差分プライベートなサンプリング

金沢大・経　星野伸明

「差分プライバシー」はランダムな出力（匿名データなど）から秘密が漏洩する危険性を評価するための概念である。情報学の文脈で提案された概念なので、統計家にとっては分かりづらい説明がなされることが多いが、その根本的なアイデアは統計家にとって既知である。すなわち秘密が未知母数であり、差分プライバシーは未知母数の微小変化（差分）の有無に対応する 2 つの仮説の尤度比が、母数空間と標本空間の全域で 1 に近いことを要求する。これは対数尤度関数の傾きがほとんど全ての観測値について押さえられていると解釈できるので、秘密の推定量の分散は情報量不等式から下限が定まる。従って差分プライバシーは、秘密の推測精度の上限を管理する基準である。

このように差分プライバシーは理論的に自然な概念なので、プライバシー保護の研究では便利である。しかし推測による秘密漏洩と個体識別による秘密漏洩は別概念で、差分プライバシーと個体識別の関係は明白ではない。実務的には個体識別が可能か否かの判定が必要なので、差分プライバシーの実務上の意義も明らかでない。

本報告では、差分プライバシーと個体識別の関係性を、母集団一意の推定精度の管理に求める。Marsh et al. (1991) の議論に従えば、あるレコードが母集団一意であることを確証できることは、個体識別の要件である。そして通常は、母集団一意であることは確率的な標本から推定されることである。従ってこの場合、母集団一意の推定精度を管理することは、母集団一意の確証可能性を管理することと考えられる。情報学的な発想では特定個体と他の個体が区別できないことを（文脈に応じた）尤度比が 1 に近いことで定義し、そのような個体が存在すれば識別が “deniable” と考える。母集団一意の確証を差分プライバシーで管理すれば、区別できない個体が存在しないことの確度が高くないという意味の “deniability” を達成できる。

しかしながら差分プライバシーはプロトコルの安全性を確保するための基準で、特定データの安全性を確保するための基準ではない。言い換えれば、特定の母数（現実のデータの一部分）さえ保護できればよいはずが、母数空間の全域に渡って保護することが要求される。統計的開示制限の立場からは明らかに過保護で、刊行するデータの有用性に重篤な悪影響を及ぼす。特に高次元の有用なデータ刊行は難しく、Machanavajjhala et al. (2008) は負の超幾何分布によるサンプリングを提案しているが、最終的に素の差分プライバシー概念を放棄した。

本報告では Machanavajjhala らによる素の差分プライバシーの定式化を採用した上で、負の超幾何分布より有望なサンプリング（疑似多項分布）を提案する。また公的統計の匿名化手法としてよく使われるリサンプリングないしサブサンプリングの安全性について、差分プライバシーの観点から考察する。

単純無作為抽出は非復元なら超幾何分布、復元なら多項分布が出力のランダムメカニズムである。これらについて Machanavajjhala らの差分プライバシーを考察すると、標本空間が母集団度数に依存すれば差分プライバシーは成立しないことが分かる。結局、Machanavajjhala らの意味での差分プライバシーを一般に達成するには母集団にダミーを加える必要がある。これはつまり、差分プライベートな刊行データは、模造でしかありえないということだ。実際、Machanavajjhala らも模造データの刊行を目的としている。問題はダミーの量で、多ければ有用性が悪い。そして数値的に評価すると、差分プライベートな単純無作為抽出が生成するデータは使い物にならない。負の超幾何分布も少しはましだが、やはり使い物にならない。

第二種の疑似多項分布 (Consul and Mittal, 1977) において必要なダミーの量は桁違いに少ない。刊行サンプルのサイズを m と書くとき、ダミーの量は負の超幾何分布でも $O(m)$ だが、疑似多項分布だと $O(1)$ で m に依存しない。従って大規模データの保護において、疑似多項分布は極めて有利である。