ノイズが挿入された個票データの変数の型によるリスクの差について

岡山商科大学　　佐井 至道

個票データの秘匿措置の一つに攪乱的な方法がある．ノイズの挿入やスワッピングが代表的な手法で，国内においても用いられることが増えてきた．攪乱的な秘匿措置が施されると寸法指標を用いたリスク評価ができないため，秘匿後の個体が元の個体にリンクされる確率などをリスクの指標とすることが多い．本報告では，標本調査で得られた個票データにおいて，個体を特定するために用いられるキー変数にノイズを挿入した場合を想定し，母集団との関係も考慮に入れたリスク評価について考えるが，キー変数の型，特に連続型と離散型の量的変数の場合のリスクの差について考察する．

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，ここでは量的変数に絞って議論する．

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 $\mathrm{d}(\mathbf{x}_i + \mathbf{e}_i, \mathbf{a}_j) \leq \mathrm{d}(\mathbf{x}_i + \mathbf{e}_i, \mathbf{x}_i)$ を満たす $\mathbf{a}_j$ $(j \neq i')$ 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 2 つの個体のキー変数ベクトルの距離である．

上の不等式を満たす $\mathbf{a}_j$ の領域を領域 D と呼び，ある $\mathbf{x}_i$ に対して $\mathbf{a}_j$ が領域 D に入る確率 $p_f(\mathbf{x}_i, \mathbf{a}_j)$ を領域 D の確率と呼ぶ．また $\mathbf{x}_i$ が真のリンクとなる確率を $P_t(\mathbf{x}_i)$ と書き，単に真のリンク確率と呼ぶ．それぞれの期待値 $\mathrm{E}(p_f(\mathbf{x}_i, \mathbf{a}_j))$，$\mathrm{E}(P_t(\mathbf{x}_i))$ をリスクの指標として考える．

ここで $K = 1$ の簡単な 3 つの場合を考える．(A) $a_{i,1}$, $x_{i,1}$ が区間 $[0, 1]$，$e_{i,1}$ が区間 $[-\frac{1}{2}c, \frac{1}{2}c]$ の連続型一様分布にそれぞれ従う．(B) $a_{i,1}$, $x_{i,1}$ が区間 $[0, 1]$ の連続型一様分布に従い，$e_{i,1}$ が $\pm\frac{1}{2\sqrt{3}}c$ を確率 $\frac{1}{2}$ ずつでとる二値分布に従う．(C) $a_{i,1}$, $x_{i,1}$ が $\frac{1}{M}, \frac{2}{M}, \ldots, \frac{M}{M}$ を確率 $\frac{1}{M}$ ずつでとる離散型一様分布に従い，$e_{i,1}$ が $\pm\frac{1}{M}, \pm\frac{2}{M}, \ldots, \pm(\frac{1}{2}c - \frac{1}{M})$ を確率 $\frac{1}{Mc}$ ずつでとり，0 を確率 $\frac{2}{Mc}$ でとる離散型分布に従う．なお，いずれの場合もすべての変数は互いに独立で，i についても独立とする．c はノイズの相対的な散布度を表す正の実数で，M は離散型一様分布の取り得る値の数で，自然数である．

このとき，領域 D の確率の期待値は (A) と (C) が $\mathrm{E}(p_f(\mathbf{x}_i, \mathbf{a}_j)) = \frac{1}{2}c$ となり，離散型と連続型の一様分布の値は等しくなる．なお (B) は $\mathrm{E}(p_f(\mathbf{x}_i, \mathbf{a}_j)) = \frac{1}{\sqrt{3}}c$ である．

しかし，真のリンク確率の期待値の傾向は大きく異なる．$c = 10^{-1}$ の場合の，いくつかの N と M の組み合わせに対する真のリンク確率の期待値 $\mathrm{E}(P_t(\mathbf{x}_i))$ を表に示す．他の分布や $K \geq 2$ についての検討の詳細は当日報告する．

表: 真のリンク確率の期待値 $\mathrm{E}(P_t(\mathbf{x}_i))$

	$M \setminus N$	10	10^2	10^3	10^4
	20	$6.3025 \cdot 10^{-1}$	$6.2321 \cdot 10^{-3}$	$5.5703 \cdot 10^{-23}$	$1.8127 \cdot 10^{-223}$
	40	$6.4600 \cdot 10^{-1}$	$4.1000 \cdot 10^{-2}$	$5.1830 \cdot 10^{-12}$	$5.7032 \cdot 10^{-111}$
(C)	60	$6.4895 \cdot 10^{-1}$	$6.5271 \cdot 10^{-2}$	$1.7019 \cdot 10^{-8}$	$3.4498 \cdot 10^{-74}$
	80	$6.4999 \cdot 10^{-1}$	$7.8096 \cdot 10^{-2}$	$8.7198 \cdot 10^{-7}$	$5.9490 \cdot 10^{-56}$
	100	$6.5047 \cdot 10^{-1}$	$8.5166 \cdot 10^{-2}$	$8.7215 \cdot 10^{-6}$	$4.5430 \cdot 10^{-45}$
(A)		$6.5132 \cdot 10^{-1}$	$9.9997 \cdot 10^{-2}$	$1.0000 \cdot 10^{-2}$	$1.0000 \cdot 10^{-3}$
(B)		$5.8554 \cdot 10^{-1}$	$2.7742 \cdot 10^{-3}$	$1.5808 \cdot 10^{-26}$	$5.7056 \cdot 10^{-259}$