資 料

匿名データのチェックリスト（案）

本資料は、一橋大学で行っている試行的提供における匿名化の考え方及び具体的な方法を整理し、それに基づき、海外におけるチェックリストを参考にして、匿名化の考え方、基準及びチェックリストの検討に資するためにまとめたものである。

Ⅰ　匿名化の考え方

1　匿名化とは調査単位とミクロデータの対応関係を特定されないようにすることであり、対応関係を特定される危険性の最も高いのは地理的情報である。

1-1 地理的情報は、詳細になればなるほど調査単位が特定される危険性が高くなるので、一定の範囲より小さい範囲の情報を提供しないようにする必要がある。

1-2 提供する地理的情報の匿名化の詳細さに応じて、構造分析用、地域分析用といった何種類かの匿名データを作成することも考えられる。

2　調査単位とミクロデータとの対応関係は、識別情報によって判断される。識別情報はミクロデータに含まれており、かつ外部情報からも知ることができる情報である。

2-1 識別情報のうち、調査単位とミクロデータとの対応関係が特定可能な直接的な情報（氏名、住所等のアイデンティファイア）は匿名データから除く。

2-2 性別、年齢等の間接的な識別情報については、 トップコーディング、グルーピングまたは削除などの技法によって匿名化を施し、調査単位が特定されないようにしなければならない。

2-3 既存の名簿などの外部情報により調査単位が特定される危険性がある場合、それを回避するための措置をとらなければならない。

3　匿名化の技法としては、識別情報の削除、トップコーディング、グルーピング、リサンプリング、ミクロデータの並べ替え、スワッピング、誤差の付加などがある。

3-1 これらの方法で十分でないと判断される場合には、秘密の情報自体を削除またはグルーピングするなどの処理を施すことになる。

4　データの安全性を重視すると、有用性を損なうおそれがあり、有用性を重視すると安全性に疑問が生じるおそれがある。匿名データの作成に当たっては、この両方を考慮しつつ、匿名化の方法を定めなければならない。調査によって安全性、有用性は異なるので、匿名化の方法は、調査ごとに決める必要がある。

5　ミクロデータから秘密の情報が漏洩する危険性を排除するためには、利用者に利用の条件を遵守させる必要があり、匿名化の方法はそれを前提条件として定めるものである。

（参考）匿名データの利用条件

１：特定の個人もしくは世帯を特定しようとするような試みを行わないこと。
２：匿名データは、統計作成もしくは学術研究のためにのみ用いること。
３：匿名データを第三者に提供又は利用させないこと。
４：許可なく、他のデータベース等とのマッチングを行わないこと。
５：その他、提供者が提示する利用の条件を遵守すること。

Ⅱ　匿名化の基準

１　地理的情報について

(1) 地理的情報としては、地域内に最小でも人口 50 万人以上いなければならない。
(2) 直接的な地理的情報以外で、地理的情報が明らかになる項目（例えば、サンプリング情報など）についても、上記(1)の最小人口 50 万人の基準に適合させなければならない。
(3) 地域分析用として、人口 50 万人未満の地理的情報を提供するような匿名データを作成する場合には、他の識別情報などの匿名化の程度を高めなければならない。
(4) 入手可能な外部情報により、ある特定の種類の施設であることが明らかになるようなことがないようにしなければならない。

２　個人・世帯の識別情報について

(1) 氏名、住所など個人又は世帯を直接的に識別できる情報は削除されなければならない。
(2) 間接的に個人又は世帯を識別できる情報、例えば年齢、世帯人員、居住室数などの情報については、年齢の高い個人、世帯員数が多い世帯、居住室数の多い住宅など特定される可能性が高い場合、トップコーディング、グルーピングまたは削除を施す必要がある。トップコーディングにおいては、母集団（個人又は世帯）全体の 0.5%を目安にすることが望ましい。
(3) 少数の特定の集団を対象とする場合、トップコーディングの基準を３～５%にすることを考慮すべである。
(4) トップコーディングするデータ項目については、その情報（平均値や中央値など）を明らかにすることが望ましい。
(5) 世帯単位のデータを提供する場合、調査単位が特定されることがないよう、必要があれば、匿名化を考慮する必要がある。

３　誤差（ノイズ）

(1) ミクロデータに誤差を加えることによって、調査データと外部情報との対応関係を特定する可能性を低めることができる。他に適当な匿名化の技法がない場合には、研究・分析上の有用性を損なわない範囲で誤差を付加することを考慮すべきである。
(2) 誤差を加える方法としては、①乱数による誤差の付加（random noise）、②調査単位間の調査情報の交換（swapping）、③ブランク（blank）への置換え又は補定（imputation）がある。

４　リサンプリング

ミクロデータを全て提供する場合は、その一部を提供する場合に比べて、調査単位の特定の可能が高くなる。例えば、ある人が調査を受けたことがわかっている場合には、ミクロデータの中に必ずその人のデータがあるはずとの前提で探すことができる。したがって、必要に応じて、ミクロデータの全てではなく、一部のデータだけを提供することを考慮すべきである。

