資料６

持ち出し審査のイメージ・内容について（素案）

総務省政策統括官（統計基準担当）、統計局
独立行政法人統計センター

利用者から、リモートアクセス型オンサイト施設から研究成果を外部に持ち出したい旨の申出があった際の持ち出し審査のイメージと内容について取りまとめた。

１　持ち出し審査のイメージ

（１）利用者は、持ち出し審査に係るチェックシートの項目に従い、統計表や分析結果について、秘匿性に問題がないか自己チェック

（２）統計センターにおいて、同じチェックシートの項目に沿って同様に確認

２　持ち出し審査の内容

分析結果の種類とチェック内容について、ESS（European Statistical System）Net SDC(Statistical Disclosure Control）の”Guidelines for the checking of output based on microdata research”の経験則に関する記述を参考に作成。

分析結果の種類	チェック内容
度数表	・10 未満のセルがないこと ・行計又は列計の 90%超のセルがないこと
数量表、パーセンタイル値、集中度	・10 未満のセルがないこと ・行計又は列計の 90%超のセルがないこと ・各セルにおいて、50%を超えて寄与する調査客体がないこと
最大値、最小値	・不可（通常ただ 1 つの調査客体を指していることから）
最頻値	・行計又は列計の 90%超のセルがないこと
平均、指数、比率、指標	・10 以上の調査客体から算出した値であること ・算出した値において、50%を超えて寄与する調査客体がないこと
分布の高次モーメント[1]	・自由度が 10 以上であること
グラフ	・不可（他の許可を受けた分析結果から作成すべきことから）
線形回帰係数、非線形回帰係数	・1 つ以上の推定係数を削除していること（例：切片）
推定残差	・不可
要約統計量及び検定統計量[2]	・自由度が 10 以上であること
相関係数	・10 以上の調査客体から算出した値であること

○　分析結果と併せて、使用したデータ、原変数及び利用者自身が作成した変数の説明、分析対象（対象の選択基準及び数）等の情報が必要。

○　分析結果以外（プログラム等）を持ち出す場合についても、個別データに関する記述やデータが含まれていないことを確認する必要。

---

1 分散、歪度、尖度等

2 決定係数、変動係数、分散、情報量規準、t 検定、F 検定、χ^2検定、Wald 検定、Hausman 検定等

参考

ESSNet SDC "Guidelines for the checking of output based on microdata research"の一部を翻訳

2　アウトプット・チェッキングの規則

2.1　アウトプットの分類

前述したとおり、RDC 動物園は、すべてのアウトプットを限定された数のクラスに分類することによって多少構造化することができる。下表は、様々なアウトプットの種類を列挙したものである。各クラスには、安全か安全でないかが表示されている（「安全/安全でない」の分類の説明については、第 1.2 項を参照されたい）。

統計の種類	アウトプットの種類	分類
記述統計	度数表	安全でない
	数量表	安全でない
	最大値、最小値及びパーセンタイル値（中央値を含む）	安全でない
	最頻値	安全
	平均、指数、比率、指標	安全でない
	集中度	安全
	分布の高次モーメント（分散、共分散、尖度、歪度を含む）	安全
	グラフ：図形による実データの表示	安全でない
相関及び回帰分析	線形回帰係数	安全
	非線形回帰係数	安全
	推定残差	安全でない
	推定値から得られる要約及び検定統計量（R^2, X^2その他）	安全
	相関係数	安全

2.2　全般的な経験則

前述したとおり、経験則は単純かつ明快（及び厳格）なルールに基づいている。これらのルールは、アウトプットの様々なクラスによってもわずかしか異ならないことから、全般的な経験則を設定することができる。この全般的な経験則がまず提示され、その後で、アウトプットの各クラスに適用するルールを記述する際に、この全般的な経験則の解釈がアウトプットの個別のクラスに関して与えられる。

全般的な経験則は、4 つの要素で構成されている。

1　**10 ユニット**：表形式による及びそれに類似するアウトプットはすべて、提示されたセ

ル又はデータ点の基礎となるユニットを少なくとも 10（加重されていないもの）有すべきである。そのようなルールを一般的に「閾値ルール」（セル数値は、特定の閾値を超えなければならない）と呼ぶ。

2　**自由度 10**：モデル化されたアウトプットはすべて、自由度 10 を持つべきであり、また､モデルを生成するのに少なくとも 10 ユニットが使われていなければならない。
自由度＝（観測数）-/-（パラメーター数）-/-（モデルのその他の制約）

