回答拒否を許す統計調査におけるLDP推定問題の精度解析

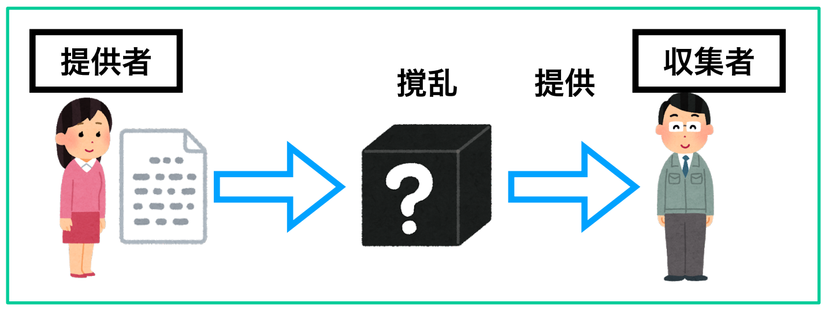
小野　元　　総合研究大学院大学　統計科学専攻　博士課程(5年一貫制)4年

【概要】
本研究では回答拒否を許す自由度と形式的なプライバシー保護を両立する統計調査フレームワークにおいて，パラメータ推定問題を解いた場合の推定誤差のminimax下界の解析を行う．
さらに，具体例としてlogistic回帰の係数推定を考え，その下界を高々定数倍で達成可能なアルゴリズムが存在することを示した．

【動機】
近年国勢調査等の調査票回収率の低下が顕著になっている
国民のプライバシー意識の高まりが原因の一つ見られる．
これに対して，次の2つの疑問が生じた：
- 疑問1：回答拒否はプライバシ保護になっているのか？
- 疑問2：回答拒否は統計の精度にどう関係するか？

本研究では，パラメータ推定，特にロジスティック回帰の係数推定，を題材にこれらの疑問を解消することで，実務者が回答拒否のプライバシ保護・精度にどの程度悪影響を与えるのかの定量的な理解の助けになることを意図する．

【プライバシ定義：local differential privacy】
本研究では，プライバシ定義としてlocal differential privacy (LDP) [1,2]を採用し，それに基づいてプライバシの議論を行う．
LDPは「データを収集する実体 (収集者)は全能でもないし，悪意を持ってデータを取り扱うかもしれない」という考えのもと，
レコードを提供する実体 (提供者) が提供前にレコードをランダムに撹乱することを要請する．
その撹乱が撹乱前のレコードの特定をどの程度難しくするかを定量的に定義するのがLDPである．



ε -LDP
ある撹乱メカニズムQが，任意の入力対r, r'と任意の出力集合Sに対して，以下の不等式を満たすとき，Qはε -LDPであるという:

どんな入力対を考えても，
ある値を出力する確率の比は常に
一定以下に抑えられている
→出力を見ても，入力がわからない

$$\frac{\Pr(Q(r) \in S)}{\Pr(Q(r') \in S)} \leq e^{\epsilon}$$

ε が小さいほど「似てる」
→より安全

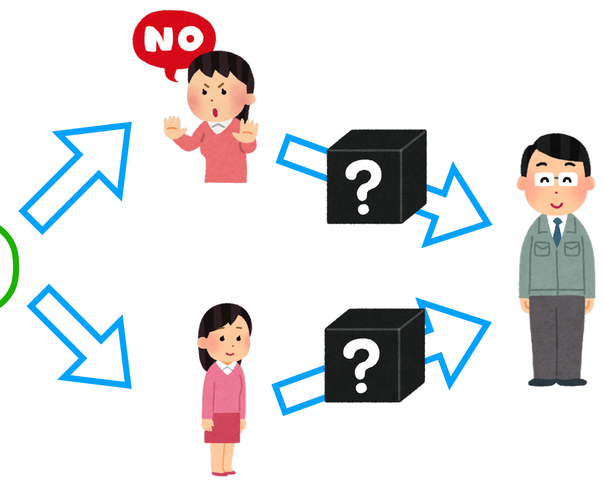
【関連研究】
Differential privacyは統計の公開に際して，統計にランダムな摂動を加えることで，低い確率でしかレコードを特定できないことを保証するプライバシ定義である．
このプライバシ定義ではデータを収集し統計を算出する実体(収集者)が善良であることを仮定しており，収集者が悪意を持ってプライバシ侵害を行なった場合は防ぐ手立てはない．
悪意あるデータ収集者からデータ提供者を守るための定義の1つとして，local differential privacy(LDP)がある．LDPはデータ提供者自身がレコードを提供する前にレコードを撹乱することを前提としており，仮に収集者がレコードの悪用を試みたとしても，生のレコードがわからないため，レコードは保護されている．
DuchiらがLocal differential privacy制約下でのパラメータ推定でのminimax下界を導出した[2]．彼らは情報理論的なアプローチでプライバシ保護による情報量の損失を定量的に評価した．

