疑似ミクロデータの教育・研究での利用について

公的統計のミクロデータの利用に関する研究集会

2012年11月16日　統計数理研究所

岡山商科大学経済学部　　佐井至道

待望の疑似ミクロデータ

- 目的外使用による調査票情報
- 二次利用による匿名データ，オーダーメード集計
- 上記のデータとともに，比較的自由に使用できる疑似ミクロデータの提供が，研究者からも強く要望されていた
- 新統計法の施行からわずか2年で試行的提供を始めたことに感謝します

私自身の研究上の興味（1）

- 官庁統計の個票データ（ミクロデータ）の秘匿やリスク評価に関する研究を行っている
- 理論的な研究が中心であるが，手法を検証するために実データに対する適用が不可欠
- 調査票情報の目的外使用は，使用期間の制約や，事前の目的の決定が必要
- 手法の検証において，疑似ミクロデータはミクロデータの代わりとなり得るか？

私自身の研究上の興味(2)

- 科研費の目的として，疑似個票データ(疑似ミクロデータ)作成に関する研究も掲げている
- 疑似ミクロデータの作成方法そのものにも興味がある
- 秘匿方法は適切か？ 他に代わる方法はないか？
- 疑似ミクロデータに相当するものとして，海外では PUMS，IPUMS が有名

PUMS (Public Use Microdata Samples)

- アメリカ合衆国のセンサスの疑似ミクロデータ
- 2000年，1990年の1%，5%抽出データなどが無料でダウンロード可能

  http://www2.census.gov/census_2000/datasets/PUMS/

  http://www2.census.gov/census_1990/
- トップコーディングなどの嘘をつかない秘匿措置の他，スワッピングなどの嘘をつくタイプの秘匿措置も使われている

PUMS (Public Use Microdata Samples)

- データは州単位にまとめられている
- ワシントン州の1%抽出データで約24MB
- 個人レコード（先頭Ｐ）は，フラグを含めて160程度の項目からなる
- Excel などでの利用のためには，変数の切り出しが必要
- 2010年のセンサスについては2012年12月以降に提供開始の予定

IPUMS-International website

- Minnesota Population Center がホスト

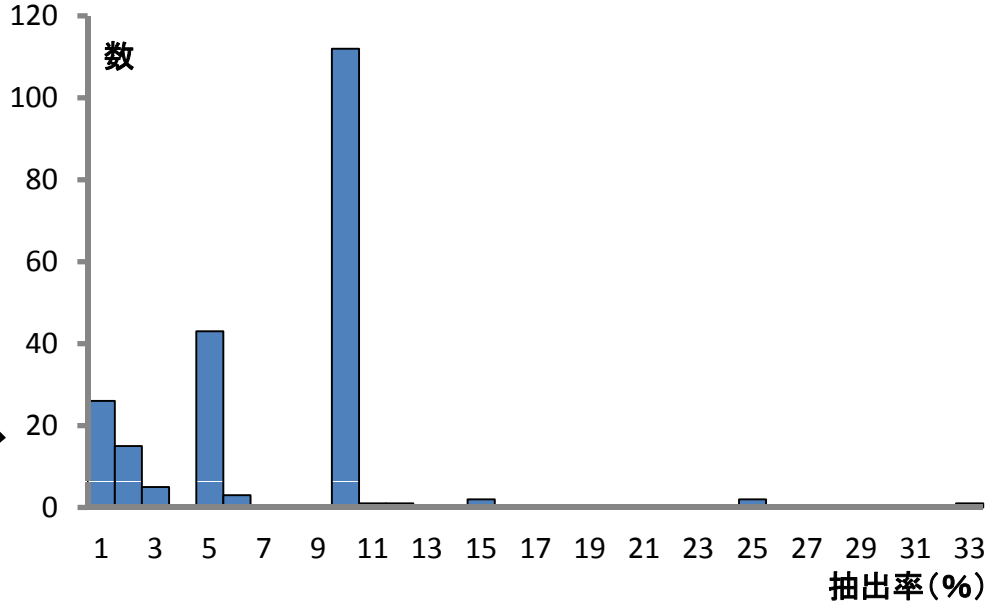
  https://international.ipums.org/international/
- 無料でダウンロード可能
- 比較的簡単な登録が必要
  - 氏名，メールアドレス，研究分野，職業，結果の発表方法，所属機関の名称　など
  - 75words 以上の研究計画も必要
  - 研究計画で 1/3 は拒否される

