資料１

諮問第 34 号の答申
国民生活基礎調査に係る匿名データの作成について（案）

本委員会は、厚生労働省が作成を予定している平成 16 年国民生活基礎調査（以下「本調査」という。）に係る匿名データの作成方法の計画について審議した結果、下記の結論を得たので答申する。

記

１　計画の適否

本計画については、これにより作成される匿名データにおいて、本調査の客体の匿名性及び学術研究等における有用性がおおむね確保されるものと認められることから、適当である。

ただし、以下の「２　理由等」で指摘した事項については、修正が必要である。

２　理由等

（１）情報の削除

ア　レコードのリサンプリング及び地域情報の削除

本調査の匿名データの作成に当たっては、複数の調査票情報を組み合わせることにより、世帯票及び健康票から構成される匿名データ（以下「匿名データＡ」という。）並びに世帯票、健康票、所得票及び貯蓄票から構成される匿名データ（以下「匿名データＢ」という。）の２種類の匿名データを作成することとし、それぞれ、地域区分を「全国」のみとするとともに、国勢調査区（又は単位区）及び世帯の二段階で再抽出（以下「リサンプリング」という。）したもの（以下「サブサンプル」という。）を用いる計画である。

このうち、匿名データＡ及び匿名データＢのそれぞれについて、リサンプリングを国勢調査区（又は単位区）及び世帯の二段階で行うことについては、特に本調査が採用している集落抽出法が、集団を単位として抽出し、抽出された集団内の全ての標本を調査する方法であるという特性を踏まえ、匿名データの中に特定の調査区（又は単位区）が含まれるか否か、また特定の世帯が含まれるか否かの判別を困難とする措置であることから、適当である。

また、地域区分を「全国」のみとし、全国一律の拡大乗数を付与することについては、サブサンプル中に含まれる多くの属性情報と詳細な地域情報を組み合わせた場合に調査客体が特定される可能性が高まること、各レコードが保持する拡大乗数から抽出地域が特定されてしまうことを防ぐ必要があること、本調査の匿名データの作成は今回が初回であり、調査客体の匿名性の確保を十分に図るよう慎重を期す必要があることから、やむを得ない措置である。

以上の措置を講じた結果として、本調査のサブサンプルの抽出率については、約２割となっているものの、サブサンプルの大きさは中間年調査の集計客体数と

同程度であること、作成された匿名データＡ及び匿名データＢによる統計と全レコードから作成された公表統計との間で代表的な項目の平均値や分布に大きな乖離はないことから、適当である。

イ　識別情報の削除等

（ア）直接的な識別情報の削除等

本調査のサブサンプル中のレコードに含まれる情報のうち地区番号等の直接識別できる情報は、これを削除するとともに、レコードは乱数により並び替える計画である。

これらについては、調査客体の特定や探索を防止するために効果的な措置であること等から、適当である。

（イ）所得の内訳等の削除

所得票に含まれる所得等の情報については、世帯の総所得、課税等の状況及び掛金のみに限定し、その内訳や世帯員別の情報は削除して提供する計画である。

これについては、所得等の内訳や世帯員別の情報を全て提供すると、これらを合計することにより、（２）のアの（イ）による匿名化措置の効果を担保できなくなること、所得等の詳細な内訳の提供は調査客体が特定される可能性を高めることから、やむを得ない措置である。

ウ　裾切りによるレコード削除

本調査のサブサンプル中のレコードのうち、世帯人員８人以上の世帯、同一年齢の子供が３人以上いる世帯、父子世帯、要介護者が２名以上いる世帯、年齢差の大きい夫婦や親子のいる世帯に係るものは、匿名データから削除する計画である。

世帯人員８人以上の世帯のレコードを削除することについては、世帯員の人数の情報が世帯の外部から比較的容易に把握可能な属性であり、それが極端に大きい場合は出現頻度が低く調査客体が特定される可能性が生じること等から、適当である。