3　**グループ開示**：表形式による及びそれに類似するすべてのアウトプットにおいて、グループ開示を防止するための表の行又は列のユニット数の合計の 90%以上を保有するセルはない。グループ開示とは、表内の一部の変数（通常スパニング変数）が統計ユニットのグループを特定し、表内の他の変数が当該グループ内の各メンバーに有効な情報を暴露する状況を指す。個々のユニットは認識されないものの、情報はグループの各メンバーに関して有効であり、グループは認識可能であるために、機密保持違反となる。

4　**占有性**：表形式による及びそれに類似するすべてのデータにおいて、セルの最大寄与因子がセル合計の 50%を超えてはならない。

実際的な問題：占有性ルール

極めて単純に見える全般的経験則でさえも 1 つの大きな難点を抱えている。これは 4 番目の構成要素である占有性ルールに関するものである。

研究者は、このルールのチェックを受けるために、各セルの最大の寄与因子の値に関して追加の情報を提供しなければならない。これに伴い、研究者は余分な多くの労力を費やさざるを得なくなる。また、最大の寄与因子の値を公表すれば、多大な開示リスクを負うことから、公表前にアウトプットから開示につながる余分な情報を削除しなければならないのは明白である。

したがって、占有性ルールは経験則に含まれているものの、多くの国々おいては現在、積極的なチェックを行っていないというのが実情である。たとえそうであっても、研究者は、そのアウトプットを生成する際に、このルールを考慮に入れるべきであると指導されている。

通常、NSI は、特定の状況、例えば、以下のような状況においてのみ、積極的に占有性ルールをチェックすることを決定している。

- ビジネス・データに関する数量表
- 極めてセンシティブな変数に基づくアウトプット
- 歪みが極めて大きい分布を持った変数

しかし、原則ベース・モデルではなく、経験則のみに従いたいと思う人々は、占有性ルールに関するチェックをベスト・プラクティスと考えるべきである。

本章の残りでは、上記のとおり分類された各アウトプットを個別に検討するとともに全般的な経験則について論じ、さらに、原則ベースのモデルに関してより詳細な情報を提供していく。

2.3 クラス 1：度数表

2.3.1 経験則

度数表の場合、全般的な経験則の要素 1 及び 3 のみが適用される。

- 表内の各セルには、少なくとも 10 ユニット（加重されていないもの）が含まれなければならない。研究者は、このルールのチェックができるようにするため、表内に加重されていないセル数値を含めるべきである。
- 任意の行又は列のすべてのセルにおけるユニットの分布状況に関して、特定の行又は列の総ユニット数の 90%超を含むセルがないようにする（行又は列のたった 1 つのセルにユニットが集中してはならない)。このルールによって、グループ開示が防止される。

2.3.2 原則ベース・モデルに関する詳細情報

表内のすべてのセルは潜在的に安全でない。表を安全にするための一般ルールはない。しかしながら、上述したとおり、表をより安全にするための様々な選択肢は存在する。そのような状況において、多数の問題を考慮に入れるべきである。

- データ自体に開示リスクがあるかどうか（新たな変数に変換されたかどうか、詳細度その他)。
- データ又はサブセットを構成するユニットは識別することができるかどうか。
- データと経験則の要素 1（閾値)、3（グループ開示）及び 4（占有性）との近接性
- 寄与因子の等級序列が周知であるかどうか（換言すれば、最大値/最小値/最高値その他が周知である又は推測できるかどうか？)。
- セル・ユニットの選択。ユニットは人々、世帯、地域その他か？
- 標本の選択
- 加重

分析が行われる背景は重要である。考慮に入れるべき要素には、以下が含まれる。

- 地理的な細分化のレベル
- 産業/職業分類の詳細
- 世界との関係：国内活動対国際活動

最も重要な基準は、合理的に判断して識別され得る回答者が皆無になるようにすること及びグループに関して従前に利用できなかったいかなる機密情報も推測され得ないようにすることである。

「合理的に判断して」という言葉は、明示的に定義されていない。いかなる定義であっても批判に晒されることになるからである。また、原則ベースのアウトプット・チェッキングは、すべての定義を特定の文脈の中で捉えるということを目的としている。しかしながら、考慮に入れる要因の中には、以下のような内容の識別が含められている。

- 相当な量の時間と努力を必要とする。
- 大半の個人が取得するとは見込まれない又は容易に取得できないと考えられる追加の情報を必要とする。
- 一定の技術的能力を必要とする。