５　外部ファイルとのマッチングの可能性

(1) ミクロデータと外部の既存ファイルのデータを突き合わせることにより調査単位が識別されるような可能性があれば、それを回避するための措置をとらなければならない。

(2) 調査のための標本フレームが、国勢調査の母集団情報以外の情報によって提供されている場合には、調査データと標本フレームの元の情報とを一致させることが可能となるおそれがあるので、事前に回避する措置をとらなければならない。

６　その他の問題

(1) データの一連番号、データの並び順によって、およその地域範囲が推測されるおそれがあるので、削除、付替え又は並べ替えをするべきである。

(2) サンプリングに関する情報によっては、地理的情報以外に特定の地域や集団であることが明らかになるおそれがあるので、そのような情報は削除すべきである。

(3) 秘密の情報のうち秘匿の必要性の高い調査項目については、その調査項目自体についてグルーピング、削除等の匿名化を施す必要がある。

(4) 時間の経過とともに、調査データを外部情報と照合することは困難になる。提供時期は調査時点から最低限２年間以上は離すべきである。

Ⅲ　チェックリスト

統計委員会に匿名データについて諮問する際には、匿名データのチェックリストだけではなく、当該統計調査の基本的情報（調査概要、調査票の様式、標本抽出法、提供データの形式等）、匿名データに関する情報（匿名データの作成方針、匿名データのレイアウトフォーム・符号表、匿名化に当たっての留意事項、実施した匿名化の概要）を添付することが必要である。

匿名データのチェックリスト（案）

―全国消費実態調査を例として―

匿名データを作成する統計データの名称および年次

全国消費実態調査（平成元年、６年、11 年、16 年）

1　地理的情報

(1) 提供するファイルにはどのレベルの地理的情報が含まれていますか。匿名化のために地理的情報を加工していますか。

全国を６地域に区分した地域ブロック。 全国 47 都道府県を６ブロックに集約しています。 ブロックの構成、人口、世帯数は別添１（省略）を参照。

(2) 直接的な地理的情報以外に地理的情報が明らかになるような情報がありますか。

標本データを母集団に復元するための乗率は、都市階級別にそれぞれ固有の値になっているために、地理的情報と組み合わせると、市区町村レベルまで判明するおそれがあります。そのため、匿名化措置として、乗率を階級別に区分し、階級別のその平均値を乗率としています。 乗率の階級、平均値等は別添２（省略）を参照。

(3) 地域分析用に詳細な地理的情報を提供していますか。

特に地域分析用のファイルは作成していません。

(4) ある特定の種類の施設であることが明らかになることはありますか。

特にそのようなことはありません。

2　世帯の識別情報

(1) 世帯の識別情報として考えられるデータ項目を挙げてください。

世帯符号、世帯人員

(2) それぞれの識別情報について、どのような匿名化措置をとっていますか。

世帯符号について、オリジナルの符号は削除し、新たに世帯単位に一連番号を付与しています。
世帯人員が９人以上の世帯は削除しています。

(3) 匿名化措置を施した場合には、その情報を明示してください。

世帯人員９人以上の世帯は母集団全体の約 0.07%を占めています。
世帯人員分布は別添３（省略）を参照。

(4) 世帯単位のデータを提供することに対応して特別な匿名化措置を施していますか。

特別な匿名化措置は施していません。

３　個人の識別情報

(1) 個人の識別情報として考えられるデータ項目を挙げてください。

性別、年齢

(2) それぞれの識別情報について、どのような匿名化措置をとっていますか。

年齢が 80 歳以上のデータについては、すべて 80 歳としています。

(3) 匿名化措置を施した場合には、その情報を明示してください。

年齢が 80 歳以上の人は、母集団全体の約５%を占めています。
年齢分布は別添４（省略）を参照。

４　誤差（ノイズ）

匿名化措置として、誤差を付加する方法を採っていますか。誤差を付加する方法を採っている場合には、その方法を記載してください。

誤差を加える方法は採用していません。

5　リサンプリング

匿名化措置として、リサンプリングをしていますか。リサンプリングをしている場合には、その抽出方法と抽出率を記載してください。

リサンプリングを行っています。
抽出方法は乗率階級別に標本数を比例配分し、乗率階級内は乗率を考慮した確率比例抽出法を採用しています。抽出率は 80%です。

6　外部ファイル

(1) ミクロデータを特定できる可能性のある外部ファイルは存在しますか。

そのような外部ファイルは存在しません。

(2) 母集団情報として利用している情報は何ですか。

母集団情報として利用しているのは国勢調査の調査区情報です。
調査区内の世帯名簿は調査の一環として作成し、その世帯名簿は調査関係者以外見ることはできません。

7　その他

(1) データの一連番号、データの並び順について、何らかの匿名化措置を施していますか。

オリジナルのデータ一連番号は削除しています。
データの並び順は、世帯単位に、乱数によりランダムな並びにしています。
ランダムな並びにしてから、データの一連番号を付与しています。