同様に、父子世帯、要介護者が２名以上いる世帯、年齢差の大きい夫婦のいる世帯【案１】、年齢差の大きいまたは小さい親子のいる世帯のレコードを削除することについても、外部から比較的容易に把握可能な属性である一方で出現頻度が低く、調査客体が特定される可能性が生じること等から、適当である。

【案２】なお、年齢差の大きいまたは小さい親子のいる世帯についても、同様に外部から比較的容易に把握可能な属性であり、調査客体が特定される可能性が生じることから、当該レコードを削除する必要がある。

また、同一年齢の子供が３人以上いる世帯については、本計画では、世帯員の年齢は各歳ではなく年齢階級別に提供されるため、同一年齢階級の世帯員数に着目して、当該階級に一定以上の者がいる世帯についてレコードを削除するよう、変更する必要がある。

（２）識別情報の階級区分の統合

ア　トップコーディング及びボトムコーディング

（ア）高齢者の年齢

世帯員の年齢については、一定の値を上限値とし、それを上回る場合に上限値以上でまとめる措置（以下「トップコーディング」という。）を行うこととし、当該上限値は 85 歳以上とする計画である。

これについては、出現頻度が低い一定年齢以上の高齢者をトップコーディングすることにより、性別等の他の属性情報との組み合わせによる調査客体の特定を防ぐことから、適当である。

（イ）総所得及び貯蓄現在高等

世帯の総所得及び貯蓄現在高等については、一定の金額を上限値としてトップコーディングを行う計画である。

これについては、トップコーディングにより、所得等が極端に大きい調査客体の特定を防ぐことから、適当である。

イ　リコーディング（分類区分の再付与）

（ア）世帯員の年齢

世帯員の年齢（トップコーディングを行う高齢者を除く。）については、その分類の程度を粗いものにする措置（以下「リコーディング」という。）を講じることとし、15 歳以上の者は５歳階級別とし、15 歳未満の者は「０～５歳」「６～11 歳」及び「12～14 歳」の３区分とする計画である。

これについては、各歳別のデータ提供に比べて匿名データの有用性が低下するものの、各歳別の年齢が明らかになると、世帯員に関する他の多くの属性情報との組み合わせにより調査客体が特定される可能性が生じることから、やむを得ない措置である。

なお、15 歳未満の者については、健康票の記入対象項目が年齢により異なることから、年齢の特定を防ぐために健康票の回答区分である３区分としたものであり、やむを得ない措置である。

（イ）出現頻度の低い選択肢のある項目

出現頻度の低い選択肢のある項目については、当該選択肢を「その他」等に統合する計画である。

これについては、調査客体の特定を防ぐことから適当な措置であるが、「希望する仕事の形」、「悩みやストレスの原因」、「最も気になる悩みやストレスの原因（主原因）」及び「健診を受けなかった理由」については、専ら本人の意識を問う項目であって外観から識別される可能性が低く、当該情報によって調査客体が特定される可能性が低いと考えられることから、匿名化措置を緩和し、匿名データの有用性の向上を図る必要がある。

ウ　トップコーディング等の基準

トップコーディング、一定の値を下限値としこれを下回る場合に下限値以下でまとめる措置（以下「ボトムコーディング」という。）等を講じる場合、本調査では対象サンプル全体の１％未満を対象とする計画である。

これについては、本調査が集落抽出法で実施され、各世帯及び世帯員に関する多様な項目が把握されていること、有用性の観点から閾値は可能な限り継続した方が望ましく、他年次の匿名データの作成においても当該閾値により匿名性が確保されることを考慮したものであり、やむを得ない措置である。

3　今後の課題

本計画については、本調査に係る匿名データの作成は初回であって、多様な調査項目や抽出方法を考慮した場合、調査客体の匿名性の確保により慎重を期する必要があることから、厳格な匿名化措置を講じていることはやむを得ない。

しかしながら、匿名データの利用者のニーズについては様々なものが考えられることから、以下の課題等について速やかに検討を進め、当該データのより一層の充実に努める必要がある。