2.4 クラス 2：数量表

2.4.1 経験則

数量表の場合、全般的な経験則の要素 1 、3 及び 4 が適用される。

- 表内の各セルには、少なくとも 10 ユニット（加重されていないもの）が含まれているものとする。研究者は、このルールのチェックができるようにするため、表内に加重されていないセル数値を含めるべきである。
- 任意の行又は列のすべてのセルにおけるユニットの分布状況に関して、特定の行又は列の総ユニット数の 90%超を含むセルがないようにする（行又は列のたった 1 つのセルにユニットが集中してはならない)。このルールによって、グループ開示が防止される。
- すべてのセルにおいて、最大寄与因子がセル合計の 50%を超えてはならない。

2.4.2 原則ベース・モデルに関する詳細情報

セルの 10 ユニット・ルール（要素 1）が不適切となる状況があるかも知れない。例えば、研究者は、セル数値が要求される 10 ユニットという閾値を満たさない場合であっても、アウトプットは開示リスクを抱えていないと確信する可能性がある。この場合、研究者が立

証責任を負い、なぜそのように言えるのかについて説明しなければならず、アウトプットのチェックに責任を負う担当者は、このアウトプットをリリースできるかどうかについて決定することを義務付けられる。

この決定を下すに当たって、以下の要素が重要である。

- アウトプットを巡る背景
- 付帯情報（これは従前のアウトプットに依存する場合がある）
- 特にアウトプットに占有度の高い観測値（要素 4 参照）が含まれている場合、アウトプットをチェックする担当者が、アウトプットから特定の個人又は企業を特定することができるか？

この判断を下すに当たって、第 2.3.2 項のガイドラインを念頭に置くべきである。

2.5 クラス 3：最大値、最小値、パーセンタイル値

2.5.1 経験則

最大値と最小値は、通常ただ 1 つのユニットを指していることから、公表されない。

パーセンタイル値は、数量表の特別な場合として扱われる。各パーセンタイル・バンドは、その順位によって決定されるセルから構成される表のセルとして扱われるべきである。ユニットの順位が知られている場合、そのユニットに関する情報は収集することができる。これがどの程度有用である/開示リスクがあるかは、セルの規模やバンド幅次第である。この各値については、数量表に関する経験則が適用される。

2.5.2 原則ベース・モデルに関する詳細情報

パーセンタイル値については、数量表の場合と同様の原則が適用される。

原則ベース・モデルにおいては、最大値と最小値に関しても数量表に関するルールが適用される。換言すれば、最大値/最小値は、個々のデータ点と関連付けられない場合、公表することができる。

2.6 クラス 4：最頻値

2.6.1 経験則

最頻値は、全般的な経験則の要素 3（グループ開示）が満たされる場合は安全である。これによって、すべての観測が同一の値を持つ場合が防止される。

2.6.2　原則ベース・モデルに関する詳細情報

経験則が適用される。

2.7　クラス 5：平均、指数、比率、指標

2.7.1　経験則

平均、指数、比率及び指標については、n 個の観測値を単一の値に集約したものであることから、一般的な SDC（Statistical Disclosure Control：統計的開示抑制）ルールを適用することができる。数量表のセルについても同様の考え方、特に、全般的な経験則の要素 1 及び 3 が適用される。

- 単一の値はそれぞれ、少なくとも 10 ユニットを集約したものから生成されるべきである。（加重なし）研究者は、このルールのチェックができるようにするため、そのアウトプットに必要な情報を含めるべきである。
- 単一の値をそれぞれ公表する場合、当該集約値に含まれる最大寄与因子が合計の 50%を超えてはならない。研究者は、このルールのチェックができるようにするため、そのアウトプットに必要な情報を含めるべきである。

2.7.2　原則ベース・モデルに関する詳細情報

平均、指数、比率及び指標を評価するに当たって、以下の検討を行うことを考慮すべきである。

まず、指数算式を検討すべきである。

一般に、単純な指数は、所与の母集団内の統計ユニットに関して、個々の変数の値を要約したものである。

$$I=f(X, n)$$

アウトプットを評価するに当たって、指数算式(f)と母集団の規模(n)を特定しなければならない。

指数算式は、各ユニットに関する多くの属性が関係し、単一の値から遡って算出することが非現実となるような方法で各値を結合する形をとるなど、極めて複雑になる場合もある。

例として、下記のフィッシャー物価指数の複雑さを考えられたい。

$$P_F = \sqrt{P_L \cdot P_P} = \sqrt{\frac{\sum_{j=1}^{m} p_{1,j} \cdot q_{0,j}}{\sum_{j=1}^{m} p_{0,j} \cdot q_{0,j}} \cdot \frac{\sum_{j=1}^{m} p_{1,j} \cdot q_{1,j}}{\sum_{j=1}^{m} p_{0,j} \cdot q_{1,j}}}$$

