匿名データを用いた統計教育事例と直交変換による合成データ生成法の紹介

東京情報大学

佐野夏樹

概要

1. 住宅・土地統計調査の匿名データを利用した基礎分析教育の紹介
   - 変数の種類と尺度
   - 変数間の関連性の尺度
   - クラメールの連関係数が計算出来ないケース
2. 全国消費実態調査の匿名データを利用した研究紹介
   - 直交変換による合成データ生成法
   - 全国消費実態調査の匿名データによるリスク評価と有用性評価

変数の種類と尺度

質的変数と量的変数
- 質的変数：離散的な値をとる変数，カテゴリカル変数とも呼ばれる．
- 量的変数：連続的な値をとる変数

質・量	尺度の種類	特徴
質的	名義尺度	性別や職業の様にカテゴリーの違いだけを表す
質的	順序尺度	優・良・可・不可のように順序に意味があるが，カテゴリー間の差は同じではない
量的	間隔尺度	温度の様に，順序も間隔も意味があるが，原点に意味はない
量的	比率尺度	長さや重さの様に，間隔尺度であり，さらに原点に意味がある．

平成25年住宅・土地統計調査匿名データにおける変数の例（1）

名義尺度の例：住宅以外の建物の種類4区分

会社等の寮	学校等の寮	旅館・宿泊所	その他の建物	対象外
90	51	56	461	340971

順序尺度の例：子の居住地6区分

一緒に住んでいる	徒歩5分程度の場所	片道15分未満の場所	片道1時間未満の場所	片道1時間以上の場所	子はいない	不詳	対象外
123630	7049	12576	23315	27818	53742	45183	48316

平成25年住宅・土地統計調査匿名データにおける変数の例（２）

間隔尺度の例：建物の階数（うち一戸建・長屋建３区分）

1階建	2階建	3階建以上	対象外
41273	165291	6629	128436

比率尺度の例：住宅の延べ面積

回答者	不詳	対象外
286801	6379	48449

平均: 102.1868 m^2

問題：変数の種類と尺度

平成25年住宅・土地統計調査匿名データの変数を
（１）名義尺度（２）順序尺度（３）間隔尺度（４）比率尺度
のいずれかに分類せよ.

全変数の数（157変数）

- 政府統計コード, 一連番号等のメタ変数関係（８変数）
- 名義尺度（118変数）
- 順序尺度（13変数）
- 間隔尺度（3変数）
- 比率尺度（15変数）

代表的な関連性の尺度（１）

量的変数↔量的変数　　ピアソン相関係数

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$-1 \leq r \leq 1, \bar{x}: x$の平均,　$\bar{y}: y$の平均

質的変数↔質的変数　クラメールの連関係数

$$V = \sqrt{\frac{\chi_0^2}{n\{(\min(a, b) - 1\}}}$$

$$0 \leq V \leq 1$$

χ_0^2：カイ２乗値
n：サンプル数
a：分割表の行数
b：分割表の列数

代表的な関連性の尺度（２）

量的変数↔質的変数　相関比

質的変数 x	量的変数y	平均
x_1	$y_{11}, y_{12}, \cdots, y_{1n_1}$	$\bar{y}_1$
x_2	$y_{21}, y_{22}, \cdots, y_{2n_2}$	$\bar{y}_2$
x_3	$y_{31}, y_{32}, \cdots, y_{3n_3}$	$\bar{y}_3$
x_4	$y_{41}, y_{42}, \cdots, y_{4n_4}$	$\bar{y}_4$

$\bar{\bar{y}}$：総平均

総平方和S_T：$\sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{\bar{y}}\right)^2$

級間平方和S_A：$\sum_{i=1}^{a} n_i (\bar{y}_{i\cdot} - \bar{\bar{y}})^2$

相関比：$\eta^2 = \frac{S_A}{S_T}$

$0 \leq \eta^2 \leq 1$

平成25年住宅・土地統計調査匿名データに対するRを使った計算例

相関係数
１ヶ月当たり家賃・間代（量的）と１ヶ月当たり共益費・管理費（量的）

```
> cor(yachin,kyoeki,use="complete.obs")
[1] 0.215262
```

