統計研究彙報　第76号　2019年３月　(81-100)

オンサイト利用における持ち出し安全性基準及び審査方法

南　和宏 †

菊池　亮 ‡

Safety Rules for Output Checking for On-site Use in Japan

MINAMI Kazuhiro
KIKUCHI Ryo

オンサイト利用は公的調査票情報の２次的利用を推進する有力な手段である。ただし利用者が持ち出す分析結果に対して安全性審査を実施することで、機密情報の漏洩を防止する必要がある。本論文では、データ持ち出しの安全性基準を整備する際の基本的な指針を紹介し、その際参照した欧州統計連合の基準との差異、拡張部分を概説する。また審査作業を効率化するために *R* 言語で表データのセル秘匿処理ツールを開発したのでその概要を紹介する。最後に今後安全かつ柔軟な持ち出し審査を実施するための検討課題を議論する。

キーワード：統計開示抑制、オンサイト利用安全性審査、セル秘匿問題、機密性ルール

In Japan, we have been establishing the on-site service that allows researchers to access microdata of various public surveys at secure on-site facilities. However, we must prevent any unintended disclosure of sensitive information on survey participants by verifying the safety of researchers' output data. In this paper, we introduce basic principles for establishing output checking rules while clarifying the difference from those in the Eurostat SDC handbook. We also describe our R-based tool for performing statistical disclosure control on tabular data, which bridges the gap between researchers and output checkers. We finally discuss possible modifications of current conservative safety rules to meet the needs of researchers in a more flexible way.

Keywords: Statistical disclosure control, Output checking, Cell suppression problems, Sensitivity rules

† 統計数理研究所
‡ NTT セキュアプラットフォーム研究所

1　はじめに

近年、我が国は公的調査票情報の二次的利用を推進し、平成 30 年 1 月よりオンサイト利用制度 [（独）統計センター, 2018]の本格運用が開始している。この制度により、学術研究を目的とする研究者はオンサイト施設の端末から調査票情報に対する探索的分析を行なうことが可能となる。しかし調査票情報には調査客体の機密情報を含まれるため、研究者が分析結果を学術論文として公表する際、個人のプライバシーに関する情報が漏洩するのを防止する必要がある。そのため、 同様のオンサイト利用制度を既に運用している欧米諸国では、分析結果が安全に公開できるかを検証する安全性審査を実施している。したがって、日本でもオンサイト利用の運用開始にあたり、安全性審査に用いる安全性基準（ルール）を事前に整備することが不可欠であった。

本論文は、著者らが 2016 年 5 月～2017 年 1 月にかけての準備過程で作成したオンサイト利用のデータ持ち出しに関する安全性基準の提言をまとめたものであり、その基本的な指針の多くは本格運用における安全性基準に反映されている。ただし、実務レベル、特に審査作業の効率化の点からの検討を経た結果、個々の安全性ルールの詳細において、 制度で運用される公式の安全性基準とは異なる部分も生じている。その結果、統計開示抑制技術 (Hundepool, et al., 2012) の専門性を必ずしも有しないオンサイト利用制度の利用者、審査担当者の双方にとって、個々の安全性基準のルールが意図する目的を理解するのは容易ではない。そこで、本論文は安全性基準を策定した際の基本理念、検討課題を解説し、制度に関係する各ステークホルダの機密情報保護に対する理解を深め、円滑なオンサイト利用の運営の実現に資することを目的とする。

我々は安全性基準の策定に際し、欧州連合統計局 (以下、Eurostat) が編纂した統計開示抑制技術のハンドブック (Hundepool, et al., 2010)を精査した。基本的には、記載されている安全性ルールには長年の実績があり、安全性の観点からは妥当と判断した。しかし、一部の安全性ルールの記述に具体性が不足している点があり、安全性審査の実施に必要となる説明資料の要件に関する記述が不足していることも判明した。また、表データに関しては、オランダ統計局が開発したｩ-ARGUS と呼ばれるセル秘匿処理ツール (Statistics Netherlands, 2018) はオンサイト利用における安全性審査には不向きであることが判明した。

そこで我々は本論文で紹介する下記の貢献をした。

i. Eurostat 基準における安全性ルールの詳細を明示的に補完し、安全性審査に必要な説明資料の要件を提示した。
ii. 表データの秘匿処理に関しては、*R* 言語によるセル秘匿処理ツールを開発し、データの持ち出す研究者と安全性を検証する審査者の円滑な連携による審査プロセスを実現した。
iii. 柔軟な持ち出し審査を実施する際の検討課題に関する考察を行った。

本論文の構成は以下の通りである。第 2 章はオンサイト利用制度の概要を紹介し、オンサイトのデータ持ち出しに関する安全性基準の策定に際しての基本的な方針を述べる。第 3 章は安全性基準の基本となる 5 つの原則を紹介し、想定する情報漏えいの防護策との関係を説明する。第 4 章は基本的なデータ形式に関する安全性ルールを紹介し、特に Eurostat 基準との差異に言及する。第 5 章は審査の効率化とのために開発した R 言語による秘匿処理ツールを紹介する。第 6 章は研究者のニーズに答える柔軟な持ち出し審査を実施する際の今後の検討事項をまとめ、第 7 章で全体のまとめとする。

2　安全性基準決定の方針

本章では、日本のオンサイト利用における持ち出しデータの安全性基準を決める前提となる重要な方針を述べる。最初に、オンサイト利用制度の利用形態と安全性審査の構成について紹介す

る。次に今回のオンサイト利用で想定する攻撃者モデルを定義し、複数回のデータ持ち出しに対する統一的な基準適用の原則を紹介する。最後に、今回の安全性基準の策定にあたり参照した Eurostat ハンドブックの構成を説明する。

2.1　オンサイト利用と安全性審査

図 1 にオンサイト利用制度の利用形態を示す。利用者は公益性の高い学術研究を行なう研究者であり、物理的な安全性が確保されたオンサイト施設を訪問する。研究者は施設に設置されたシンクライアント端末を介して、中央データ管理施設のサーバーに格納された調査票情報に対して探索型のデータ分析を実施する。ただし、機密情報の漏洩を防ぐため、端末にはデータを保管することは許されず、分析結果はサーバー内に格納される。研究者が分析結果を持ち出す場合、審査者による安全性審査を受け、審査要件を満足するデータのみが研究者に提供される。

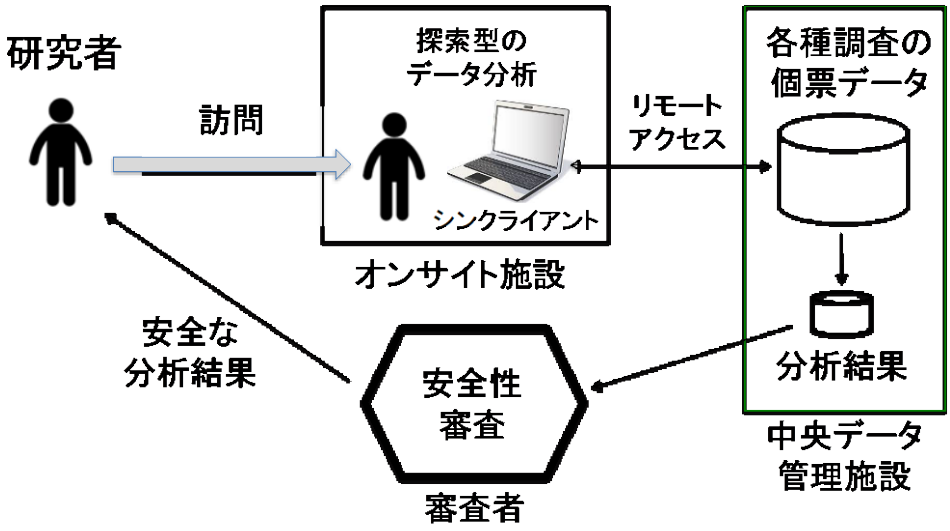


図 1. オンサイト利用制度の利用形態

2.2　攻撃者モデル

安全性基準の内容は想定する情報漏洩シナリオに大きく依存する。したがってオンサイト施設を利用する研究者がどのように情報漏えいに関わるか、その行動モデルを決める必要がある。攻撃者モデルは 2 つに分類でき、1 つは意図的に機密情報を持ち出そうとする**悪意**の研究者であり、もう 1 つはそのような悪意はないが秘匿処理に習熟していない等の理由からのミスによる情報漏えいを行う**善意**の研究者である。