したがって、指数がほんの 2、3 のユニットに基づき算出されているものではないと仮定した場合、データ変換の複雑性自体が、指数の値から個々の情報が開示されるのを合理的なレベルで防止していると言える。同様にして、一般に複雑な指数（すなわち、比率を含む単純な指数を複数結合したもの）は、単純な指数よりも開示リスクは低い。

しかしながら、指数算式は極めて単純なものにもなり得る。下記算式はその例である。

a) $I = \frac{\sum_{i=1}^{n} X_i}{n}$ ， b) $I = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{N} Y_i}$ ， c) $Range = X_{MAX} - X_{MIN}$

また、指数算式の 1 つ又は複数の独立変数の値は、たやすく公的に入手できる可能性がある。a)は算術平均に相当することに留意されたい。

例えば、a)及び b)において、分母は知られている可能性がある。このような場合、問題は分子（Σ*Xi*）の評価というところまで絞られる。これは、数量表（*X*が数量の場合）又は度数表（*X* が 2 種類の値しかとらない 2 分変数の場合）のセルをチェックする問題と同じである。

特に上記の 2 つ目のケースにおいて、度数表の事例に合わせて、*X*=｛0,1｝である場合、適用ルールは、ΣXi≧10 とすべきである。また、「0」の値を持つ度数も導き出せることから、
n－ΣXi≧10 ともすべきである。

例：*X*が 2 分変数の場合、 母集団の規模を n=20 とし、平均値を $\overline{X}$=0.7 とした場合、下記の度数表のセルに相当することになる。

X=0	*X-1*	合計
6	14	20

この事例において、*X* の平均が 20 ユニット（10 以上）から算出されており、*X* の値 1

に対応する度数が 14＞10（要素 1：度数表のセル数値ルール）となっている場合であっても、*X*=0 を有しているユニットが 6（＜10！）しかないことを導き出すことができる。しかしながら、この種の状況において、個人に関する情報が開示されるリスクが実質的に存在するかどうかについて検討しなければならない（度数表に関する第 2.3 項参照）。

c)のケースについては、レンジが最大値と最小値の線形関数であれば、その構成要素（最大値及び最小値）を公表できる場合にのみ、公表すべきである（第 2.5 項参照）。同様にして、既述のとおり、その他の比較又は分散指数を評価する際、算式（線形か否かを問わない）とその独立変数を検討すべきである。

しかしながら、関係する変数の知名度も検討すべきである。

「指数」カテゴリーは、多様な統計結果（平均値、合計値その他の含む）で構成されている。アウトプットをチェックする観点から見れば、指数は、実際的に極めて対処しやすい SDC 問題しか提起しないものの、理論的には、表形式のデータと類似している。したがって、同様の項目、特に、指数を算定する際に関係する変数の種類（社会指標によく見られるように、変数が公的に入手可能な場合）と n 個のユニットの亜母集団（センシティブ又は識別できるスパニング変数に従って選択される場合）の特徴を考慮に入れなければならない（第 2.3 項及び第 2.4 項参照）。

2.8 クラス 6：集中度

2.8.1 経験則

集中度は、経験則の要素 1（閾値） 、3（グループ開示）及び 4（占有性）を満たしている限り安全である。

2.8.2 原則ベース・モデルに関する詳細情報

結果が開示リスクを持たないことを示すことができる場合、CR10（最大の 10 ユニット）より下の集中度は許容される。具体的には以下を満たす場合がこの事例に該当する。

- アウトプットの基礎となっている標本に少なくとも 10 ユニットが含まれていること。
- 要素 4：占有性ルールを満たしていること。
- パーセンテージが小数点以下の値のない形で表示されていること。

2.9 クラス 7：分布の高次モーメント

2.9.1 経験則

高次モーメント（分散、歪度、尖度）は、経験則の要素 2 を満たしている限り安全である。

- モデル化されたアウトプットはすべて、少なくとも自由度 10 を持つべきであり、また、モデルを生成するのに少なくとも 10 ユニットが使われていなければならない。
  自由度＝（観測数）-/-（パラメーター数）-/-（モデルのその他の制約）

2.9.2　原則ベース・モデルに関する詳細情報

単純なルールが適用される。研究者と協議する際、自由度 10 未満であれば、アウトプットの統計値はいずれにしても疑わしいということに触れておくべきであろう。

2.10　クラス 8：グラフ（記述統計又は調整済み値）

2.10.1　経験則

グラフ自体はアウトプットとして許可されていない。基礎となるデータはアウトプットとして提出することができる。このアウトプットは、チェックを受けた後であれば、研究者自身の環境でグラフを再作成するために使用できる。

