「立教大学教育用疑似匿名データの作成と教育利用事例について」

立教大学社会情報教育研究センター
政府統計部会

公的統計の二次利用に関する教育コンテンツ・教育用匿名データの作成目的と概要

・ミクロデータを用いたデータ解析のリテラシーを養うためには、実際のデータを扱う演習を通じて、集計及び統計的分析に関する理解を深めることが必要である。しかしながら、学部学生等の初学者に対するデータ解析の教育を考えた場合、匿名データは膨大な情報を含むことから実際のデータ利用においてはデータ内容の確認と分析に適した項目の絞り込みに悩む学生が多いことが想像され、限られた講義の中では、肝心のデータ解析まで至らず、十分な分析が行えない可能性がある。

・また、統計の初学者に対しては、統計データを扱う際に知っておくべき法制度面や倫理面の教育も必要である。

・新統計法の下では匿名データが高等教育目的で利用可能になり、高等教育機関又はそれに所属する教員が学生や大学院生への講義や演習を行う目的で匿名データを利用することが可能になったが、特に学部教育の場面では現実的ではない。

（申請の煩雑・困難さ、利用者の限定性（申請時に利用者を確定する必要）など）

・教育用の疑似データなら、学部の学生に直接利用させる教育形態も可能

・このようなことから、初学者がデータに慣れ親しみながら、必要な知識を得るという観点から、統計データの二次利用の仕組みや手続きについて理解させるとともに、項目数を限定した教育用擬似ミクロデータを用いた集計及びデータ解析の学習をさせることで匿名データを扱うための基礎的能力を養うことを目的とした教育を行うことが必要と考えられる。

（学習コンテンツと教育用疑似匿名データの関係）

・本センターが作成した学習コンテンツ『教育用擬似匿名データを利用した学習コンテンツ「公的統計の二次利用とデータ分析入門」』は、

1）「公的統計二次利用制度の活用」（二次利用制度に関する学習マニュアル部分）と、

2）「ミクロデータ統計分析入門」（統計分析実習部分）から構成される。

・今回作成した「教育用疑似匿名データ」は上記学習コンテンツ 2) の各章における解説や、練習問題、演習で、利用データとして指定されている。

※上記 1）コンテンツは、（財）統計情報研究開発センターと、上記 2）コンテンツは法政大学日本統計研究所と共同で開発した。

（作成データ）

１）教育用匿名データ（全国消費実態調査平成 16 年単身世帯）

２）教育用匿名データ（全国消費実態調査平成 16 年二人以上世帯）

３）教育用匿名データ（社会生活基本調査平成 13 年時間編）

４）教育用匿名データ（就業構造基本調査平成 14 年）

※（全国消費実態調査平成 16 年単身世帯）データは 3396 レコード

（全国消費実態調査平成 16 年二人以上世帯）データは 43861 レコード

（社会生活基本調査平成 13 年時間編）データは 285855 レコード

（就業構造基本調査平成 14 年）データは 752068 レコード

それぞれ、データ・符号表・レイアウト表を作成

２．教育用疑似データ作成方針と方法論

（教育的観点からのデータ作成方針）

・初学者の理解が複雑にならない程度に変数は絞る一方、ある程度の多重クロス集計や多変量解析ができるだけの変数は残すため、公表統計表でよく使われている分類事項、集計事項を参考に変数を選択。

・推測統計の学習での利用を想定し、データ量が大量になったときのデータ処理技術を段階的に身につけさせることができるようにするため、データ量は原データと同じとする。

・教育用データはできるだけ実際のミクロデータに近似している方が望ましいが、教員があらかじめ教育用データの特徴を理解した上で学生等に利用させるのであれば、近似の程度が低いデータであっても教育効果は期待できるため、教育用データは必ずしも原データとの近似の程度が高くなくともよい。

・教育用データの形式は、匿名データに準じた形式にし、政府統計個票データレイアウト標準記法に沿って作成する。

- 匿名データの利用時に、符号表とデータレイアウトからデータ構造を理解するための基本的能力が身につけられるようにするため
- 統計データ利用に関して、公表統計表(&詳細集計表)→教育用データ→匿名データ→調査票情報の利用といった段階的な教育の一環としての位置付けを想定

（作成方法）

・原匿名データをいくつかの変数に着目して類別し、同値類となるデータ集合をさらに統計量を求める変数に着目して類似のレコードからなる部分集合(クラスター)に分割する。

・クラスターごとに統計量を作成(量的変数の場合は総計値、質的変数の場合は最頻値)し、量的変数の場合は総計値をクラスターに含まれるレコード数分に分割する。

・作成した統計量(又は分割した統計量)をクラスター内の各レコードに配分して疑似個別データを作成する。

３．教育用疑似匿名データの作成作業とデータの特徴（全国消費実態調査の例）

・公表統計表での分類事項を参考にし、世帯区分、地域、世帯人員の組み合わせで原データを類別する。同値類の中は支出総額を昇順にソートし、k レコードづつに分割する単一軸法ミクロアグリゲーションを適用。

- ✧ データの有用性の点では個別ランキング法がよいとされているが、量的変数に全消の収支項目のような上位加法性と収支バランスといった制約がある場合、収支バランスの保存ができず、乗率の扱いも複雑になるなど不向きなため、支出総額で家計収支を代表させることにし、乗率を含む全変数を統一的に扱える最もシンプルな単一軸法を採用した。

・どのクラスターも k 未満の大きさにならないようにする(最後のクラスターサイズは k または k+1～2k-1 となる)。クラスターサイズ k を 3,5,7,9 として年間収入などの量的項目の基本統計量等を求め、作成する教育用データのクラスターサイズは探索的に決定した。