本制度においては、善意の研究者のミスによる情報漏えいを検知し、防止する安全性基準を策定する。なぜなら悪意の研究者の機密情報持ち出しに対する審査基準を策定することは技術的に困難であり、審査作業の負荷を考えると現実的ではないからである。一例を挙げると、悪意をもつ研究者が暗号化した機密情報を持ち出すことの防止するのはほぼ不可能である。したがって、悪意の研究者に対しては、審査とは別途、本制度利用の契約時の機密情報保護遵守への同意、事前教育、罰則等で対策を講じる。

我々が想定する攻撃者は、分析結果を持ち出す研究者でなく、個人情報の特定を目的とする**悪意をもつ学術論文の読者**である。学術論文の読者の多くは善意の研究者であるが、学術論文が広く一般に公開される状況を鑑みると、悪意の読者が論文に示される分析結果と別途入手した

外部知識を組み合わせて、個人情報を識別する可能性は無視できない。攻撃者が利用する外部知識には、一般に公開される公的統計データ及び匿名データ、さらに我々が容易に入手できる第三者に関する外観識別性の高い属性情報（例えば、多人数の子供をもつ世帯等）が考えられる。また、標本調査の場合、標的とする客体が標本に含まれるかどうかの情報 (Response Knowledge) を得ることが重要であるが、クラスター抽出法における調査回答者は同じ調査地区の住人も同様に参加していることが推察できる。

2.3 審査基準の統一性

探索型のデータ分析を行なう研究者は、分析プロセスの途中で中間成果物を何度か生成し、最終的な分析結果に到達する。したがって、オンサイト施設の利用者は中間成果物に対する分析作業を進めるために、中間成果物の持出しを申請することは十分想定される。したがってオンサイト施設の利用者は、中間/最終成果物を問わず、施設からの持ち出し申請を行うことができる。中間成果物には、論文に掲載する最終成果物よりも詳細な情報が含まれるが、そのままの形での論文掲載は通常予定していない。したがって、中間成果物に対して安全性基準を緩和する考え方もあるが、現行のオンサイト利用制度では中間及び最終成果物に対して、統一的な共通の基準を適用する。なぜなら、最終成果物の安全性基準に満たない中間成果物に対して、研究者が常に適切な秘匿処理を行なうことは期待できず、いったん持ち出しを承諾した中間成果物から生成した最終成果物の審査を確実に実施することは制度の運用上困難であり、また煩雑な手続きを伴うからである。

今回の安全性基準は、持ち出したデータにどのような分析、加工を行って論文に公表しても、個人情報が漏えいしないことを目的とする。しかしながら、全てのデータ加工を想定した安全性基準を策定することは技術的に困難であり、また提供までの時間短縮のため、審査を簡便化することも制度運用上考慮する必要がある。したがって、最終的に論文に掲載する情報からの個人情報の漏えいを防止する最終的な責任は研究者に課すものとする。

2.4 欧州連合統計局（Eurostat）の安全性基準の踏襲

Eurostat ハンドブック (Hundepool, et al., 2010)は欧州諸国の公的統計機関の統計開示抑制技術に関する知見を編纂したものであり、このガイドブックに基づき、統計データに関する具体的な安全性基準が「Guidelines for the checking of output based on microdata research」 (Maurice, 2009) に策定されている。この安全性基準を見ると、全ての審査ルールが科学的な根拠から論理的に導き出されるのではなく、過去の審査事例から帰納的に抽出された**経験則**が重要な役割を担う。したがって、我が国の基準策定にあたり、基本的には欧州諸国で既に実績のある (Maurice, 2009)の経験則を踏襲することとした。さらに (Maurice, 2009)が提唱する２種類の審査モデルを使う手法 (a two-model approach) を採用し、

- 明示的でかつ厳格な安全性を要求する**経験則**
- 申請の状況を考慮し柔軟な審査を行うための**原則ルール**

の 2 つからの安全性基準を構成する。

経験則は、審査の開始地点であり、通常の審査は明示的な経験則を適用するところから始める。この経験則は明示的なルールで構成され、統計的開示抑制技術に精通していない審査者も機械的に審査に利用できる簡便性を備える。その一方、審査基準は安全性を重視した厳格なものであり、精査すれば持ち出し可能と判断されるべきデータであっても持ち出し不可になる場合もある。

原則ルールは、持ち出しの判断をするための原則を抽象的に述べたものであり、研究者の目的とする分析結果を可能な限り安全に持ち出すことを目的とする。そのためには、経験則で考慮で

きない様々な要因を考慮する必要があり、データを持ち出す研究者と審査者の密なコミュニケーションが必要になる。原則ルールの主な役割は、厳格な経験則で持ち出し不可となった審査案件のうち、妥当な理由が認められる場合に持ち出しを認めることにある。したがって、原則ルールの審査者は、統計的開示抑制技術 (Hundepool, et al., 2012)に精通している必要がある。

日本のオンサイト利用では、2 種類の審査モデルに対応する形で、図 2 に示すような 2 段階の安全性審査を実施する。1 次審査は 経験則に基づく明示的なルールによる審査である。この 1 次審査を経た申請データはその可否にかかわらず、原則ルールに基づく 2 次審査にかけられる。2 次審査は、原則ルールに基づくデータのセマンティクス（意味）を考慮した審査であり、調査票情報の内容に精通したデータ管理者が担当する。

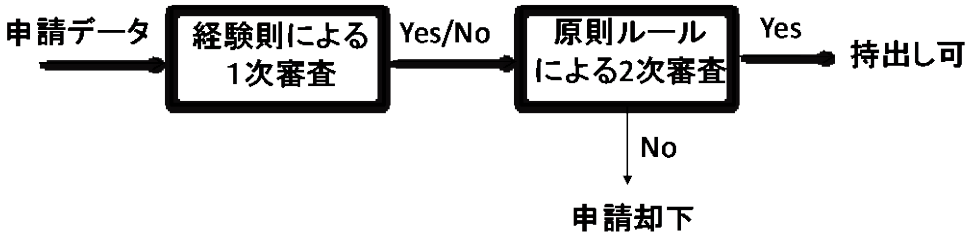


図 2. 安全性審査の２段階構成

3　安全性基準の原則

調査表情報は、各行が調査客体、各列が調査項目に対応する表データである。本論文では、調査票情報の各行を**個票レコード**、また個票レコードに含まれる各項目の値を**個票値**と呼ぶ。なぜ識別子情報を削除した複数の個票レコードの個票値を集計した統計情報から、元の個票値が識別子情報と紐付いて漏えいするのであろうか？代表的な情報漏えいのシナリオを紹介し、その対策として導かれ 5 つの安全性基準の原則を紹介する。これらの原則は Eurostat ハンドブックに記載されるものを採用しているが、本論文では情報漏えいシナリオとの関連性を明らかにすし、その存在理由を説明する。次章以降で説明する安全性基準ルールは、本章で紹介する５つの原則を個々の形式のデータについて具体的に適用したものである。

原則1　個票値は機密情報として扱うこと

例えば、ある世帯の個票レコードから名前、住所等の識別子情報を削除し、家計収入のみを公開するとする。その場合、家計収入の情報から対応する世帯を識別する直接的な方法はない。しかし、もしその収入がある狭い地域で最も高収入な世帯のデータであった場合はどうであろうか？その地域の住人は誰が一番の高額所得者であるか、住居、所有する車、生活スタイル等の情報から高い確率で推測できる可能性がある。その場合、その高額所得者の正確な世帯収入が漏洩することになる。