(2) サンプリング情報によって、地理的情報以外に特定の地域や集団であることが明らかになる可能性はありますか。

そのような情報はありません。

(3) 秘密の情報のうち、特に秘匿する必要性の高い調査項目がありますか。ある場合には、どのような匿名化措置をとっていますか。

秘密の情報のうち、年間収入について秘匿の必要性を検討したが、年間収入から調査単位が特定される可能性は低いとして、匿名化措置は特に施していません。
また、年間収入は回帰分析などで説明変数としてよく利用され、ジニ係数の計算のため

にも実数でないと困ることから、利用の面も考慮してそのまま提供しています。

(4) 提供時期と調査時点とはどの程度の期間が開いていますか。

調査による結果がすべて公表されてから、匿名データを提供しています。したがって、最短の期間でも調査時点から２年以上は開いています。

(5) そのほか、データを匿名化するに当たり、措置していることがありますか。

特にありません。

匿名データのチェックリスト（案）

―就業構造基本調査を例として―

匿名データを作成する統計データの名称および年次

就業構造基本調査（平成４年、９年、14 年）

1　地理的情報

(1) 提供するファイルにはどのレベルの地理的情報が含まれていますか。匿名化のために地理的情報を加工していますか。

全国を６地域に区分した地域ブロック。 全国 47 都道府県を６ブロックに集約しています。 ブロックの構成、人口、世帯数は別添１（省略）を参照。

(2) 直接的な地理的情報以外に地理的情報が明らかになるような情報がありますか。

特にありません。

(3) 地域分析用に詳細な地理的情報を提供していますか。

特に地域分析用のファイルは作成していません。

(4) ある特定の種類の施設であることが明らかになることはありますか。

特にそのようなことはありません。

2　世帯の識別情報

(1) 世帯の識別情報として考えられるデータ項目を挙げてください。

世帯符号、世帯人員

(2) それぞれの識別情報について、どのような匿名化措置をとっていますか。

世帯符号について、オリジナルの符号は削除し、新たに世帯単位に一連番号を付与しています。 世帯人員が９人以上の世帯は削除しています。

(3) 匿名化措置を施した場合には、その情報を明示してください。

世帯人員９人以上の世帯は母集団全体の約 0.07%を占めています。
世帯人員の分布は別添２（省略）を参照。

(4) 世帯単位のデータを提供することに対応して特別な匿名化措置を施していますか。

特別な匿名化措置は施していません。

３　個人の識別情報

(1) 個人の識別情報として考えられるデータ項目を挙げてください。

性別、年齢

(2) それぞれの識別情報について、どのような匿名化措置をとっていますか。

年齢が 80 歳以上のデータについては、すべて 80 歳としています。

(3) 匿名化措置を施した場合には、その情報を明示してください。

年齢が 80 歳以上の人は、母集団全体の約５%を占めています。
年齢分布は別添３（省略）を参照。

４　誤差（ノイズ）

匿名化措置として、誤差を付加する方法を採っていますか。誤差を付加する方法を採っている場合には、その方法を記載してください。

誤差を加える方法は採用していません。

５　リサンプリング

匿名化措置として、リサンプリングをしていますか。リサンプリングをしている場合には、その抽出方法と抽出率を記載してください。

リサンプリングを行っている。
抽出方法は単純任意抽出法を採用し、抽出率は 80%です。

6　外部ファイル

(1) ミクロデータを特定できる可能性のある外部ファイルは存在しますか。

そのような外部ファイルは存在しません。

(2) 母集団情報として利用している情報は何ですか。

母集団情報として利用しているのは国勢調査の調査区情報です。
調査区内の世帯名簿は調査の一環として作成し、その世帯名簿は調査関係者以外見ることはできません。

7　その他

(1) データの一連番号、データの並び順について、何らかの匿名化措置を施していますか。

オリジナルのデータ一連番号は削除しています。
データの並び順は、世帯単位に、乱数によりランダムな並びにしています。
ランダムな並びにしてから、データの一連番号を付与しています。

(2) サンプリング情報によって、地理的情報以外に特定の地域や集団であることが明らかになる可能性はありますか。

そのような情報はありません。

(3) 秘密の情報のうち、特に秘匿する必要性の高い調査項目がありますか。ある場合には、どのような匿名化措置をとっていますか。

特に秘匿する必要性の高い調査項目はありません。

(4) 提供時期と調査時点とはどの程度の期間が開いていますか。

調査による結果がすべて公表されてから、匿名データを提供しています。したがって、最短の期間でも調査時点から２年以上は開いています。

(5) そのほか、データを匿名化するに当たり、措置していることがありますか。

特にありません。