（特徴点）

・平均値は乗率なしのデータでは変化しない。乗率ありのデータでは一部の項目を除き原データに比べ疑似個別データの平均値の方が小さくなる傾向がある。

・クラスターサイズが大きくなると標準偏差は小さくなるが、減少率は逓減の傾向。

・分布のピークは右にシフト、高くなる傾向。形状が大きく変わるものもある(負債現在高)支出総額の分布はほぼ一致。
→多重クロス表によっては原データとかい離大の可能性あり。

・分布の左裾はクラスターサイズが大きくなるに伴い右にシフトする(値が大きくなる)傾向があるが、右裾は左にシフトするものの必ずしも一定の傾向はみられない。

・疑似データではクラスターサイズが大きくなると変数間の相関係数が大きくなる傾向。
→多重共線性の可能性は高まっているかもしれない。

・原データと疑似個別データ間では同一変数間の相関は小さくなる傾向。異なる変数間では一定の傾向は見られない。

・作成した教育用データと原データを比べると、276 の相関係数のうち、符号が逆転した数は二人以上世帯で 9 であった(すべてマイナスからプラスに変化。ただし、原データの相関係数の絶対値は極めて小さい)。→おおむね相関構造は保存されていると言える。

４．教育用疑似匿名データの講義利用例

・学習テキストでの利用

学習コンテンツ「ミクロデータ統計分析入門」を作成し、各章の解説や、練習問題、演習問題などでの利用データとして指定した。

（参考：「**ミクロデータ統計分析入門**」目次）

序論　　この教材の活用にあたって

第 1 部

第 1 章　教育用ミクロデータの特徴/　第 2 章　調査票と統計データ/　第 3 章　データの読み込み/　第 4 章　解析結果を見やすくする工夫/　補論1　シンタックスによる実行/

第 2 部

第 5 章　基本統計量/　第 6 章　度数分布とヒストグラム/　第 7 章　散布図と相関/　第 8 章　クロス集計分析の基礎/　第 9 章　ケース選択によるクロス表の作成/　第 10 章　値の再割り当てを用いたクロス集計/　第 11 章　新変数の作成によるデータ解析（1）/　第 12 章　新変数の作成によるデータ解析（2）/　第 13 章　新変数の作成によるデータ解析（3）/　第 14 章　クロス集計の有意性検定/　第 15 章　クラスタ分析/　補論 2　外れ値とその処理/　補論 3　記述統計と乗率の利用

第 3 部

第 16 章　確率変数と確率分布/　第 17 章　正規分布/　第 18 章　標本理論と標本分布/　第 19 章　　統計的推定/　第 20 章　統計的検定(1) 平均値の検定/　第 21 章　統計的検定(2) クロス表の独立性検定/　第 22 章　単回帰分析(1) モデルとOLS/　第 23 章　単回帰分析(2) モデルと推定値の評価/　第 24 章　重回帰分析(1) ダミー変数の利用/　第 25 章　重回帰分析(2) モデルの比較検証/　第 26 章　ロジスティックス回帰モデル(1) モデルとML/　第 27 章　ロジスティックス回帰モデル(2) モデルと推定値の評価/　第 28 章　ロジスティックス回帰モデル(3) 推定結果の解釈/　補論 4　　推測統計と乗率の利用

第 4 部

第 29 章　プレゼン資料の作成（1）　分析結果を文書にまとめる/　第 30 章　プレゼン資料の作成（2）　パワポの作り方/　第 31 章　プレゼン資料の作成（3）　レジメの作り方/　第 32 章　いろいろなグラフとその利用/　第 33 章　レポートの書き方/　第 34 章　論文の書き方

・本学ならびに他大学（協力校）学部開講の講義での利用（2011-12 年度）

- 立教大学全学共通科目「統計情報で社会経済を診断する」、中央大学経済学部「入門統計演習」では、各種統計情報の学習の一環として「公的統計二次利用制度」を取り上げ、制度の概要や申請・利用方法を学習するとともに、その例として教育用疑似データを例示し、匿名データの特徴点や利用上の限界点などを学習した。
- 中央大学経済学部「データ分析演習」、東洋大学経済学部「統計分析論」では、上記の統計学習に加え、学生等は、教育用擬似データ上の項目を適宜選択し、集計表を作成して統計表の分析(作成した統計表に一定の傾向が見られるか否か、はずれ値の考察など)や、項目を適宜選択して重回帰モデルを作成し、統計的分析(変数選択が適切か否かの考察など)について学習し、レポート作成までを行った。

・その他、教育用匿名データ、「ミクロデータ統計分析入門」を活用した関連講義は以下の通り

立教大学経済学部・計量経済学（SPSS による分析実習に利用）
立教大学経済学部・調査実習（SPSS 実習の利用データとして）
立教大学全カリ・統計情報で社会経済を診断する（公的統計二次利用の例として解説）
中央大学経済学部・データ分析演習（SPSS による分析実習に利用）
中央大学経済学部・入門統計演習（公的統計二次利用の例として解説）
東洋大学経済学部・統計分析論（Excel ファイルを分析実習・レポートに利用）
東洋大学経済学部・経済データ分析（公的統計二次利用の例として解説）

・学習コンテンツに関する講習会の開催

教育用擬似匿名データを利用した学習コンテンツの講習会を開催

・SPSS・R に関する講習会での利用

立教大学社会情報教育研究センター・統計教育部会が開催する「SPSS 統計分析セミナー」内で参照データとして利用

（参考：「公的統計二次利用制度の活用」目次）

1　統計法とは
2　二次利用とは
3　調査票情報と匿名データの違いとは
4　匿名データ利用
5　オーダーメード集計とは
6　二次利用の仕方、手続きと手順、注意点など

参考：二次利用の申請形態別の特徴（平成 23 年 2 月現在）

以上