最大値、最小値のような極端な識別性の高い情報でなくとも年齢、性別、職業等、1 つの個票レコードに含まれる複数の異なる個票値が公開された場合、特定の個人に絞り込まれて識別されることがある。実際、そのような情報漏えい事件は近年米国では何度か起きており (Daniel, 2012)、個々の情報は個人を識別しない安全な情報に見えても、複数の情報の組み合わせで個人の識別が生じる危険性に留意する必要がある。個票値に関して、どこまで情報を開示して安全か明確な線

引きは困難であり、本審査基準では安全性を第一に考え、個票レコードに含まれる全ての個票値は非公開とすべき機密情報と判断する。

原則2　統計値は 10 個以上の客体の個票値から算出されていること。（客体 10 の原則）

議論を簡略化するために、n個の個票レコードに含まれる個票値 $x_1, x_2, \ldots, x_n$ を集計する関数 $f(x_1, x_2, \ldots, x_n)$ を考える。$y = f(x_1, x_2, \ldots, x_n)$ とすると、審査の対象は出力値 y である。ここで考慮すべきは、その出力 y から入力データである個票値$x_1, x_2, \ldots, x_n$ の値が推定できるかという問題であり、関数 f の出力から入力値を求める逆問題に相当する。例えば、関数 f が合計を求める場合、関数 f は、

$$f(x_1, x_2, \ldots, x_n) = x_1 + x_2 + \cdots + x_n \tag{1}$$

となり、出力値 y から元の個票値 $x_1, x_2, \ldots, x_n$ を求めることは不可能に見える。ただし、入力の数が少ない場合、例えば入力が x_1 と x_2 の２つしかない場合はどうであろうか？この場合でも、合計値 y から x_1 または x_2 の正確な値を推論すること、つまりどのように y の値を両者に配分するかを知ることは不可能に見える。

しかし、個票値x_2を提供した調査客体が合計値 y を知った場合には状況が変わる。この場合、x_2 を知る客体は合計値 y から自身の値 x_2 を引くことで、もう一つの客体の値 $x_1 = y - x_2$ と求めることができる。つまり、一見安全と思える統計データであってもその計算の入力データを提供した**内部者**を攻撃者として想定すると不可能な逆問題が簡単に解けてしまう。よって、複数の調査客体が**共謀**して、残りの客体の個票値を算出するリスクを考慮し、統計情報には 10 個以上の客体の個票値が貢献していることを原則とする。この原則では個票値を提供する最大 8 個の客体が共謀しても残りの 2 つの客体の個票値の機密性を守ることができる。

原則3　回帰係数等の数理モデルの情報を取得する場合、自由度 10 以上を残すこと。（自由度 10 の原則）

数理モデルは、複数の変数とその関係を規定する関係式で構成される。個票値の変数 $x_1, x_2, \ldots, x_n$を入力とする関数の出力を持ち出す場合、その関数と入出力の値の組は変数間の関係式とみなすことができ、図３に示すような個票値変数に関する数理モデルが定式化できる。

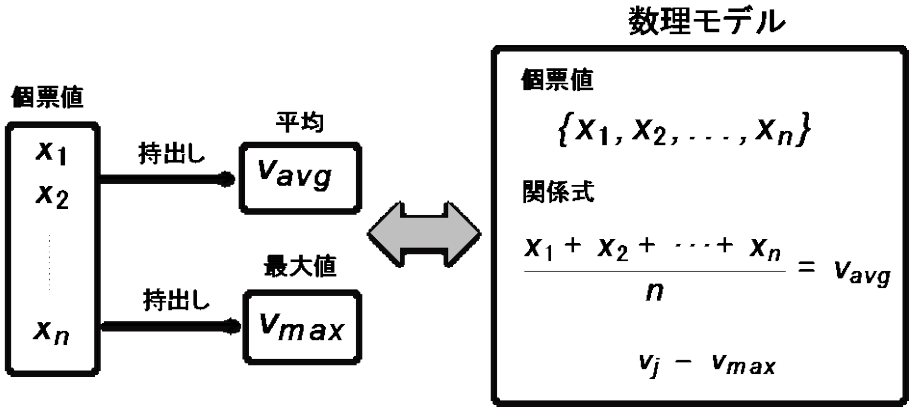


図 3. 個票値の組に関する数理モデル

数理モデルは複数の関係式の存在を想定しており、**自由度**の概念を用いることで同じ個票値の組から計算した複数の異なる統計値を持ち出す場合の安全性を定義することができる。それら関係式を拘束条件として自由に値を決められるモデル変数の数をそのモデルの自由度と呼ぶ。

例えば、３つ個票値の変数x_1,x_2,x_3を考える。この個票値の組に関する統計値が持ち出されていない段階では、各変数の値は不確定であり、数理モデルの自由度は 3 である。次に３つの個票値の合計を算出したとする。もし合計が 6 なら、３つの変数間に$x_1+x_2+x_3=6$が成り立つ。ここでx_1とx_2の値を決めると$x_3=6-x_1-x_2$とx_3の値が一意に決まるので、合計値が与えられた場合の自由度は 2 となる。さらに３つの変数の最大値 $\max(x_1,x_2,x_3)$ =4 が与えられると自由度は 1 になる。

原則 3 は数理モデルがいかなる関係式のセットを含んでいても、自由度が 10 以上を残すことを要件とする。これは、原則 2 で説明した内部者攻撃を数理モデルに適用したものであり、内部者による共謀を防ぐという目的は同じである。

原則4　特定の客体の個票値が合計の 50%以上を占有しないこと（占有性の原則）

原則 2 は、計算に必要な個票値を提供する内部者である客体が統計関数 f の逆問題を解くリスクを防ぐことが目的であった。しかし、原則 2 には重大な抜け道があり、特定の入力 x_iが出力値の大部分を占める場合、入力 x_iのおおよその値が出力 y で近似できてしまう問題が存在する。例えば、入力 $x_1,x_2,\dots,x_n$ の合計 y を求める場合、もし x_i が y の値の大半を占めるとすると、それはその個票値そのものを公開したのとほぼ同義となり、原則 1 に反する。したがって、個々の個票値の占有度に 50%という上限を設けることでこの占有性の問題を防止する。

原則5　ある属性でグループ化された個票の 90%以上が、別の属性に関して同一の区分（または値）に属することを禁ずる。（グループ開示の原則）

この原則は、特定の個票値が識別されなくても個人の機密情報が漏えいする可能性を指摘している。つまり、ある個人があるグループに属することが分かった（グループ開示）だけで、「がん患者である」等、そのグループに属する人に共通の機密属性が漏えいする問題である。

4　安全性基準

本章では、図 2 に示す経験則による 1 次審査のための安全性基準を紹介する。研究者は、調査票情報に対して様々な分析を行うため、どのようなデータの持ち出しを希望するか、事前に予測することは困難である。したがって、日本のオンサイト利用制度の安全性基準は 、多くの研究者のニーズが高いと思われる、度数表、数量表、回帰分析、要約統計量に対象を絞っている。我々の安全性基準は基本的に Eurostat ガイドライン (Hundepool, et al., 2010) の経験則を踏襲し、具体性が足りない部分のみ追加の要件を明示している。

ただし、Eurostat ガイドラインでは、安全性基準の各ルールを審査者が検証するために必要な説明資料に対する要件がほとんど示されていない。これはハンドブックの作成に携わった欧州諸国の公的統計機関では、職員である内部の研究者が自身で分析し、秘匿処理をするという自己完結したプロセスが主だったことが理由と考えられる。しかし、我が国のオンサイト利用において分析者と審査者の役割は完全に分かれており、説明資料の要件の明確化が制度の円滑な運用に不可欠である。よって、我々の安全性基準は審査における具体的なプロセスを考察し、安全性の検証に必要な説明資料の要件を明らかにした。その点が Eurostat ガイドラインと我々の安全性基準の大きな差分と言える。