（１）地域区分及びリサンプリングの単位

本計画では、匿名性を確保するため、調査客体である世帯の特定につながる可能性が高い地域情報を削除し、地域区分を「全国」のみとする厳格な匿名化措置を講じることとしている。

しかしながら、地域区分については、有用性の観点から極めて重要な情報であることから、調査客体の匿名性の確保を十分に図りつつ、匿名データの利用者のニーズを踏まえて、何らかの地域表章の可能性について検討する必要がある。

また、リサンプリングの単位については、今回、世帯単位のみとしているが、世帯員単位でリサンプリングを行うことで地域情報の付与やリサンプリング率の向上の可能性があること、公衆衛生や疫学分野の研究においては、世帯員単位での健康状態や生活習慣の分析が重要となること等から、利用者のニーズを十分に考慮したうえで、世帯員単位でのリサンプリングによる匿名データの作成の可能性について、速やかに検討を開始する必要がある。

（２）所得票の情報の提供

本計画では、所得票に含まれる情報については、世帯の総所得、課税等の状況及び掛金のみに限定して提供することとしている。

しかしながら、近年、社会保障や所得格差等に関する研究の重要性が増しており、その分析には所得等に関する内訳や世帯員別の情報が重要であること、一方、本計画で適用されていないトップコーディング等以外の匿名化措置の適用も考えられることから、今後、匿名化措置に関する研究等の進展や利用者のニーズを十分に考慮したうえで、所得等の内訳や世帯員別の情報の提供の可能性について検討する必要がある。

（３）匿名データの作成対象年次の拡大

本計画では、匿名データの作成対象調査を調査実施後５年以上経過したものとしており、今回は平成16年に実施したものを作成対象とするとともに、今後、順次拡大することとしている。

しかしながら、研究には経年的な分析が重要であるとともに、近年の経済・社会状況の急激な変化に伴い直近の統計に基づく分析の重要性が増していること、さらに、本調査については３年ごとに大規模調査が実施されていることを踏まえれば、提供時期の短縮について検討する必要がある。

（４）年齢のトップコーディング

本計画では、世帯員の年齢については、85 歳以上でトップコーディングを行うこととしている。

しかし、トップコーディングの上限値については、近年の急速な高齢化の進展及び高齢者に関する分析の重要性等を踏まえ、今後、匿名データの作成対象年次を拡大する際には、当該年次の人口構成に応じて検討する必要がある。

（５）トップコーディング等が行われた変数

本計画により作成された匿名データの各レコード上の変数のうち、トップコーディング及びボトムコーディングが行われている変数については、利用者の利便性向上の観点から、海外における提供事例も踏まえ、当該トップコーディング等を行った変数の平均値等の提供可能性を速やかに検討する必要がある。

（別添 見え消し版）

諮問第 34 号の答申
国民生活基礎調査に係る匿名データの作成について（案）

本委員会は、厚生労働省が作成を予定している平成 16 年国民生活基礎調査（以下「本調査」という。）に係る匿名データの作成方法の計画について審議した結果、下記の結論を得たので答申する。

記

1　計画の適否

本計画については、これにより作成される匿名データにおいて、本調査の調査客体の匿名性及び学術研究等における有用性がおおむね確保されるものと認められることから、適当である。

ただし、以下の「2　理由等」で指摘した事項については、修正が必要である。

コメント [　1]: 重複の削除（厚労省）

2　理由等

（1）情報の削除

ア　レコードのリサンプリング及び地域情報の削除

本調査の匿名データの作成に当たっては、複数の調査票情報を組み合わせることにより、世帯票及び健康票から構成される匿名データ（以下「匿名データＡ」という。）並びに世帯票、健康票、所得票及び貯蓄票から構成される匿名データ（以下「匿名データＢ」という。）の２種類の匿名データを作成することとし、それぞれ、地域区分を「全国」のみとするとともに、国勢調査区（又は単位区）及び世帯の二段階で再抽出（以下「リサンプリング」という。）したもの（以下「サブサンプル」という。）を用いる計画である。

