リスクの高いデータの秘匿について

岡山商科大学　　佐井 至道

大規模なデータの公開に対する要望が高まっている．大規模データとしては官庁で収集された公的統計を初めとして，自治体，医療機関，一般企業などで収集されたものも増えてきており，その形式も多岐にわたる．特に 1 つのレコードに対する情報が多い，いわば次元の高いデータの場合，識別子と呼ばれる氏名や住所のような項目を削除しても，年齢や性別のような準識別子（キー変数）と呼ばれる項目の値の組み合わせによって個人が特定されたり，特定されないまでもセンシティブな値が特定されたりする．本報告ではこのようなリスクの高いデータについて，どのような秘匿が効果的かを，リスク評価の観点から考える．

最も一般的なデータの形式は，個体（レコード）ごとに変数（項目）の値が並べられたものである．まず，変数の数がすべての個体で等しい場合を考える．データに含まれる個体数を n，キー変数の数を K として，このデータが取られた母集団の大きさを N とする．

秘匿方法としては，カテゴリーの併合に一般化される非攪乱的な方法と，ノイズの挿入のような攪乱的な方法があるが，n，N，K，秘匿の方法と程度により，リスクがどのように変化するか，またどのようなパラメータの組み合わせが同じリスクになるかを検討する．

まず，キー変数が離散型の量的変数であることを想定して，各キー変数が互いに独立に $1, 2, \ldots, M$ を等確率でとる単純な場合を想定する．表 1 に，$M = 10$，$K = 1, 2, \ldots, 10$ の場合，各キー変数の値に ± 1 のノイズを挿入したとき，標本の 1 個の個体が母集団で元の個体にリンクされる確率（真のリンク確率と呼ぶ）の期待値を例示する．他の検討結果については当日報告する．なお “–” は 10^{-300} 未満を意味する．

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

$K \setminus N$	10	10^2	10^3	10^4	10^5	10^6	10^7	10^8
1	$4.04 \cdot 10^{-2}$	$4.62 \cdot 10^{-16}$	$1.79 \cdot 10^{-155}$	–	–	–	–	–
2	$4.28 \cdot 10^{-1}$	$8.81 \cdot 10^{-5}$	$1.21 \cdot 10^{-41}$	–	–	–	–	–
3	$7.82 \cdot 10^{-1}$	$6.66 \cdot 10^{-2}$	$1.33 \cdot 10^{-12}$	$1.38 \cdot 10^{-119}$	–	–	–	–
4	$9.23 \cdot 10^{-1}$	$4.13 \cdot 10^{-1}$	$1.32 \cdot 10^{-4}$	$1.51 \cdot 10^{-39}$	–	–	–	–
5	$9.70 \cdot 10^{-1}$	$7.19 \cdot 10^{-1}$	$3.57 \cdot 10^{-2}$	$3.28 \cdot 10^{-15}$	$1.38 \cdot 10^{-145}$	–	–	–
6	$9.88 \cdot 10^{-1}$	$8.76 \cdot 10^{-1}$	$2.62 \cdot 10^{-1}$	$1.49 \cdot 10^{-6}$	$5.28 \cdot 10^{-59}$	–	–	–
7	$9.95 \cdot 10^{-1}$	$9.47 \cdot 10^{-1}$	$5.80 \cdot 10^{-1}$	$4.30 \cdot 10^{-3}$	$2.13 \cdot 10^{-24}$	$1.94 \cdot 10^{-237}$	–	–
8	$9.98 \cdot 10^{-1}$	$9.79 \cdot 10^{-1}$	$8.05 \cdot 10^{-1}$	$1.14 \cdot 10^{-1}$	$3.77 \cdot 10^{-10}$	$5.77 \cdot 10^{-95}$	–	–
9	$9.99 \cdot 10^{-1}$	$9.92 \cdot 10^{-1}$	$9.19 \cdot 10^{-1}$	$4.29 \cdot 10^{-1}$	$2.10 \cdot 10^{-4}$	$1.70 \cdot 10^{-37}$	–	–
10	1.00	$9.97 \cdot 10^{-1}$	$9.68 \cdot 10^{-1}$	$7.21 \cdot 10^{-1}$	$3.77 \cdot 10^{-2}$	$5.79 \cdot 10^{-15}$	$4.20 \cdot 10^{-143}$	–

レコードごとに変数の数が異なるデータも少なくない．例えば労働力調査では，就業者のレコードには週労働時間などの就業に関する変数が付随して，非就業者とは変数の数が異なる．その場合，非就業者の項目をダミーの値で置き換えることもできるが，就業状態別に秘匿やリスク評価を行った方が効果的かもしれない．移動履歴データでも，長時間の停止が多数存在する場合になどには，工夫が必要になろう．

官庁統計でも，個体が世帯で，そのレコードに全世帯員の情報が付随しているものがあり，世帯員数によって変数の数は異なる．医療関係でも，レセプトデータでは期間が個体ごとに大きく異なることがある．特にその中にキー変数が多数存在する場合には，個体の特定を防ぐことを断念し，センシティブな値の漏洩を防ぐように方針を変更したり，特に後者の場合には，診療期間の表示を 1 週間ずらすなどの別のタイプの秘匿措置も候補に入れるできであろう．