4.1　説明資料の要件

図 4 はオンサイト利用制度における研究者と審査者の役割分担を示す。Eurostat ハンドブックは図 4 における「持ち出し基準（ポリシー）」に相当し、持ち出す分析結果が満たすべき要件を規定している。この安全性基準を踏まえ、研究者は分析結果に必要な秘匿処理を施し、審査者は基準の要件が満たされているかを審査する。しかし、秘匿処理した分析結果のみを渡されても審査者はその安全性を判断することができない場合が多い。

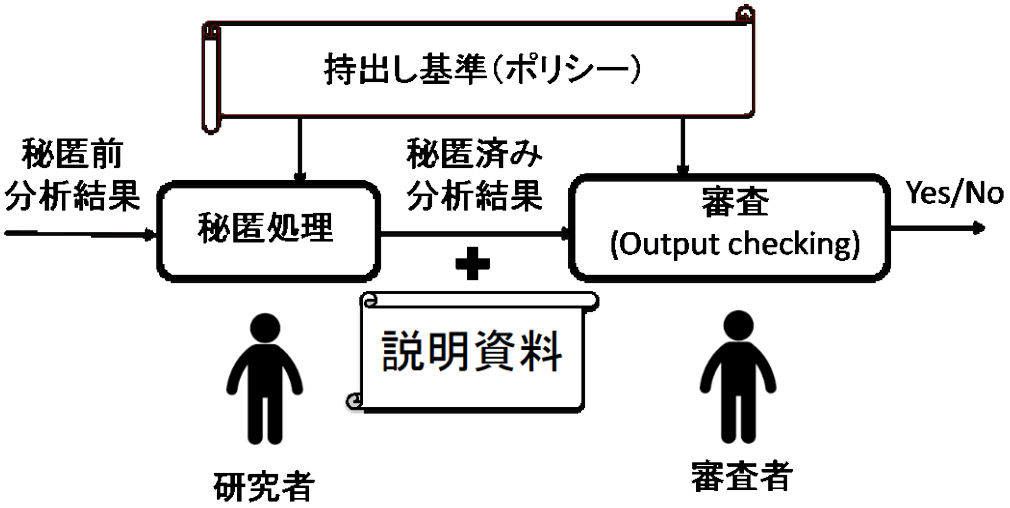


図 4. 安全性審査のための研究者と審査者の役割分担.

したがって、オンサイト利用制度においては、持ち出すデータの形式に応じて、研究者が作成すべき説明資料にどのような情報が提供されるべきか、分析結果の安全性を検証するための説明資料の要件も明示する必要がある。原則として、以下の情報は全ての持ち出しデータについて提出する必要がある。

- 持ち出しデータの使用目的
- 持ち出しデータのファイル名とその形式
- データに含まれる変数の説明
    - 利用者が定義した変数については導出に用いた式、分析方法の説明を含むこと

4.2　表データ

度数表、集計表は基本となる分析結果である。度数表を公開する場合、ある調査客体（これを攻撃対象の客体と呼ぶ）の属性情報を知る攻撃者が、度数表の特定のセルと攻撃対象の客体を結びつける危険性が存在する。攻撃対象の客体が属する度数表のセルの値（度数）が 1 であり、かつそのデータが全数調査であれば、攻撃者はその対象とする人物を一意に識別したことになる。もし、攻撃者が対応する数量表を入手できれば、度数 1 に対応する数量表のセルの情報から、攻撃対象の客体の個票値を知ることができてしまう。そのような度数 1 のセルからの情報漏えいの例を図 5 に示す。図の上の度数表の (M_2, P_3) に相当するセルの度数は 1 である。もし、同じカテ

ゴリー属性の区分をもつ数量表が入手できれば、識別された客体の収入が 22 であることが分かる。

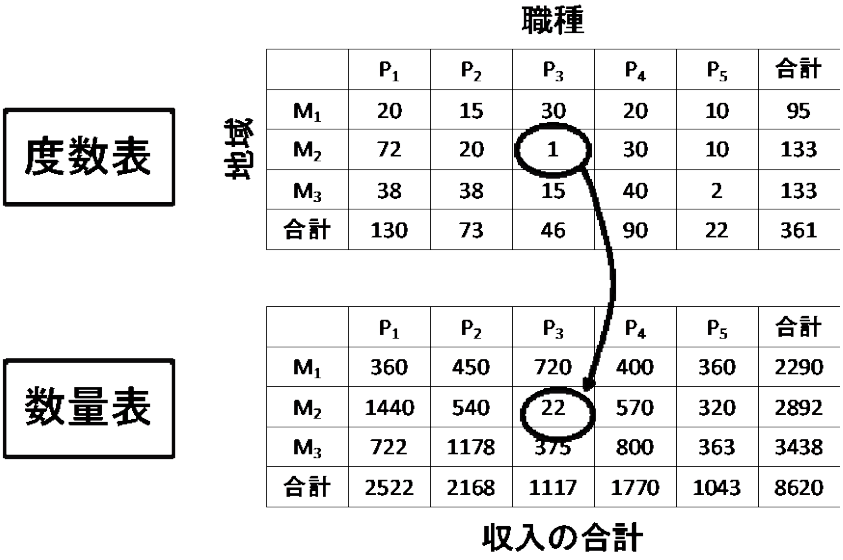
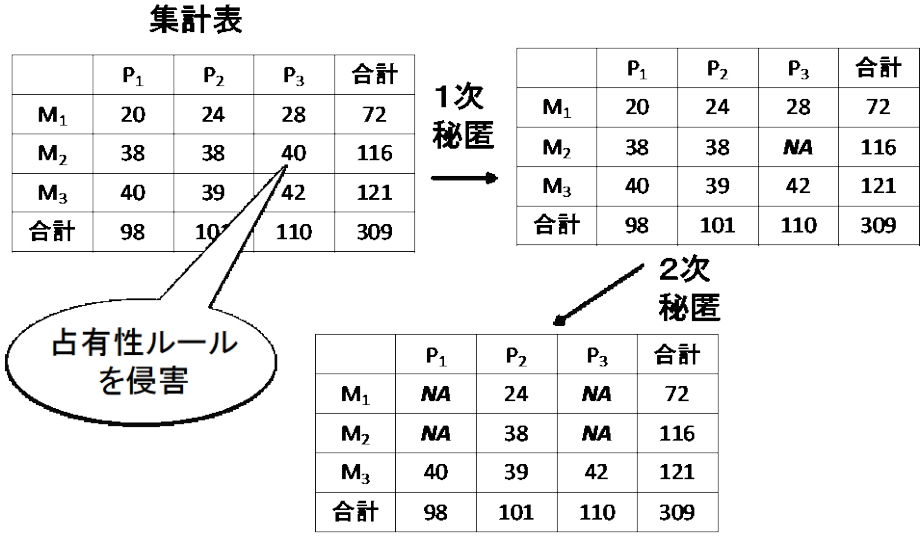


図 5. 表データからの情報漏えい

このような表データの場合、与えられた機密性ルールを各セルに適用し、値を秘匿すべき機密セルを決定する。表データの秘匿処理は機密セルの値を隠すことであり、図 6 の集計表の例が示すような 2 段階のセル秘匿処理を行う。図 6 のような集計表の場合、特定の客体の個票値がセル値の大部分を占めていないかを占有性ルールで確認する。客体の個票値の分布に極端な偏りがあり、特定の客体の値が他の客体よりも突出して大きいときは、大まかな値の推論が可能であり、そのような確率的な情報漏えいに対する防御策が必要だからである。



	P_1	P_2	P_3	合計
M_1	20	24	28	72
M_2	38	38	40	116
M_3	40	39	42	121
合計	98	101	110	309