このうち、匿名データＡ及び匿名データＢのそれぞれについて、リサンプリングを国勢調査区（又は単位区）及び世帯の二段階で行うことについては、特に本調査が採用している集落抽出法が、集団を単位として抽出し、抽出された集団内の全ての標本を調査する方法であるという特性を踏まえ、匿名データの中に特定の調査区（又は単位区）が含まれるか否か、また特定の世帯が含まれるか否かの判別を困難とする措置であることから、適当である。

また、地域区分を「全国」のみとし、全国一律の拡大乗数を付与することについては、サブサンプル中に含まれる多くの属性情報と詳細な地域情報を組み合わせた場合に調査客体が特定される可能性が高まること、各レコードが保持する拡大乗数から抽出地域が特定されてしまうことを防ぐ必要があること、本調査の匿名データの作成は今回が初回であり、調査客体の匿名性の確保を十分に図るよう慎重を期す必要があることから、やむを得ない措置である。

以上の措置を講じた結果として、本調査のサブサンプルの抽出率については、約２割となっているものの、サブサンプルの大きさは中間年調査の集計客体数と

同程度であること、作成された匿名データＡ及び匿名データＢによる統計と全レコードから作成された公表統計との間で代表的な項目の平均値や分布に大きな乖離はないことから、適当である。

イ　識別情報の削除等

（ア）直接的な識別情報の削除等

本調査のサブサンプル中のレコードに含まれる情報のうち地区番号等の直接識別できる情報は、これを削除するとともに、レコードは乱数により並び替える計画である。

これらについては、調査客体の特定や探索を防止するために効果的な措置であること等から、適当である。

（イ）所得の内訳等の削除

所得票に含まれる所得等の情報については、世帯の総所得、課税等の状況及び掛金のみに限定し、その内訳や世帯員別の情報は削除して提供する計画である。

これについては、~~所得が極端に大きい場合は調査客体が特定される可能性が生じるが、~~所得等の内訳や世帯員別の情報を全て提供すると、これらを合計することにより、（２）のアの（イ）による匿名化措置の効果を担保できなくなること、所得等の詳細な内訳の提供は調査客体が特定される可能性を高めることから、やむを得ない措置である。

ウ　裾切りによるレコード削除

本調査のサブサンプル中のレコードのうち、世帯人員８人以上の世帯、同一年齢の子供が３人以上いる世帯、父子世帯、要介護者が２名以上いる世帯、年齢差の大きい夫婦や親子のいる世帯に係るものは、匿名データから削除する計画である。

~~これ~~世帯人員８人以上の世帯のレコードを削除することについては、世帯員の人数等の情報が世帯の外部から比較的容易に把握可能な属性であり、それが極端に大きい場合は出現頻度が低く調査客体が特定される可能性が生じること等から、適当である。

同様に、父子世帯、要介護者が２名以上いる世帯、年齢差の大きい夫婦のいる世帯【案１】、年齢差の大きいまたは小さい親子のいる世帯のレコードを削除することについても、外部から比較的容易に把握可能な属性である一方で出現頻度が低く、調査客体が特定される可能性が生じること等から、~~レコードを削除することが~~適当である。

【案２】なお、年齢差の大きいまたは小さい親子のいる世帯についても、同様に外部から比較的容易に把握可能な属性であり、調査客体が特定される可能性が生じることから、当該レコードを削除する必要がある。

【新規追加】また、同一年齢の子供が３人以上いる世帯については、本計画では、世帯員の年齢は各歳ではなく年齢階級別に提供されるため、同一年齢階級の

コメント [2]: 分かりにくいため、削除（事務局）

コメント [3]: コメント８の挿入に伴う修正（事務局）

コメント [4]: コメント８の挿入に伴う削除（事務局）

コメント [5]: 用語の統一（追加）

コメント [6]: 修正に関して部会において説明（厚労省）

コメント [7]: 表現の統一

コメント [8]: 用語の統一

コメント [9]: 表現の統一

コメント [10]: 案１を採用し、案２の方は削除してもよいのではないか（委員）

世帯員数に着目して、当該階級に一定以上の者がいる世帯についてレコードを削除するよう、変更する必要がある。

コメント [11]: 修正に関して部会において説明（厚労省）

（２）識別情報の階級区分の統合

ア　トップコーディング及びボトムコーディング

（ア）高齢者の年齢

世帯員の年齢については、一定の値を上限値とし、それを上回る場合に上限値以上でまとめる措置（以下「トップコーディング」という。）を行うこととし、当該上限値は 85 歳以上とする計画である。