IPUMS-International website

- 2012年10月末現在，68カ国（211センサス）の疑似ミクロデータを利用することが可能
  - 1年半前よりも13カ国増加
  - アメリカ合衆国のデータは1960年のものから
  - アジアの国も多い

抽出率：0.06%～33%

5%以上：164センサス



岡山商科大学･経済学部 について

- 統計学関係の講義は多い

  統計学総論Ⅰ･Ⅱ，経済統計論Ⅰ･Ⅱ，
  市場調査論，計量経済学Ⅰ･Ⅱ，
  経営統計学Ⅰ･Ⅱ，社会調査実践，
  経済データ分析　など
- 統計手法の習得の講義が多く，実データに対する実証分析は十分行われていない
- 疑似ミクロデータはゼミで使用（予定も含む）
  - 3年のゼミ：約10名
  - 大学院修士課程のゼミ：1名

疑似ミクロデータ利用のメリット

- そもそもデータを勘違いしている学生が多い
  - 生データを見る機会が少ない
  - 調査を頻繁に行えればよいが時間的には難しい
- 分析の方針から自分で考えるきっかけとなる
  - 「問題があるから分析をする」からの脱却
- 分析に対するモチベーションも高まる
  - 海外のデータよりは国内のデータの方が身近で興味は高い

使用した疑似ミクロデータ

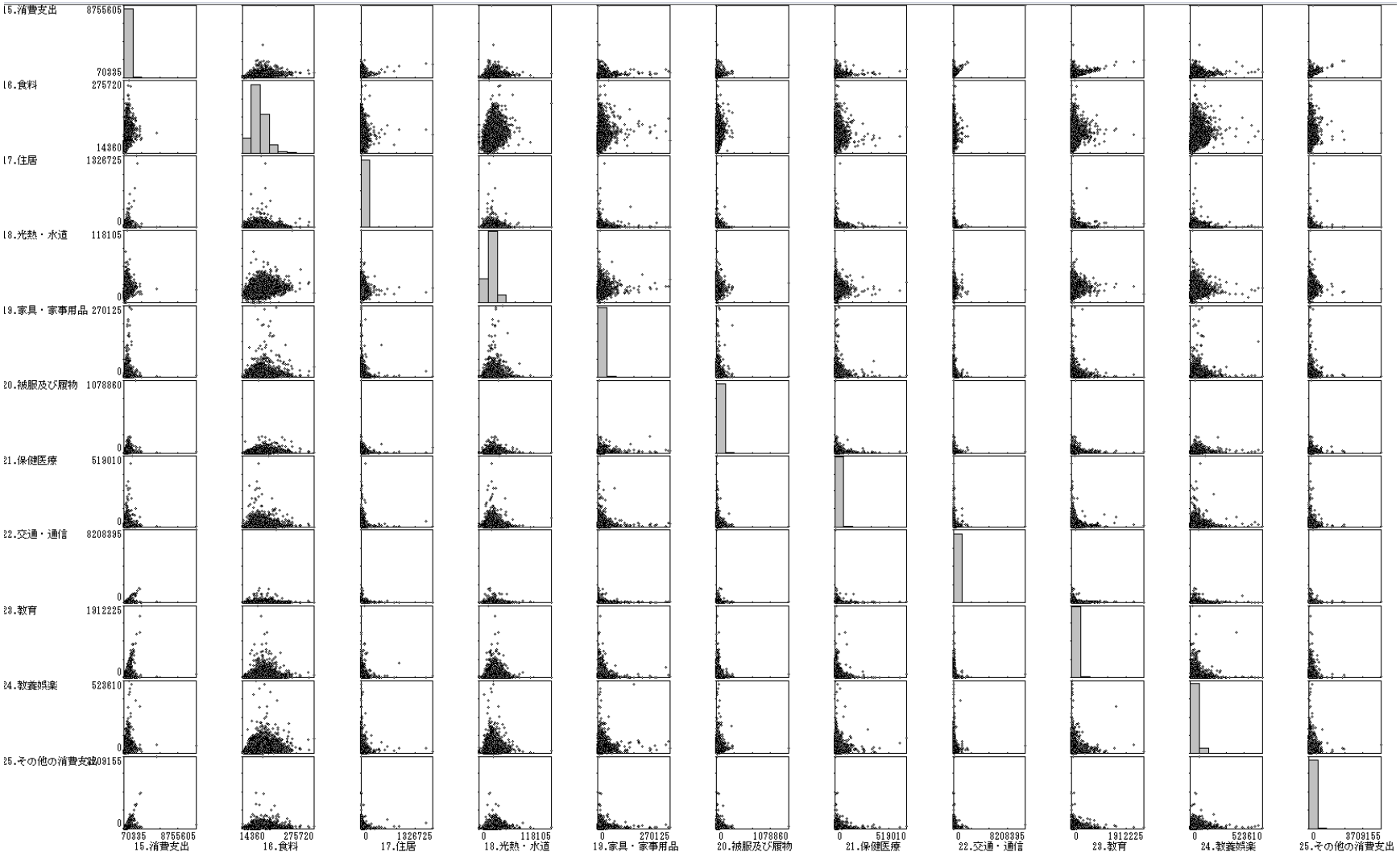
	レコード数	世帯属性	支出項目	収入項目
大規模データ	32,027 二人以上の勤労者世帯	14項目	149項目	34項目
簡易データ	8,333 世帯人員:4名 有業人員: 1～2名	14項目	11項目	なし