	P_1	P_2	P_3	合計
M_1	20	24	28	72
M_2	38	38	*NA*	116
M_3	40	39	42	121
合計	98	101	110	309

	P_1	P_2	P_3	合計
M_1	*NA*	24	*NA*	72
M_2	*NA*	38	*NA*	116
M_3	40	39	42	121
合計	98	101	110	309

図 6. 表データの 2 段階のセル秘匿処理

もしこのルールの条件を侵害すると機密セルと判断され、1 次秘匿処理では機密セルの値を削除する。ただし、表データの場合、行計、列計の関係式から秘匿したセル値を復元することは容易なため、追加で秘密でないセルの値も秘匿する。この処理を 2 次秘匿処理と呼ぶ。

しかし一般には、表データにはセル値が満足すべき複数の制約条件があるので、秘匿したセルの取り得る可能な値の範囲が狭い範囲に絞り込まれる可能性がある。よって、1 次秘匿したセルの取りうる値の上限と下限を線形計画法で計算し、秘匿インターバルと呼ばれる可能な値の幅が与えられたしきい値より大きいことを確認することが必要である。

Eurostat ハンドブックでは、機密セルの秘匿インターバルの幅が狭く、値が絞り込まれてしまうリスクを指摘しつつも、それを防止する具体的な対策が安全性基準に盛り込まれていなかった。したがって、我々の安全性基準では、秘匿インターバルの最小幅に関する条件を追加することとした。また 1 次秘匿セルを決定する占有性ルールの確認には、各セルの第 1 位、第 2 位の客体の値の情報が必要であり、説明資料として提出することを分析者に義務付けることとした。

4. 3　線形回帰係数、非線形回帰係数

線形回帰係数、非線形回帰係数とは、複数の変数の関係性を式で表現した際の各係数である。例えば線形回帰であれば、図 7 にあるように、いくつかの個票レコードについて年齢と年収を点で図示したとき、年収を y 、年齢を x として線形回帰分析を行うと、年収 y と年齢 x の間に $y = 15x + 200$ という関係式が当てはまることがわかり、線形回帰係数は 15 及び 200 となる。

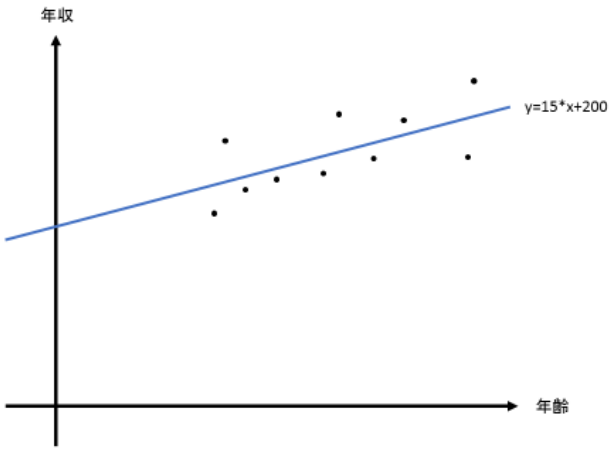


図 7. 線形回帰分析（係数）の例

Eurostat との違いは、線形回帰係数、非線形回帰係数は似たリスクを持っているため簡単のため同じ項目として纏めたこと、およびチェック基準は、その他の統計値と異なり、Eurostat における経験則ではなく原則ルールを用いている。これは経験則では回帰係数を完全な形で持ち出すことができず、利用者のニーズを満たすことが難しいと判断したためである。また、Eurostat では（非）線形回帰係数のリスクについて具体的な説明が無いため、ここでは具体例を挙げて説明する。

一般に（非）線形回帰係数は、多くの個票レコードの全体の傾向を表すものであり、線形回帰係数から個票値が知られてしまうようなリスクは一般的に低いと考えられる。しかしながら、

・　入力する個票値の数が少ない場合
・　個票値の取り得る値が限られる場合

では、個票値を推測できるリスクが無視できない。例えば、上記の例において２人の調査客体から線形回帰係数を作成し、片方の客体がもう片方の客体の年齢を知っているような場合、もう片方の客体の年収が分かってしまう。また、個票値の取り得る値が限られる場合は出力の係数になるような個票値が一意に定まってしまう場合がある。

例えば、図 8 の年齢と「年収が 2000 万以上か否か」の 2 値からなる線形回帰分析を考える。年収 2000 万円以上の人が殆どいないこと、及び 80 歳の客体が 1 人しかいないことを知っていると、80 歳の客体が 2000 万以上のとき（図 左）とそうでないとき（図 右）で大きく回帰分析結果が異なるため、回帰分析結果を見ることで 80 歳の客体の年収が 2000 万以上か否かがわかってしまう。

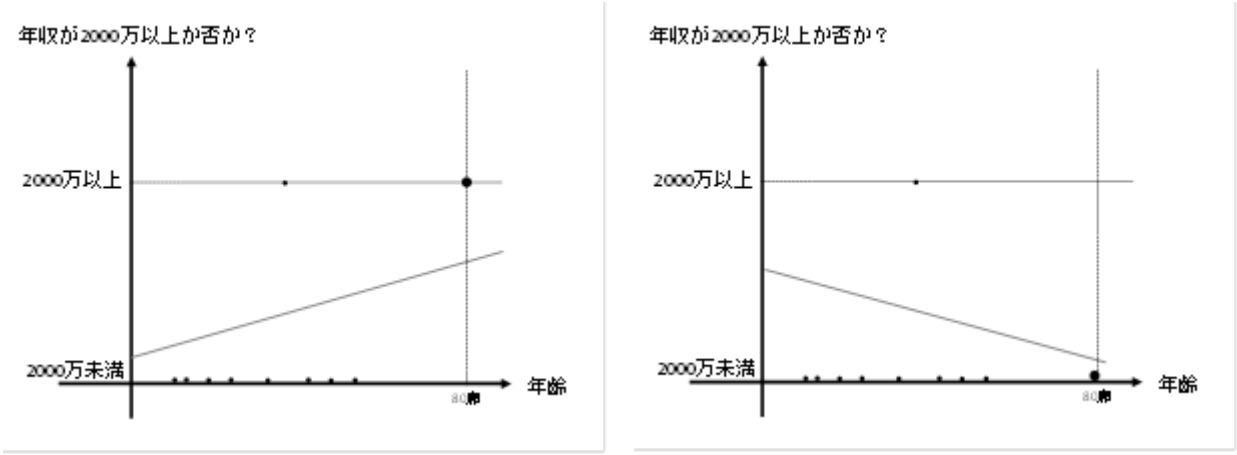


図 8. 個票値の取りうる値が限られる時のリスク例

4.4　パーセンタイル値

xパーセンタイル値とは、通常、複数の値を小さいほうから順に並べ、全体を 100 と見たときに、下から数えてx番目の数である。しかし Eurostat ハンドブックにおける“パーセンタイル値”は、明確に書かれてはいないものの違うものを指していると考えられる。そのため、チェック基準ではこの Eurostat ハンドブックにおける“パーセンタイル値”を、パーセンタイル値そのものではなくそれによって分類された数量表と定義しチェック基準を策定した。本来のパーセンタイル値は 4.8 節にて論じる。

通常のパーセンタイル値はある個票値そのもの、もしくは 2 つの個票値の平均が出力されるため個票値を推測されるリスクは大きい。一方で、パーセンタイル値そのものを持ち出すよりも、パーセンタイル値で分類された何らかの数量表を持ち出すことが多いと考えられる。例えば、全体を 4 つに分け、それぞれのグループに含まれる個票レコードの年収の総和を計算するようなものであり、以下のような出力である。このような出力であるとき、そのリスクは数量表と同等となる。

パーセンタイル値	0-25	26-50	51-75	76-100	Total
総和	100	200	250	450	1000

実際に Eurostat ハンドブックにおけるパーセンタイル値の安全性基準には、本来のパーセンタイル値では満たすことのできない度数 10 以上とあることからも、明示的には書いていないものの、パーセンタイル値そのものを持ち出すことはあまり想定されていないと考えられる。