これについては、~~トップコーディングにより、極めて高齢であるという特殊な属性はまとめられ~~出現頻度が低い一定年齢以上の高齢者をトップコーディングすることにより、性別等の他の属性情報との組み合わせによる調査客体の特定を防ぐことから、適当である。

コメント [12]: 分かりにくいため修正（事務局）

（イ）総所得及び貯蓄現在高等

世帯の総所得及び貯蓄現在高等については、一定の金額を上限値としてトップコーディングを行う計画である。

これについては、トップコーディングにより、所得等が極端に大きい~~という特殊な属性をまとめられ、~~調査客体の特定を防ぐことから、適当である。

コメント [13]: 分かりにくいため、削除（事務局）

イ　リコーディング（分類区分の再付与）

（ア）世帯員の年齢

世帯員の年齢（トップコーディングを行う高齢者を除く。）については、その分類の程度を粗いものにする措置（以下「リコーディング」という。）を講じることとし、15 歳以上の者は５歳階級別とし、15 歳未満の者は「０～５歳」「６～11 歳」及び「12～14 歳」の３区分とする計画である。

これについては、各歳別のデータ提供に比べて匿名データの有用性が低下するものの、各歳別の年齢が明らかになると、世帯員に関する他の多くの属性情報との組み合わせにより調査客体が特定される可能性が生じることから、やむを得ない措置である。

なお、15 歳未満の者については、健康票の記入対象項目が年齢により異なることから、年齢の特定を防ぐために健康票の回答区分である３区分としたものであり、やむを得ない措置である。

（イ）~~外観から識別される可能性の低い項目~~出現頻度の低い選択肢のある項目

出現頻度の低い選択肢のある項目については、当該選択肢を「その他」等に統合する計画である。

これについては、調査客体の特定を防ぐことから適当な措置であるが、「希望する仕事の形」、「悩みやストレスの原因」、「最も気になる悩みやストレスの原因（主原因）」及び「健診を受けなかった理由」については、専ら本人

コメント [14]: 項の立て方の間違いを修正（事務局）

コメント [15]: 厚労省の確認により提供可能なため追加（厚労省）

の意識を問う項目であって外観から識別される可能性が低く、当該情報によって調査客体が特定される可能性が低いと考えられることから、匿名化措置を緩和し、匿名データの有用性の向上を図る必要がある。

ウ　トップコーディング等の基準

トップコーディング、一定の値を下限値としこれを下回る場合に下限値以下でまとめる措置（以下「ボトムコーディング」という。）等を講じる場合、本調査では対象サンプル全体の１％未満を対象とする計画である。

これについては、本調査が集落抽出法で実施され、各世帯及び世帯員に関する多様な項目が把握されていること、有用性の観点から閾値は可能な限り継続した方が望ましく、他年次の匿名データの作成においても当該閾値により匿名性が確保されることを考慮したものであり、やむを得ない措置である。

3　今後の課題

本計画については、本調査に係る匿名データの作成は初回であって、多様な調査項目や抽出方法を考慮した場合、調査客体の匿名性の確保により慎重を期する必要があることから、厳格な匿名化措置を講じていることはやむを得ない。

しかしながら、匿名データの利用者のニーズについては様々なものが考えられることから、以下の課題等について速やかに検討を進め、当該データのより一層の充実に努める必要がある。

（１）地域区分及びリサンプリングの単位

本計画では、匿名性を確保するため、調査客体である世帯の特定につながる可能性が高い地域情報を削除し、地域区分を「全国」のみとする厳格な匿名化措置を講じることとしている。