クラメールの連関係数
建物の階数_うち共同住宅・その他（9区分）（質的）とエレベーターの有無（質的）

```
> library(vcd)
> assocstats(table(kaisu,elevator))$cramer
[1] 0.8628889
```

相関比
１ヶ月当たり家賃・間代（量的）と建築の時期１４区分（質的）

```
> summary(lm(rent.fee~as.factor(year)))$r.squared
[1] 0.08834161
```

問題：クラメール連関係数

以下の３つのパターンの行列に対してクラメール連関係数を計算出来ない理由を計算式とデータの側面から考察し，対策を考えよ.

（A）

V9 \ V11	1	2	3	4	5
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0

（B）

V10 \ V39	1
	125609
1	0
2	0
V	0

（C）

V10 \ V11	1	2	3	4	5
	108102	106753	100192	24847	1077
1	0	0	0	0	0
2	0	0	0	0	0
V	0	0	0	0	0

パターン（A）

$$\chi^2 = \sum_{i=1}^{l} \sum_{j=1}^{k} \frac{(f_{ij} - \hat{f}_{ij})^2}{\hat{f}_{ij}}$$

f_{ij}：(i,j)セルの観測度数
$\hat{f}_{ij}$：(i,j)セルの期待度数
l：分割表の行数
k：分割表の列数

V11

V9	1	2	3	4	5
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0

計算出来ない理由
特定の行（列）が全て0だと，その行（列）確率が0になり，期待度数が0となるセルが生じるため．

欠損値NAも含めた分割表

V11

V9	1	2	3	4	5	<NA>
1	0	0	0	0	0	90
2	0	0	0	0	0	51
3	0	0	0	0	0	56
4	0	0	0	0	0	461
<NA>	108102	106753	100192	24847	1077	0

２つの質問項目の対象が異なり，対象外がNAとして処理されていると，このパターンになる．

V9：住宅以外の建物の種類4区分（対象：住宅以外の建物に居住する世帯）
V11：建物の構造5区分（対象：住宅）

パターン（B）

$$V=\sqrt{\frac{\chi_0^2}{n\{(\min(a,b)-1\}}}$$

$$0\le V\le 1$$

	V39
V10	1
	125609
1	0
2	0
V	0

計算出来ない理由
分割表の行（列）数が1だと
クラメール連関係数の平方
根中の分母が0となるため

欠損値NAも含めた分割表

	V39	
V10	1	<NA>
	125609	215887
1	0	60
2	0	70
V	0	3

「無し」に該当するカテゴリがなく，「無し」と対象外が区別されず，NAとして処理されている場合，このパターンになる.

V10：住宅以外の建物の所有の関係2区分
V39：高齢者等のための設備状況_手すりの有無

パターン（C）

V10 \ V11	1	2	3	4	5
	108102	106753	100192	24847	1077
1	0	0	0	0	0
2	0	0	0	0	0
V	0	0	0	0	0

欠損値NAも含めた分割表

V10 \ V11	1	2	3	4	5	<NA>
	108102	106753	100192	24847	1077	525
1	0	0	0	0	0	60
2	0	0	0	0	0	70
V	0	0	0	0	0	3

計算出来ない理由
特定の行（列）が全て0だと，その行（列）確率が0になり，期待度数が0となるセルが生じるため．

２つの質問項目の対象が異なり，片方の質問は，対象外をNAとして処理し，もう片方の質問項目は，対象外を空白として処理している場合，このパターンになる．

V10：住宅以外の建物の所有の関係2区分
(対象：住宅以外の建物に居住する世帯)
V11：建物の構造5区分（対象：住宅）

基礎分析教育のまとめ

- 住宅・土地統計調査の匿名データを用いて,名義尺度, 順位尺度, 間隔尺度, 比率尺度のデータとは, 具体的にどの様なデータか, 教育
- 変数の尺度によって, 基礎分析においても, 異なる分析手法を適用する必要があることを教育
- 単に, 分析パッケージ利用するだけでなく, NaN等の計算不可の出力結果から, 分析指標の意味やデータの性質について理解することの重要性を教育