4.5　最頻値

最頻値とは、ある複数の値のうち、最も出現する数である。例えば {1, 1, 2, 3, 4, 4, 4, 5} の最頻値は 4 であり、{東京都, 東京都, 東京都, 神奈川県, 埼玉県}の最頻値は東京都である。最頻値のチェック基準は Eurostat ハンドブックと同様とした。

少ない個票レコードから最頻値を計算した場合、個票値がわかってしまう場合がある。また、多くの個票レコードから最頻値を計算したとしても、すべてが同じ値をとる場合は個票値と同じになるため、これも避ける必要がある。

4.6　総和、平均、集中度

Eurostat ハンドブックでは平均、総和、指数、集中度、分布の高次モーメント、相関係数、要約統計量など多種な統計量に対して基準が定められているが、それぞれの統計量の違いが曖昧であること、チェック基準が似通っていること、および利用者・審査者が扱いやすいように、それらの統計量を整理し、平均、総和、集中度などの「平均のような、複数の数を足し合わせて得られるような統計値」と、次節で記述するその他の統計量との 2 つに整理した。もし、ある値を本節か次節のどちらに分類するか困った場合は、要約統計量として扱うこととする。

平均、総和、集中度などでは、値が 1 つもしくは数個の個票値によって決まるような場合を避けるべきである。例えば、2 つの個票値から平均を計算した場合、片方の個票値を引くことでもう片方の個票値を得ることができてしまう。また、ある 10 人の年収で 1 人だけ飛びぬけていることがわかっている場合、10 人の年収平均を 10 倍することで、飛びぬけた 1 人の年収を推測することが可能となってしまう。これらのリスクは一般に、度数ルールや占有性ルールなどを用いて確認する。

4.7　分布の高次モーメント、相関係数、要約統計量及び検定統計量

ここでは指数等を含む非常に広い概念について言及する。これらは複数の個票値からある特徴を表す値を計算するものである。そのため、そのリスクは出力（例えば相関係数など）からどの程度の個票値が推測可能であるかということが重要であり、出力からどの程度入力値が取りうる値があるのかを測る尺度である自由度を用いてリスクを測ることとする。

4.8　非審査対象（持ち出し不可）のデータ

単一の調査客体に関するデータは、持ち出し不可とする。また持ち出すデータから容易に作成できるデータも、持ち出し審査の対象とはしない。以下のデータは審査の対象とはせず、持ち出し不可とする。Eurostat ハンドブックで持ち出し不可の値にパーセンタイル値を加えたものとなっている。

- 最大値、最小値、パーセンタイル値
- 中間値（同じ中間値をもつ客体数が十分な数あれば、提供可の可能性あり）
- グラフ
- 推定残差

最大値、最小値、中間値は単一客体に関するデータである。またパーセンタイル値は多くとも 2 つの客体から算出されるデータである。グラフは元の数値データから計算可能なので審査対象とはせず、研究者は持ち出し許可を得たデータから生成するものとする。また、推定残差は線形回帰式などの統計モデルと実際の値の差であり、統計モデルと同時に持ち出すことで個票値がわかってしまうため、持ち出し不可とする。

ただし、上記の非審査対象のデータであっても、個票値の推測が十分に難しいと判断される場合は持ち出し可能とする場合がある。一例としては、最大値・最小値が一般的常識から推測可能である場合や、グラフが細かい数値を読み取れないように曖昧に書かれている場合、推定残差を計算する前の元の値自体が、そもそも 10 以上の客体からなる平均だった場合などが挙げられる。

5　審査作業の効率化

オンサイト利用の安全性審査はデータを持ち出す利用者と審査担当者の共同作業であり、円滑な遂行のためには、審査に必要な追加資料を安全性基準の中に明確に示す必要がある。また追加資

料の準備には非常な労力を要するため、表データに関してその作業を軽減する秘匿処理ツールを開発したのでその概要を示す。

5.1　審査自動化の必要性

表データの秘匿処理を手作業で行うのは 2 つの理由で困難である。第 1 は、4.2 章で説明した機密セルの秘匿インターバルを確認するには、その上限値、下限値を求めるために線形計画法の問題を解く必要があり、更に情報損失を最小化するために最適な 2 次秘匿セルの組み合わせを決める計算量は膨大である。第 2 は、集計表の安全性を占有性ルールで確認するには、審査者が表データの元の調査票情報を参照する必要がある点である。図 9 は、集計表に占有性ルールを適用する手順を示す。研究者はまず、個票調査票の複数のレコードの値を合計して集計表を作成し、審査担当者に提出する。しかし、審査者は表の各セルの安全性を検証するためにセルに寄与する客体の個票値が必要になる。

したがって、占有性ルールの審査においては、審査対象の数量表に加えて、説明資料として各セルの第 1 位、第 2 位の客体の値の表を追加で提出する必要がある。これらの表を作成するために、利用者は今までの分析とは異なる処理を行う必要があり、利用者に対して過大な負荷となる可能性がある。

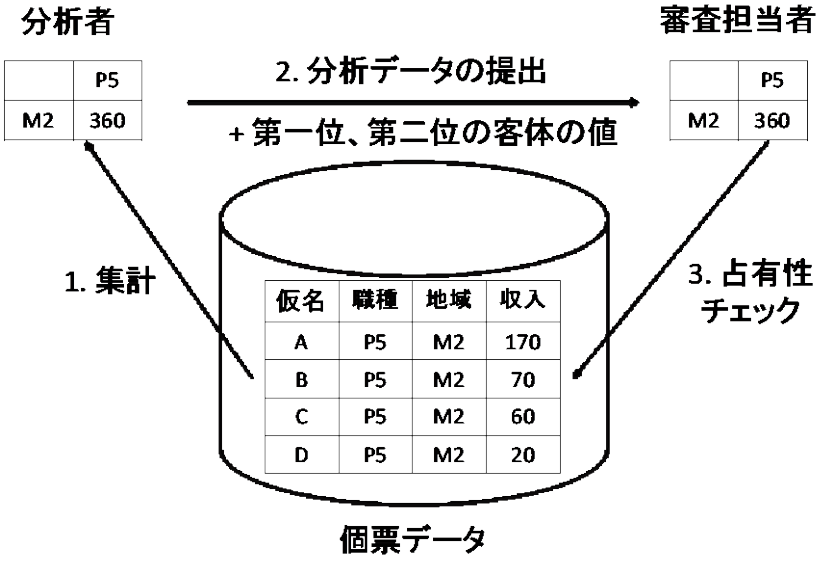


図 9. 占有性ルールの審査プロセス

オランダ統計局が開発した τ-ARGUS (Statistics Netherlands, 2018)の表データの 2 次秘匿処理を行う機能を提供し、秘匿インターバルを計算する問題を解決することができる。しかし、τ-ARGUS は利用者として統計局内の職員を想定しており、作成した秘匿処理済みの表データの安全性を証明するための追加情報を作成する機能は備えていない。

5.2　R による秘匿処理ツールの開発

我々は τ-ARGUS の課題を解決するために、R 言語で表データの秘匿処理ツール (Minami & Abe, Statistical Disclosure Control for Tabular Data in R, 2017) を開発した。研究者が R で表データを作成する場合、その表データに対してそのまま秘匿処理を行うことが可能である。我々のツールは、τ-ARGUS と同等の機能を R 関数のライブラリとして提供し、さらに審査に必要な追加の情報を出力する。

例として度数表に 1 次及び 2 次秘匿を実行するフローを示す。図 10 の秘匿処理関数は、元の表と、度数閾値、行／列占有度閾値、秘匿インターバル閾値の 4 つのセキュリティに関するパラメータを入力とし、2 次秘匿された表を出力する。また同時に、説明資料として、対応する度数表、1 次秘匿された表、さらに秘匿セルの秘匿インターバルの情報を出力する。