回答拒否は統計学の文脈において，欠測の一種と分類される．
欠測を含むデータを用いたパラメータ推定のためのアルゴリズムはここ数十年統計学者によって盛んに研究されているが，そのminimax下界の研究は限られている[3]．
LDPに基づいたプライバシ保護をしながらの欠測データ解析のminimax下界は非自明である．

【疑問1への回答：回答拒否はプライバシ保護になっていない】
回答拒否するかどうかは2値の確率変数であると見なすことができる．
回答拒否の有無と特定の機微な属性の間に強い相関や決定的な関係があることがわかっている場合には回答拒否したという事実が，その機微な属性を暗示してしまう．
したがって，一般には回答拒否はプライバシ保護の手段として不適切である．LDPの意味でプライバシ保護をするには，回答拒否したかどうかも隠蔽する必要がある．

【疑問2への回答：ロジスティック回帰の係数推定のMinimax Excess Risk解析】
撹乱される前のデータをD，撹乱されたデータをZとする．
Dはn人の提供者のレコードの集合で，レコードはそれぞれd個の説明変数と1つの目的変数からなる．
Zはε-LDPである撹乱ルールQによって撹乱されたものとする．

x_iを生成 → β^*でパラメトライズされたロジスティックモデルからY_iを生成 → 確率的に非協力的に



$$\Pr(Y_i = y_i \mid X_i = x_i) = \frac{1}{\exp(-y_i \beta^{*\top} x_i) + 1}$$

β^*でパラメトライズされたロジスティックモデルからデータが生成されたとして，
ある推定アルゴリズムAのexcess riskを次のように定義する：

$$ER(A, \beta^*) \equiv \mathbb{E}_D \left[\mathbb{E}_A \left[\mathbb{E}_{(X,Y)} \left[\log \frac{p(Y \mid X; \beta^*)}{p(Y \mid X; A(D))} \right] \right] \right]$$

これを用いて，minimax excess riskを次のように定義する:

$$\mathscr{M} \equiv \inf_A \sup_{\beta^*} ER(A, \beta^*)$$

Table 1. Summary of bounds for logistic regression. $\bar{\epsilon} = \min\{\epsilon^2, \epsilon, d\}$. These upper bounds hold with $0 < \epsilon \leq d$. $\underline{\alpha}^{(\mathrm{U})}$ and $\underline{\alpha}^{(\mathrm{Y})}$ are lower bounds of non-response rates in each case. $\overline{\alpha}^{(\mathrm{U})}$ and $\overline{\alpha}^{(\mathrm{Y})}$ are upper bounds. $\check{\alpha}^{(\mathrm{Y})}$ is the average non-response rate.

Non-response	Lower bound	Upper bound
$\mathbf{X}$ Unit	$\frac{d^2}{\bar{\epsilon} n (1 - \underline{\alpha}^{(\mathrm{U})})}$	$\frac{d^2}{(1 - \overline{\alpha}^{(\mathrm{U})})(\epsilon^2 \wedge \epsilon) n} \vee \frac{1}{(1 - \overline{\alpha}^{(\mathrm{U})})(\epsilon \wedge 1)} \sqrt{\frac{\log(1/\delta)}{n}}$
Y	$\frac{d^2}{\bar{\epsilon} n (1 - \underline{\alpha}^{(\mathrm{Y})})}$	$\frac{(1 - \check{\alpha}^{(\mathrm{Y})}) d G^2}{\epsilon (1 - \overline{\alpha}^{(\mathrm{Y})})^2 n} \vee \frac{d G^2}{\epsilon^2 n (1 - \overline{\alpha}^{(\mathrm{Y})})}$

【参考文献】
[1] Kasiviswanathan, S.P., Lee, H.K., Nissim, K., Raskhodnikova, S., Smith, A.: What can we learn privately? SIAM Journal on Computing 40(3), 793–826 (2011). https://doi.org/10.1137/090756090
[2] Duchi, J.C., Jordan, M.I., Wainwright, M.J.: Local privacy and statistical minimax rates. In: 2013 IEEE 54th Annual Symposium on Foundations of Computer Science. pp. 429–438 (2013)
[3] Loh, P., Wainwright, M.J.: Corrupted and missing predictors: Minimax bounds for high-dimensional linear regression. In: 2012 IEEE International Symposium on Information Theory Proceedings. pp. 2601–2605 (2012)