※ 疑似ミクロデータ利用の手引より

3年ゼミでの教育内容

- 変数間の相関関係に関する分析（相関図の作成，相関係数の計算など）
- 重回帰分析，数量化理論Ｉ類
  - 基準変数：何にすべきか？
  - 説明変数：5個程度（カテゴリ分け可）
  - 自由度調整済み重相関係数を最大とすることを目的とする
- 分割表の作成と独立性の検定

支出項目間の相関図(簡易データ)



重回帰分析での利用

- アメリカ合衆国のセンサスのデータで，「基準変数：年間収入」，「説明変数：5個」として，自由度調整済み重相関係数を最大化する
- これまでの最大値は 0.36641
  - 性別：2カテゴリ（質的）
  - 回復可能障害：2カテゴリ（質的）
  - 通勤手段：4カテゴリ（質的）
  - 週労働時間：量的 → 4カテゴリ（質的）
  - 労働週数：量的

重回帰分析での利用（簡易データ）

- 簡易データで「基準変数：消費支出」，「説明変数：支出項目5個」として自由度調整済み重相関係数を最大化すると・・・
- 最大値は 0.97335
  - 交通通信（支出）：量的
  - その他の消費支出（支出）：量的
  - 教育（支出）：量的
  - 食料（支出）：量的
  - 住居（支出）：量的

分割表の作成と検定(大規模データ)

- 住居の建て方:一戸建
- 住居の所有関係:持ち家(世帯員名義)

設備修繕・維持への支出

住居の構造	0円	～1000円	～10000円	10001円～	計
木造	5100	2109	3345	1486	12040
防火木造	1403	618	847	310	3178
鉄筋コンクリート	460	187	211	85	943
その他	1	0	4	1	6
不詳	1713	666	953	371	3703
計	8677	3580	5360	2253	19870

