局所的な分布を用いた個票データのリスク評価

岡山商科大学　　佐井 至道

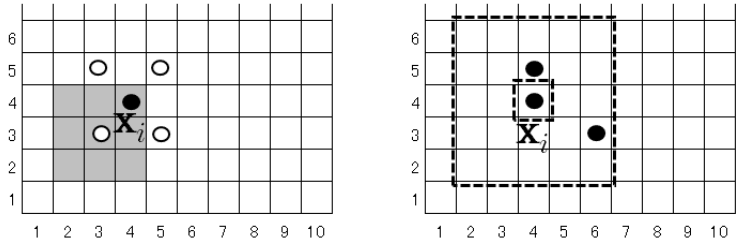
個票データを安全に公開するために様々なタイプの秘匿措置が施されるが，国内においてもノイズの挿入のような攪乱的な方法が用いられることが増えてきた．本報告では，標本調査で得られた個票データにおいて，キー変数（個体を特定するために用いられる変数）にノイズを挿入した場合に，ノイズ挿入後の個体が母集団でも元の個体にリンクされる確率をリスクの指標とするが，その推定のためにキー変数の局所的な分布を用いる方法を提案する．

母集団の大きさを N，標本の大きさを n として，標本から個票データが作成されているとする．キー変数の個数を K として，すべて離散型の量的変数とする．

標本の i 番目の個体のキー変数ベクトルを $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,K})'$，挿入するノイズ変数ベクトルを $\mathbf{e}_i = (e_{i,1}, \ldots, e_{i,K})'$，母集団の i 番目の個体のキー変数ベクトルを $\mathbf{a}_i = (a_{i,1}, \ldots, a_{i,K})'$ として，$\mathbf{x}_i$ に対応する母集団のキー変数ベクトルを $\mathbf{a}_{i'}$ とする．ここで $\mathrm{d}(\mathbf{x}_i + \mathbf{e}_i, \mathbf{a}_{i''}) \leq \mathrm{d}(\mathbf{x}_i + \mathbf{e}_i, \mathbf{x}_i)$ を満たす $\mathbf{a}_{i''}$ $(i'' \neq i')$ が少なくとも 1 つ存在すれば，ノイズを挿入した個体について母集団内で間違ったリンクが発生し，そうでない場合には真のリンクが発生したと考える．ここで $\mathrm{d}(\cdot, \cdot)$ は距離関数である．

上の不等式を満たす $\mathbf{a}_{i''}$ の領域を領域 D と呼び，$\mathbf{a}_{i''}$ が領域 D に入る確率を $p_f(\mathbf{x}_i, \mathbf{a}_{i''})$ と表す．このとき $\mathbf{x}_i$ に対応する母集団の $\mathbf{a}_{i'}$ 以外の $N-1$ 個のキー変数ベクトルが間違ったリンクとならない確率，すなわち真のリンクとなる確率は $P_t(\mathbf{x}_i) = \prod_{i''\,(i'' \neq i')} \{1 - p_f(\mathbf{x}_i, \mathbf{a}_{i''})\}$ と表せ，その期待値 $\mathrm{E}\big[P_t(\mathbf{x}_i)\big] = \frac{1}{n}\sum_{i=1}^{n} P_t(\mathbf{x}_i)$ をリスクの指標とする．これが推定目標である．

下の図に，$K = 2$ で $x_{i,k}$ が値 $1, 2, \ldots, 10$ をとりうる場合のイメージを示す．キー変数ベクトルの 1 つの $\mathbf{x}_i = (4, 4)$ を例として，各ノイズ変数 $e_{i,k}$ は ± 1 の値を確率 $1/2$ で独立にとるものとする．



左の図では $\mathbf{x}_i + \mathbf{e}_i$ として可能性のある値を白抜きの点で示している．$\mathbf{e}_i = (-1, -1)$ の場合，領域 D は図の網掛けされた部分となるが，可能性のある領域 D の和集合を周辺セル H と呼ぶことにする．右の図では，周辺セル H を大きく点線で囲んでいるが，キー変数ベクトルの値が含まれている小さな点線内の中央のセル $(4, 4)$ は周辺セルには含めないことにする．

セルに含まれる個体数はサイズと呼ばれる．$(4, 4)$ の中央セルはサイズ 1，その周辺セルはサイズ 2 で，サイズの組は $(1, 2)$ となる．一般に標本においてサイズの組 (l, h) となる組の数を $s_{(l,h)}$ と表して多重標本寸法指標と呼び，母集団では $S_{(l,h)}$ と表して多重母集団寸法指標と呼ぶ．

非攪乱的な秘匿措置が施された個票データについては，標本寸法指標 s_l を基に母集団寸法指標 S_l を推定するのがリスク評価方法の主流であるが，同様に，多重標本寸法指標を基に多重母集団寸法指標を推定できれば，対象となる個体が入る中央セルの近くにおける母集団での個体の集散状況が推測できるため，真のリンク確率の期待値を推定することが可能となる．

具体的な推定方法と数値例を用いた推定結果については当日報告するが，推定には制約付きノンパラメトリック最尤推定法を基にした探索的な方法を用いる．