直交変換による合成データ生成法

統計的開示制御の方法

分類	手法	概要
非撹乱手法	グローバルリコーディング	いくつかのカテゴリを結合し，新しいカテゴリーを作成する.
	局所抑制	特定に結びつく危険な組み合わせは，開示せず，抑制する.
撹乱手法	ノイズ追加	連続値のデータにノイズを付加し，異なるデータに撹乱する.
	PRAM	カテゴリカルデータを確率的に撹乱する.
	マイクロアグリゲーション	連続値のデータをクラスタリングし，同じグループに所属するクラスタは，クラスタの平均値で置き換える.
	データシャッフリング	連続値のデータに対して，原データの周辺分布と同じ周辺分布になることを保証しながら，異なるサンプル間のデータを取り替える.
合成データの生成	多重代入法	原データを多重代入法によって補間したデータで置き換える.

合成データの生成法

1. 完全合成データ：原データは無く, 合成データのみが公開される.
2. 部分的合成データ: 機密性の高い変数やサンプル（レコード）のみ合成データで置き換えられ, 公開される.
3. ハイブリッドデータ：原データと合成データを組み合わせて生成されるデータ. 組み合わせる割合によって, 原データ寄り, もしくは合成データ寄りのデータになる.

主成分分析

直交変換による多変量データの情報縮約手法

$$\boldsymbol{C} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T = \sum_{i=1}^{p} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$$

$\boldsymbol{C}$ ：共分散行列（$p \times p$）
$\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, \cdots, \lambda_p)$: 対角行列
$\boldsymbol{U} = (\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_p)$: 直交行列

主成分得点：新しい基底上の座標

$f_i = \boldsymbol{u}_i^T \boldsymbol{z},\ i = 1, \cdots, p$: 第 i 主成分
$\boldsymbol{z} = \boldsymbol{x} - \bar{\boldsymbol{x}}$: 中心化サンプル
$\boldsymbol{x}$: 原データのベクトル
$\bar{\boldsymbol{x}}$: 平均ベクトル

直交変換（OT）による合成データ生成法

q 次元空間 $(q<m)$ への射影

$$\boldsymbol{f}^{(q)} = \left(\boldsymbol{u}_1^T\boldsymbol{z}, \boldsymbol{u}_2^T\boldsymbol{z}, \cdots, \boldsymbol{u}_q^T\boldsymbol{z}\right)^T = \boldsymbol{U}_{(q)}^T\boldsymbol{z}, \qquad \boldsymbol{U}_{(q)} = (\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_q)$$

原データの空間での再構成

$$\hat{\boldsymbol{x}}^{(q)} = \bar{\boldsymbol{x}} + \boldsymbol{U}_{(q)}\boldsymbol{f}^{(q)} = \bar{\boldsymbol{x}} + \boldsymbol{U}_{(q)}\boldsymbol{U}_{(q)}^T\boldsymbol{z}$$

提案手法による合成データの情報損失指標

$$ILOT = \frac{\sum_{i=q+1}^{p} \lambda_i^R}{p}$$

λ_i^R:相関係数行列$\boldsymbol{R}$の第i固有値

$$\boldsymbol{R} = \sum_{i=1}^{p} \lambda_i^R \boldsymbol{u}_i^R {\boldsymbol{u}_i^R}^T$$

Cf. 回帰による合成データ生成

i番目の変数を目的変数, 残りの変数を説明変数とした回帰モデル

$$\boldsymbol{x}_i = \boldsymbol{X}^{-i}\boldsymbol{\beta}^{-i} + \boldsymbol{\varepsilon}_i \ , i = 1, \cdots, p, \ \boldsymbol{\varepsilon}_i \sim N\left(0, \sigma_i^2\right)$$

$$\boldsymbol{X}^{-i} = \left(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1}, \cdots, \boldsymbol{x}_p\right)$$