分割表の作成と検定（大規模データ）

世帯人員：2人　　パンの支出

米の支出	平均未満	平均以上
平均未満	4705	815
平均以上	1485	433

世帯人員：3人　　パンの支出

米の支出	平均未満	平均以上
平均未満	4057	1750
平均以上	1528	1202

世帯人員：4人　　パンの支出

米の支出	平均未満	平均以上
平均未満	3263	3087
平均以上	1286	2308

世帯人員：5人～　　パンの支出

米の支出	平均未満	平均以上
平均未満	1125	1431
平均以上	1128	2424

大学院のゼミでの教育内容

- 個票データの秘匿措置の研究
- 個票データのリスク評価方法の研究
- 個票データの有用性の評価方法の研究
- 疑似個票データの作成方法の研究

- 「秘匿措置や有用性の評価方法を限定した状況で，公開のリスクを測る」という研究の題材に利用できないかと考えている

個票データのリスク評価方法

キー変数：第三者が既に情報を持っていて，個体（個人，世帯，事業所など）を特定する際に用いられる変数

セル：すべてのキー変数の組み合わせ

（例）性別と年齢階級がキー変数の場合

年齢階級

性別

	15-19	20-29	30-39	40-49	50-59	60-69	70-
男性	2	1	1	4	0	1	3
女性	1	0	2	3	3	2	1

個票データのリスク評価方法

- 個票データが標本調査で得られている場合には母集団に関する推定が必要
- セルごとの推定は難しい

母集団

	15-19	20-29	30-39	40-49	50-59	60-69	70-
男性	?	?	?	?	?	?	?
女性	?	?	?	?	?	?	?

標本

	15-19	20-29	30-39	40-49	50-59	60-69	70-
男性	2	1	1	4	0	1	3
女性	1	0	2	3	3	2	1

個票データのリスク評価方法

- 標本寸法指標からの母集団寸法指標の推定

標本

	15-19	20-29	30-39	40-49	50-59	60-69	70-
男性	2	1	1	4	0	1	3
女性	1	0	2	3	3	2	1

s_l : l 個の個体が入っている標本のセル数

⇨ 標本寸法指標 : $(s_1, s_2, s_3, s_4) = (5, 3, 3, 1)$

- 母集団寸法指標は S_l という記号を用い，標本寸法指標を基に推定する

個票データのリスク評価方法

- 母集団寸法指標の推定に用いられる方法
  - ピットマンモデルなどのモデルを用いる推定法
  - 制約付きのノンパラメトリック推定法
- 寸法指標を用いるとセルの区別はできなくなるが, $S_1 \cdot \lambda$ によって母集団かつ標本一意の個体数が推定できる（λ は抽出率）

疑似ミクロデータの寸法指標

- 質的属性14属性（性別，年齢5歳階級，世帯区分など）で高次元集計表を作成
- セルで度数1，2のレコードは，属性の一部の値を不詳に置き換え，度数3以上に処理
- 上記の属性はキー変数と一致する

	処理前	処理後
度数1のレコード数	22,583	0
度数2のレコード数	5,806	0
度数3以上のレコード数	26,667	32,027
セル数	28,481	8,261

※ 処理後は勤労者世帯のみ

疑似ミクロデータの寸法指標

（サイズ30まで）

サイズ	s_l
1	0
2	0
3	5707
4	1370
5	482
6	227
7	130
8	72
9	51
10	41

サイズ	s_l
11	25
12	24
13	25
14	20
15	8
16	14
17	13
18	9
19	2
20	3

サイズ	s_l
21	3
22	3
23	3
24	1
25	0
26	2
27	5
28	0
29	2
30	0

疑似ミクロデータの作成方法に関して

- 疑似ミクロデータの作成途中で，一度高次元の分割表（集計表）を作成している
- 統計法に照らして適切な方法で，データは安全と考えられる
- 統計法の解釈とは別に，スワッピングなどを用いた，他の方法の検討も必要では？
- 嘘をつくタイプの秘匿措置が施されたデータに対するリスク評価手法の開発も必要
  - ➢このような研究については，研究者が取り組むべき

今後の利用に関して（おわりに）

- 目的外使用が，今後より厳しくなる可能性があるとのこと
  - 調査票情報とデータの形式が同じで，しかも分析結果がある程度近いデータの重要性は増す
- リスク評価を研究している者としては，秘匿措置の施されていないキー変数のみのデータが欲しい
- 集計用乗率を調整したデータが欲しい
- 提供する調査の拡充をお願いします