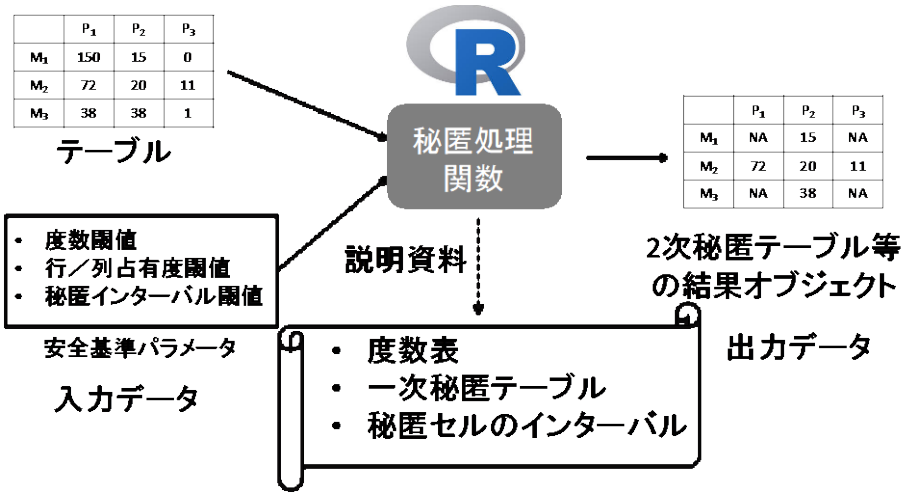


図 10. R 言語による秘匿処理関数

5.3　持ち出しデータのサイズ制限と審査頻度

オンサイト施設を利用する研究者は、分析途中の中間成果物の持ち出しを申請すると予想される。しかし、大量のデータの審査は時間を要し、審査部門の限られた人員を考慮すると、本制度の想定する期間内に審査を処理できない可能性が生じる。また 5.2 章で紹介した表データの秘匿処理ツールは表データのサイズが増大するにしたがい、著しく計算時間が増加する。

この問題を解決するために、2 つの審査の運用方針が考えられる。一つは、研究者が持ち出すデータのサイズに上限を設ける方法である。そして、制限を越えた申請については、サイズ内に収まった他の申請を先に優先することとする。この方針により、サイズ制限を守る審査については、審査終了までのスケジュール管理が容易となり、円滑な審査運営が可能になると考えられる。ただし、サイズ制限を設けるにあたり、施行期間中に利用者の要望を適時反映し、研究者の目的を妨げない上限値を選ぶことが大切である。

もう一つは、審査を実際に行う頻度を、研究者の秘匿処理に対する習熟度に応じて変える方法である。本制度においては、2.2 章で述べたように善意の研究者を想定するが、研究者がなんらかのミスにより機密情報を持ち出すことは防止しなければならない。そしてそのような機密情報の漏えいは秘匿処理に習熟しない研究者のミスによる可能性が一番高い。したがって、本制度を利用したばかりの研究者に対しては、持ち出す全てのデータについて審査を行い、利用頻度が多くなるにしたがい、それまでの審査実績を考慮し、段階的に審査頻度を下げていくのが妥当と考えられる。

審査頻度を考える対象としては、審査単位、審査の中のデータ単位、また個々のデータ（例えば数量表のセル）単位等様々な粒度が考えられ、施行期間中の審査案件を観察しつつ、研究者の習熟度合いを適切に反映した審査頻度の変更方法を決定する必要がある。

6　検討課題

オンサイト利用制度のデータ持ち出しの１次審査に採用した Eurostat の経験則はルールが明示的に定められており、実施するための十分な具体性を有する。それに対し、２次審査に用いる原則ルールは安全性の指針を抽象的に述べるに留まり、また他国における原則ルールの実際の運用事例はほとんど公開されていない。そこで本章では、今後オンサイト利用制度の運用を通して検討すべき重要な課題を議論する。

6.1　占有性ルール審査の実施可否の決定ルール

数量表の機密セルを決める占有性ルールの審査は利用者及び審査者の両者に大きな負担がかかり、自動検査ルールの導入も必要になる。一方、個人または世帯に関する調査の場合、占有性ルールで想定する情報漏洩のシナリオは現実的でないと思われる。したがって占有性ルール審査の実施可否を決定する客観的なルールの確立が望まれる。その決定に重要な役割を担う属性情報に関する 3 つの概念を紹介し、それらを機軸とする実施可否の決定手順をまとめる。

6.1.1　外観識別性の評価

4.2 章では、度数表に度数の低いセルが存在する場合、なんらかの方法でその度数表のカテゴリー変数の情報を知り得る攻撃者がそのセルに含まれる客体を個人識別するリスクを説明した。このとき攻撃者が他人のカテゴリー変数の情報を知り得る可能性があるかどうかを判断する指標として、**外観識別性**という概念を用いる。

外観識別性は定性的な概念であり、他人のある情報を外部から観察することで知り得るその容易性を示す。例えば、物理的に観察することができる住所情報、または年齢差の大きい夫婦や３つ子以上のいる世帯の外観識別性は高いと広く認知されている。また、物理的な観察を伴わなくとも、既にオープンデータに近い形で一般に流通している情報も外観識別性が高いと判断される。例えば、米国では、投票者リストという形で一般市民の郵便番号、性別等が容易に入手可能であり、州によっては州の職員の年収が公開されている。

しかし、研究者が分析する調査票情報には様々な属性情報が含まれ、その中には外観識別性が低い情報も多く含まれる。その場合、経験則をそのまま適用すると、外観識別性が低い属性でクロス集計した場合でも、セルの度数が低ければ持ち出し不可と判断され、研究者の利便性を大きく損なうことになる。したがって、外観識別性が高いと判断される属性リストを定義し、度数チェックする場合はそのリストに含まれる属性変数のみを考慮したクロス集計で安全性審査を行うことが望ましい。そのためには、属性情報の外観識別性を客観的に決める手法、手続きの整備が必要であり、社会学的な実証的手法、有識者パネルによる意思決定方法を検討する必要がある。

6.1.2　情報機密性の評価

数量表に適用する占有性ルールは、集計表のセル変数の**機密性**を守ることが目的である。しかし、全てのセル変数が機密性の高い情報を扱うわけでなく、一律に経験則を適用すると多くの安全な情報が持ち出し不可になると予想される。したがって、原則ルールに従えば、機密情報を扱う数量表のみ占有性ルールの対象とすべきであり、そのためには機密性の高い変数のグループを定義し、そのグループに含まれる変数のみを考慮したクロス集計に占有性ルールを適用すべきである。

ただし、情報の機密性は 6.1.1 章の外観識別性と同様に客観的な定量化が困難な概念であり、情報漏えいのリスク評価に関する学術的な研究の発展、及び社会における合意形成が不可欠となる。占有性ルールの審査は、5.1 章で説明したように、その実施は利用者、審査者の両方に大き

な負荷を課すことになる。したがって、機密情報を含まない数量表に対し、占有性ルールの審査を省略できるメリットは大きく、制度の円滑な運営が可能になる。

6.1.3　序列識別性の評価

この問題も 6.1.2 章同様に、数量表の占有性ルールに関する課題である。占有性ルールで想定する情報漏えいには、あるセルに含まれる客体（特に第 2 位）がそのセルに含まれる他の客体の識別子（名前等の ID 情報）とそのセル値への貢献度でみた序列に関する知識をもつことが前提条件となる。

この前提を精査すると、以下の 2 つの前提に分けることができる。1 つめは、同一セル内の他の客体を識別するには、そのセルのクロス集計を行うカテゴリー変数の外観識別性が高いという前提である。企業情報については、そのような情報が公開されている可能性が高く、その場合の外観識別性は極めて高い。2 つめは、セル内の客体が識別された場合にそれらの複数の客体間の序列情報が得られるという前提である。このような概念を指す用語はまだ確立していないが、本論文ではそれを**序列識別性**と名づける。

