資料３－１

匿名データの有用性に関する指標について

１．考え方・方向性

・有用性の指標とその目的：

①匿名データ**作成時**：複数の匿名化手法の評価のための指標

②匿名データ**利用時**：元のデータとの違いの程度を示すための指標

（匿名データの品質表示としての指標）

・（作成時）当面、定められた匿名化基準に基づき匿名データを作成

・（利用時）利用の際に、個々の分析者が要約統計量等を作成

⇒2019年度は、上記②の利用時の指標について検討・試算を予定

⇒併せて諸外国の事例についても調査を予定

２．想定される指標

(1) 元データから匿名データへの変化量を見るための指標

①元データ及び匿名データに関する要約統計量

度数分布、平均、分散、四分位範囲、レンジ（最大値－最小値）など

（※トップ（ボトム）コーディングを行った階級の基本統計量）

②元のデータ及び匿名データに基づく相関係数（順位相関係数）

(2) 解析結果の変化量を見るための指標

①回帰分析における回帰係数の符号、絶対値、決定係数など

②多変量解析（主成分分析、因子分析など）の結果、統計量の比較

３．今後の予定

2019年６月～：匿名データ有識者会議における議論

共通課題検討ワーキンググループにおける議論

海外の事例に関する調査

調査・議論の結果を踏まえた指標の検討・試算

2020年　　：提供する指標の検討・試算

（別紙）

有用性の指標の例

１．データ自体の変化量による有用性の指標

(1) 連続変数（ユークリッド距離）

$U_{Euclid}(D, D') = \sum_{i=1}^{n} \sum_{j=1}^{d} \left(x_{ij} - x'_{ij}\right)^2$.

$D = (\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n})$, $D' = (\boldsymbol{x'_1}, \boldsymbol{x'_2}, \cdots, \boldsymbol{x'_n})$

$\boldsymbol{x_i} = (x_{i1}, x_{i2}, \cdots, x_{id})$, $\boldsymbol{x'_i} = (x'_{i1}, x'_{i2}, \cdots, x'_{id})$

(2) 離散変数（カルバック・ライブラー・ダイバージェンス）

$U_{KL}(D, D') = \sum_{j=1}^{d} KL\left(p_j, p'_j\right) = \sum_{j=1}^{d} \sum_{x_j \in X_j} p_j(x) \log\left(p_j(x)/p'_j(x)\right)$.

２．解析結果の変化量に関する有用性の指標

$U(f, D, D') = f(D) - f(D')$

$f(D)$： ①回帰分析における回帰係数（絶対値・符号）、p値・t値、決定係数など

②多変量解析（主成分分析、因子分析など）の結果、統計量の比較

※平均値などの要約統計量も、「２.」に含まれる

参考文献：

[1] 伊藤伸介，村田磨理子，高野正博（2014）ミクロデータにおける匿名化技法の適用可能性：全国消費実態調査と家計調査を用いて，統計研究彙報，第71号，pp83-124.

[2] 伊藤伸介（2017）国勢調査ミクロデータにおける匿名化の誤差の評価方法に関する一考察，中央大学経済学論纂，第57巻，第３・４号併号，pp189-209.