2.10.2　原則ベース・モデルに関する詳細情報

グラフ形式のアウトプットは、完全なマイクロデータへアクセスせずにそのグラフを再生すること、換言すれば、チェックを無事通過できる（表形式の）アウトプットからそのグラフを作成することができない場合においてのみ認められる。この場合、グラフは以下の条件を満たすべきである。

- データ点からユニットを識別することができない。グラフが変換又は調整済みのデータで構成されている場合、これは通常問題とはならない。
- 大きな影響を及ぼすような外れ値がない。
- グラフがデータの添付されていない「固定した」図として提出されている。つまりグラフが.jpg -.jpeg -.bmp -.wmf といった拡張子のファイルとして提出されるべきである。

2.11　クラス 9：線形回帰係数

2.11.1　経験則

線形回帰係数は、少なくとも 1 つの推定係数が開示されない場合、安全である（例：切片）。

2.11.2　原則ベース・モデルに関する詳細情報

回帰は、以下の条件を満たす限り、完全な形で公表することができる。

- 少なくとも自由度 10 を有している。
- カテゴリー（定性的）変数のみに基づいているわけではない。
- 単一のユニットに基づいているもの（例・企業 1 社に関する時系列）ではない。

2.12 クラス 10：非線形回帰係数

2.12.1 経験則

線形回帰係数と同様に、少なくとも 1 つの推定係数が開示されない場合、非線形回帰係数は安全である（例：切片）。

2.12.2 原則ベース・モデルに関する詳細情報

非線形回帰は、従属変数が離散変数の性質を有することから、線形回帰とは異なる。回帰が推定される場合、同じ回帰が観測値を 1 つ追加しても同じ回帰が繰り返される場合、回帰結果に関連する変数平均を用いることによって、特定の 1 つの観測値に関して情報を推測することが可能になるかも知れない。

しかしながら、実際上、この可能性は低い。モデルに含まれる観測数の変更は、以下の行為の結果である。

- 標本を明示的に変更する、又は
- 仕様を変更し、したがって、許容できない値（例：見当たらない）が脱落した形で観測値を見出す。

第 1 の行為の場合、（研究者が意図的に SDC ルールを回避しようとしない限り）獲得する情報は無視し得るレベルであることから、1 つの観測値につながる可能性は事実上低い。第 2 の事例の場合、仕様が変更されていることから、観測値の様々な数は関係がない。

2.13 クラス 11：推定残差

2.13.1 経験則

残差及び残差のプロットは一切リリースすべきではない。

2.13.2 原則ベース・モデルに関する詳細情報

経験則と同様、残差をリリースすべきではない。

例えば、モデルの堅牢性を実証するために、研究者が残差プロットのリリースに関する妥当な要求を行うことがある。しかしながら、残差プロットはリリースすべきではない。代わりに、研究者は安全な環境下でプロットを分析すべきであり、堅牢性を実証するため

にプロットの形状に関する説明を公表することができる。残差プロットをリリースする必要がある場合、上記グラフ（第 2.10 項）に関してこれを評価すべきである。

2.14　クラス 12：要約統計量及び検定統計量

2.14.1　経験則

以下の要約統計量は、モデル内に少なくとも自由度 10 がある限り、公表することができる。

- R2 及び変動
- 推定分散
- 情報基準（例・AIC、BIC）
- 個体及びグループの検定及び統計（t, F、カイ二乗(chi-square), Wald, Hausman その他）

2.14.2　原則ベース・モデルに関する詳細情報

経験則ガイドラインに加える原則はない。

2.15　クラス 13：相関係数

2.15.1　経験則

相関は変数間の線形関係を測定したものである。チェッカーは、公表されたアウトプットが経験則の要素 1 を満たしている、すなわち、各相関係数の基礎となる非加重値が少なくとも 10 あることを保証しなければならない。

2.15.2　原則ベース・モデルに関する詳細情報

問題が生じ得るケースはごく少数であるが存在する（例・相関マトリックスが 0 又は 1 を含む)。こうした場合であっても、問題は要約統計量に関係する相関係数の公表に関わるものである。

2 値変数の相関は、2 値の説明変数による線形回帰又は完全飽和モデルの線形回帰として扱われる。したがって、線形回帰係数に関して同一の項目が考慮に入れられている（第 2.11 項参照)。しかしながら、大半の分析において、相関係数は安全なものとして分類されている。