序列識別性は、その情報（例えば企業の売り上げ）の過去データが公開されていて、過去の序列が分かっており、一定時間内にはその順位の変動があまりないと考えられる場合、または他の外観識別性の高い情報と対象とする情報の間で序列の関係性が保存される場合である。例えば、所有する家の占有面積、車の価格帯等の序列と、世帯の収入の序列はかなり近いことが予想される。

この序列識別性の厳密な定量化もやはり困難と考えられる。しかし、様々な属性情報間の相関性を評価する研究は多数あると考えられ、序列に関する相似性が成立する外観識別性の高い属性情報の種類の数を大まかな指標と考えることはできる。また外観識別性、機密性の評価同様、序列識別性に関する専門家の議論、合意形成の場を設けることも重要である。

6.1.4　実施是非の決定手順

外観識別性、機密性、そして序列識別性の定量的評価が困難な作業であるが、この属性情報をこの 3 つの軸で分類できるメリットは大きく、審査時の占有性ルールの審査実施の是非が明確に整理できるので、そのフローを参考までに図 11 に示す。

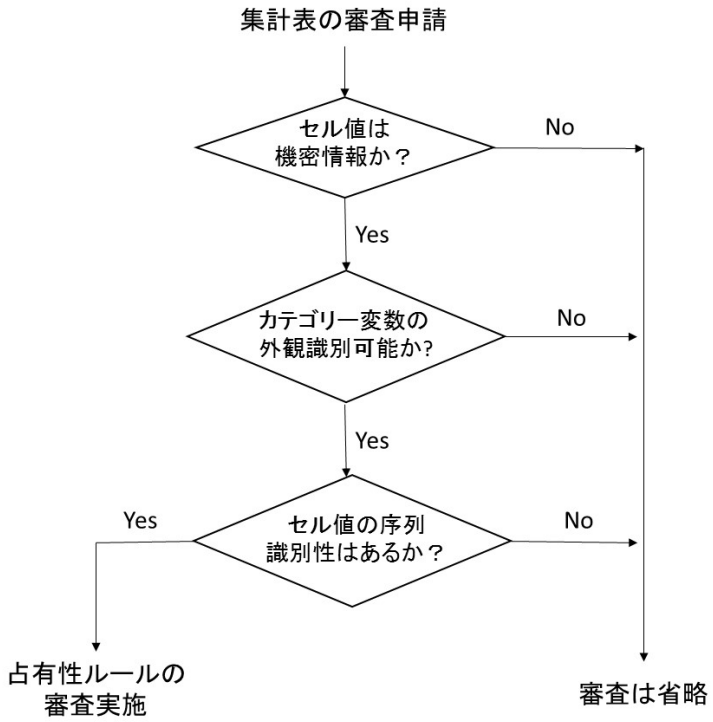


図 11. 占有性ルール実施の必要性の決定フロー

6.2　差分開示への対応

類似する複数の度数表、数量表を利用者が持ち出す場合、その差分から、より詳細な度数、集計が明らかになる問題がある。利用者単位の持ち出しについてはある程度の対応を行っているものの完全ではなく、また異なる複数の研究者が偶然そのような類似する表を別々に発表するリスクも残っている。

その対策として、一般的に広く利用される年齢、収入、売り上げ等のカテゴリー変数に関しては、審査基準のほうで標準となる区分方法を指定することで、差分開示のリスクを低減する方法や、持ち出しの際に分析結果には大きく影響しないレベルのノイズを付加するなどの対策が考えられる。今後は、広く利用されるカテゴリー変数や、どのような区分方法が多くの調査にとって十分であるかの調査、及びノイズを付加した場合のデータ分析への影響、差分開示のリスク評価をあわせて行う必要がある。

6.3　標本調査における抽出率を考慮した度数閾値の決定方法

4.2 章で説明した度数表の審査において、経験則では度数閾値として 10 の値を全ての種類の度数表に対して適用する。しかし、標本調査のデータから作成された度数表の場合、度数が 1 であっても元の母集団では複数該当する客体が存在する可能性があり、つまり標本データにおける一意（に識別）が母集団一意を意味しない。したがって、標本調査の場合、その抽出率を考慮して度数閾値を緩和し、より小さい閾値を使うのが適切と考えられる。

実際、標本一意のときに母集団一意であるリスクを評価する学術研究は多数存在し、それらの研究では母集団の分布に何らかの統計モデルを仮定し、標本一意のときに母集団一意となる確率を求めている。もし、その確率 p が導出できれば、標本データにおけるセルの度数が n の場合、

母集団における対応するセルの度数は n/p と見積もれる。したがって、母集団で最小度数 10 を必要とする場合、その条件は $n/p < 10$ であり、最小度数は $10p$ で十分になる。

今後は、本制度で対象となる調査データの母集団、特に稀な値をとる分布の周辺部分に関する統計モデルの適切な選択、及び新しい統計モデルを検討し、母集団がある程度分かる標本データについて実証的な評価をあわせて行う必要がある。

6.4　標本データにおける占有性ルールの修正

6.3 章の度数ルール同様、数量表の場合も標本データに占有性ルールをそのまま適用すると必要以上に条件が厳しくなる場合がある。例えば、ある標本データから作成した数量表のセルが 2 つの客体を含み、それぞれの値が 100 と 10 とする。この場合、客体が 2 つしかないため、p%ルールの p にいかなる正の値を選んでもp%ルールに抵触する。しかし、もしこの数量表が値 100 については重み 4、 値 10 については重み 7 で標本抽出されているとすると、もとの母集団は {100, 100, 100, 100, 10, 10, 10, 10, 10, 10, 10}に近い分布をとると考えられ、この場合、p%ルールにおける安全性は満足される。したがって、標本データの抽出率（一般には重み係数（ウエイト））を考慮した占有性ルールの修正方法の考案が必要と思われる。

7　まとめ

本論文では、オンサイト利用の安全性基準策定への取り組みの概要を紹介した。我々の安全性基準は Eurostat ハンドブックの明示的な経験則を基本としながらも、表データの安全性要件に関する詳細を補足し、さらに持ち出すデータの安全性検証に必要な説明資料の要件を各データ形式ごとに明確化した。さらに審査作業および説明資料の準備作業を軽減するための表データの秘匿処理ツールを R 言語で開発した。ただし、原則ベースの審査のための検討事項は多く残されており、今後はオンサイト利用制度の実際の運用を通して審査の知見を蓄積していくとともに、専門家による情報漏洩リスク、外観識別性等の整理に関する包括的な取り組みが必要である。

参考文献

（独）統計センター. (2018). 調査票情報のオンサイト利用. 参照先: https://www.nstac.go.jp/services/on-site.html

Daniel, C. B.-J. (2012). The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now. *SSRN Electronic Journal*.

Hundepool, A., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., Wolf, P.-P. d., & Domingo-Ferror, J. (2012). *Statistical Disclosure Control.* Wiley.

Hundepool, A., Josep, D.-F., Franconi, L., Sarah, G., Rainer, L., Jane, N., . . . Peter-Paul, D. W. (2010). *Handbook on Statistical Disclosure Control.* Luxembourg: Eurostat.

Maurice, B. (2009). *Guidelines for the checking of output based on microdata research.* ESSNet.

Minami, K., & Abe, Y. (2017). Statistical Disclosure Control for Tabular Data in R. *Romanian Statistical Review, 4*, 67--76.

Minami, K., & Abe, Y. (2018). A First Step towards Statistical Disclosure Control on Multiple Tables Under the Presence of Differential Attacks. *6th International Joint Conference New Challenges for Statistical Software - The Use of R in Official Statistics (uRos2018).* Hague, Neitherlands.

Statistics Netherlands. (2018, 3 08). *τ-ARGUS homepage*. Retrieved from http://research.cbs.nl/casc/tau.htm