資料１

データの開示リスクについて

竹村 彰通　　東大情報理工

2008 年 5 月 16 日

目次

はじめに

- 筆者はここ 10 年くらいにわたって，個票開示問題の数理統計学的側面に関する研究をおこなってきた．
- 個票開示問題には，数理統計的側面のみならず，法律的側面，文化的背景，IT 技術の進歩への対応など，さまざまの側面がある．
- 数理的側面はその一部．ミクロデータを提供には総合的な判断も必要とされる．
- 安全性と有用性のトレードオフが基本的な問題

- 諸外国では，ミクロデータの提供についてすでに長い経験がある．これに対して，わが国ではこれまで制度的な制約もありほとんどおこなわれてこなかった．
- まずは諸外国の基準にみあう形で，ミクロデータを提供を進めるべきである．
- 一方で，基準の客観性のためには数理的側面も重要．
- 以下の資料では，数理的研究に重点をおいた形で，開示リスク評価と秘匿措置のテクニックを紹介する．(以前の研究集会での発表資料のため数学的内容がメイン)

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
  $\rightarrow$ ただし推測の問題としては非正則で難しい面がある
- 基本概念: 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

$S_i =$ 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．

- **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 (Takemura(1999), Hoshino and Takemura(1998))
- Pitman sampling formula (Hoshino(2001))
- Engen's extended negative binomial model (Hoshino(2003)) **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:** Fienberg and Makov(1998), Skinner and Holmes(1998)
   - **より簡便な加法モデルを用いたアプローチ:** Takemura(2002c)
   - **分解可能モデルを用いたアプローチ (竹村・遠藤)**

– 「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
- 実用的には大域的再符号化のレベルをいくつか定めておき，レコードによってレベルの深さを変える方法がわかりやすい．

- 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．
- いくつかの周辺表を保存したままで，スワッピングをおこなう方法が竹村・原 (*Computational Statistics*, Vol.22, 173-185 (2007)) で提案された．
- 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