しかしながら、地域区分については、有用性の観点から極めて重要な情報であることから、調査客体の匿名性の確保を十分に図りつつ、匿名データの利用者のニーズを踏まえて、何らかの地域表章の可能性について検討する必要がある。

また、リサンプリングの単位については、今回、世帯単位のみとしているが、世帯員単位でリサンプリングを行うことで地域情報の付与やリサンプリング率の向上の可能性があること、公衆衛生や疫学分野の研究においては、世帯員単位での健康状態や生活習慣の分析が重要となること等から、利用者のニーズを十分に考慮したうえで、世帯員単位でのリサンプリングによる匿名データの作成の可能性について、速やかに検討を開始する必要がある。

（２）所得票の情報の提供

本計画では、所得票に含まれる情報については、世帯の総所得、課税等の状況及び掛金のみに限定して提供することとしている。

しかしながら、近年、社会保障や所得格差等に関する研究の重要性が増しており、その分析には所得等に関する内訳や世帯員別の情報が重要であること、一方、本計画

コメント [　16]: 外観識別不可能にもかかわらず、緩和措置が行われなくなるため削除（委員）

コメント [　17]: 「本人の意識にも関わる項目でもあって」に修正（委員）

コメント [　18]: 明確にする観点から追加。（厚労省）

コメント [　19]: 委員の意見を踏まえ、広く利用者（潜在的なユーザーも含む。）のニーズを踏まえるよう、強調（部会長）

で適用されていないトップコーディング等以外の匿名化措置の適用も考えられることから、今後、匿名化措置に関する研究等の進展や利用者のニーズを十分に考慮したうえで、所得等の内訳や世帯員別の情報の提供の可能性について検討する必要がある。

（３）匿名データの作成対象年次の~~拡張~~拡大

本計画では、匿名データの作成対象調査を調査実施後５年以上経過したものとしており、今回は平成16年に実施したものを作成対象とするとともに、今後、順次拡大することとしている。

しかしながら、研究には経年的な分析が重要であるとともに、近年の経済・社会状況の急激な変化に伴い直近の統計に基づく分析の重要性が増していること、さらに、本調査については３年ごとに大規模調査が実施されていることを踏まえれば、提供時期の短縮について検討する必要がある。

（４）年齢のトップコーディング

本計画では、世帯員の年齢については、85 歳以上でトップコーディングを行うこととしている。

しかし、トップコーディングの上限値については、近年の急速な高齢化の進展及び高齢者に関する分析の重要性等を踏まえ、今後、匿名データの作成対象年次を拡~~張~~大する際には、当該年次の人口構成に応じて検討する必要がある。

~~（５）匿名データＡ及び匿名データＢの閾値~~

~~トップコーディングの上限値及びボトムコーディングの下限値については、匿名データＡ及び匿名データＢで同一の値とし、他の年次に対しても可能な限り継続することとしている。~~

~~これについては、他の年次の匿名データを作成する場合には、それぞれの調査票における対象サンプル全体の分布を確認し、当該分布が著しく異なる場合には、有用性及び匿名性の観点から、匿名データＡ及び匿名データＢでトップコーディング等の閾値を変えることも検討する必要がある。~~

（５）トップコーディング等が行われた変数

本計画により作成された匿名データの各レコード上の変数のうち、トップコーディング及びボトムコーディングが行われている変数については、利用者の利便性向上の観点から、海外における提供事例も踏まえ、当該トップコーディング等を行った変数の平均値等の提供可能性を速やかに検討する必要がある。

コメント [20]: 明確にする観点から追加（厚労省）

コメント [21]: 委員の意見を踏まえ、広く利用者（潜在的なユーザーも含む。）のニーズを踏まえるよう、強調（部会長）

コメント [22]: 用語の統一

コメント [23]: 事実に沿って修正（厚労省）

コメント [24]: 分かりやすく修正（厚労省）

コメント [25]: 用語の統一

コメント [26]: 他の課題に比べやや小さい課題なので削除。（委員）

コメント [27]: 事務局からの提案

コメント [28]: 具体的な課題を明確に記載すべき（委員）