモデル，分解可能モデルの個数

m	階層	グラフィカル	分解可能
2	2	2	2
3	9	8	8
4	114	64	61
5	6894	1024	820
6	7785062	32768	18154
7	2414627396434	2097152	617675
8	561304372093703203599 66	268435456	30888596

アメリカのセンサス個票データの例

- 今回のテストデータ: ワシントン州の 1% 抽出個人データから，項目を $m = 8$ 項目, 個人を無作為に $n = 9809$ 人再抽出．

1. 続柄 (14 分類)	2. 性別 (2 分類)
3. 年齢 (91 分類)	4. 配偶関係 (5 分類)
5. 出身地 (14 分類)	6. 配偶者の有無 (7 分類)
7. 実子の有無 (2 分類)	8. 実子の年齢 (5 分類)

- $m = 8$ 元の

$$14 \times 2 \times 91 \times 5 \times 14 \times 7 \times 2 \times 5 = 12485200$$

(約 1200 万セル) 型の分割表

- 頻度の頻度 (「サイズインデックス」)

セルサイズ	1	2	3	4	5	
頻度	2243	524	275	132	104	
	6	7	8	9	10	11 以上
	60	59	34	46	19	124

センサス個票データでの計算結果

- Ewens モデル, Pitman モデルのあてはめはサイズインデックスのみに依存する.
- これらの母集団一意数の推定値は

  Ewens モデル: 6,　　Pitman モデル: 214

- 分解可能モデルのあてはめ: 無作為に 1 万個程度の分解可能モデルを推定して，AIC の高いもの 5 個程度を選んで推定結果を見る

	モデル 1	モデル 2
AIC/2	13869.07	13984.97
対数尤度	−12141.07	−12013.97
自由度	1728	1971
母集団一意数	48.867	40.515
クリーク	{1,2,6},{1,6,7},{2,6,8}, {3,6,7},{4,6},{5,6}	{1,6,7},{3,6,7},{1,6,8}, {2,8},{4,6},{5,6}
セパレータ	$\{1,6\},\{2,6\},\{6,7\},\{6\}^2$	$\{1,6\},\{6,7\},\{6\}^2,\{8\}$

	モデル 1	モデル 2
セル確率	頻度	頻度
$10^{-2} \sim 10^{-3}$	0	0
$10^{-3} \sim 10^{-4}$	352	351
$10^{-4} \sim 10^{-5}$	1092	1117
$10^{-5} \sim 10^{-6}$	599	600
$10^{-6} \sim 10^{-7}$	179	158
$10^{-7} \sim 10^{-8}$	19	15
$10^{-8} \sim 10^{-9}$	2	2
$10^{-9} \sim 10^{-10}$	0	0

スワッピングによる秘匿

- 変数値のスワップ

  (男性, 年齢 55, 看護士, 東京在住),

  (女性, 年齢 50, 警察官, 大阪在住)

  のような二つの「危ない」レコードがあったとして，
  「職業」を入れ換える:

  (男性, 年齢 55, 警察官, 東京在住),

  (女性, 年齢 50, 看護士, 大阪在住)

- **分割表の頻度で考えてみると**

$$f(\text{男性}, \text{年齢}\ 55, \text{看護士}, \text{東京在住}) \rightarrow -1$$
$$f(\text{女性}, \text{年齢}\ 50, \text{警察官}, \text{大阪在住}) \rightarrow -1$$
$$f(\text{男性}, \text{年齢}\ 55, \text{警察官}, \text{東京在住}) \rightarrow +1$$
$$f(\text{女性}, \text{年齢}\ 50, \text{看護士}, \text{大阪在住}) \rightarrow +1$$

“primitive move”

- この例では，職業という単一の変数のみを入れ換えており，1 次元の周辺頻度はすべて不変であるが，職業を含む 2 次元の周辺頻度は変わってしまっている．
- もし職業と年齢を同時に入れ換えると

  (男性, 年齢 50, 警察官, 東京在住),

  (女性, 年齢 55, 看護士, 大阪在住)

  となり，1 次元の周辺に加えて，(年齢，職業) の 2 元の周辺も保存される．

- **個票データ (変数はすべて離散化済み)**

個体＼変数	1	$\ldots$	m
1	x_{11}	$\ldots$	x_{1m}
$\vdots$	$\vdots$		$\vdots$
n	x_{n1}	$\ldots$	x_{nm}

- **個票データの第 α レコード**

$$x_\alpha = (x_{\alpha 1}, \ldots, x_{\alpha m})$$

- **第 α レコードと第 β レコードで，最初の k 個の変数を** swap:

$$x_\alpha = i = (i_1, \ldots, i_m), \quad x_\beta = j = (j_1, \ldots, j_m)$$

$$\Rightarrow$$

$$x'_\alpha = i' = (j_1, \ldots, j_k, i_{k+1}, \ldots, i_m),$$
$$x'_\beta = j' = (i_1, \ldots, i_k, j_{k+1}, \ldots, j_m)$$

周辺頻度を固定したフワッピングの条件

- すでに公表されている周辺表のセット
  $\mathcal{D} = \{D_1, \ldots, D_r\}$ の頻度をすべて固定して
  swapping ができるか?
- **重要な観察**: 二つの周辺 $D_t, D_{t'}$ について
  $D_t \cap D_{t'} \neq \emptyset$ であり，かつ

$$i_{D_t \cap D_{t'}} \neq j_{D_t \cap D_{t'}}$$

  であるならば，D_t のみを swap し，$D_{t'}$ を swap しないということでは矛盾が生じてしまう．この場合，両方を swap するか，両方とも swap しないかのいずれかでなければならない．(運命共同体)

- 逆に $D_t \cap D_{t'} = \emptyset$ であるか，あるいは
  $D_t \cap D_{t'} \neq \emptyset$ であっても

$$i_{D_t \cap D_{t'}} = j_{D_t \cap D_{t'}}$$

  であれば，これら二つの変数群については，D_t のみを swap することと $D_{t'}$ を swap することは独立に決めることができる．

- 分解可能モデルだと判定が非常に容易となる．

- センサス個票データについては，以上の判定法を用いると，危険と判断される個体も swap 可能であった．

第 II 部の参考文献

- **Conditions for swappability of records in a microdata set when some marginals are fixed", by Akimichi Takemura and Hisayuki Hara.** *Computational Statistics*, **22, 173–185.**

- **Evaluation of per-record identification risk and swappability of records in a microdata set via decomposable models", by Akimichi Takemura and Yushi Endo. Technical Report METR 06-17, March 2006.**

第 II 部のまとめ:

- 個票開示問題と多元分割表の接点でも研究の余地がある
- 個票開示問題の研究から，分割表解析の新たな視点も得られる