撹乱項無しのデータ生成（DRI）

$$\widehat{\boldsymbol{x}}_i = \boldsymbol{X}^{-i}\widehat{\boldsymbol{\beta}}^{-i} \ , i = 1, \cdots, p$$

撹乱項付きのデータ生成（SRI）

$$\widehat{\boldsymbol{x}}_i = \boldsymbol{X}^{-i}\widehat{\boldsymbol{\beta}}^{-i} + \boldsymbol{e}_i \ , i = 1, \cdots, p \ , \ \boldsymbol{e}_i \sim N\left(0, \hat{\sigma}_i^2\right)$$

レコードリンケージによるリスク評価

原データのi番目のレコード$\boldsymbol{o}_i$から生成した合成データを$\boldsymbol{s}_i$とする

原データ

$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\cdots$	$\boldsymbol{x}_p$
$\boldsymbol{o}_1$			
$\boldsymbol{o}_2$			
$\vdots$			
$\boldsymbol{o}_i$			
$\vdots$			
$\boldsymbol{o}_n$			

k近傍の探索

合成データ

$\hat{\boldsymbol{x}}_1$	$\hat{\boldsymbol{x}}_2$	$\cdots$	$\hat{\boldsymbol{x}}_p$
$\boldsymbol{s}_1$			
$\boldsymbol{s}_2$			
$\vdots$			
$\boldsymbol{s}_i$			
$\vdots$			
$\boldsymbol{s}_n$			

侵入者が$\boldsymbol{s}_i$に対して, 原データの中からk近傍$knn(s_i)$を探した時に, $\boldsymbol{o}_i$が含まれていれば, 特定されたと考え, 特定されたレコード割合をリスク評価値とする.

$$Risk(k) = \frac{\sum_{i=1}^{n} I[o_i \in knn(s_i)]}{n}$$

有用性評価指標

平均絶対誤差により, 有用性（情報損失）を評価する

$$\mathrm{MAE} = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left|o_{ij} - s_{ij}\right|$$

利用データ

- 平成16年全国消費実態調査の匿名データ
- ２人世帯データ
- サンプル数： 43861
- 利用する項目（変数）：現金で購入された食料品82項目.
- ただし, 左の肉類, 生鮮肉, 加工肉の様に, 他の項目を集計した項目は, 冗長なので除く

削除する集計項目の例

符合	項目
1.3	肉類
1.3.1	生鮮肉
220	牛肉
221	豚肉
222	鶏肉
22X	合いびき肉
224	他の生鮮肉
1.3.2	加工肉
225	ハム・ソーセージ
229	他の加工肉

全国消費実態調査の匿名データを用いた有用性とリスクの評価結果

1. 有用性評価

手法	DRI	SRI	OT(0.01)	OT(0.05)	OT(0.10)	OT(0.20)
MAE	585.114	904.207	64.791	182.304	274.295	382.404

OTの括弧内の数値はILOTの値

2. リスク評価

手法	DRI	SRI	OT(0.01)	OT(0.05)	OT(0.10)	OT(0.20)
k=1	0.000	0.000	0.996	0.930	0.694	0.271
k=3	0.001	0.000	0.997	0.949	0.775	0.377
k=5	0.002	0.000	0.997	0.955	0.801	0.425

結論

- 直交変換による合成データ生成法および, 生成データの有用性（情報損失）評価指標を提案.
- 提案手法のリスクおよび有用性を全国消費実態調査の匿名データを利用して評価.
- 提案手法は, 有用性評価指標に対応した主成分数によって, リスクおよび有用性をコントロールすることができ, 回帰による合成データ生成よりも, 高い有用性を持つデータを生成することもできる.

ご静聴ありがとうございました