総務省統計局
独立行政法人統計センター

一般用ミクロデータ利用の手引

１．はじめに

一般用ミクロデータをご利用いただき、ありがとうございます。

総務省統計局及び独立行政法人統計センターは、統計演習など教育用に利用可能な一般用ミクロデータを共同で研究、作成し、統計センターのホームページにおいて提供しています。

一般用ミクロデータは、公表済みの結果表から作成したデータセットです。したがって個別情報の秘匿を気にすることなく自由にご利用いただけるデータとなっていますが、一般用ミクロデータから導かれた分析結果は実証研究の結果と見なすことはできません。

実証研究として分析を行う場合には、一般用ミクロデータではなく、匿名データ又はオーダーメード集計のご利用などをご検討ください。

統計センター 公的統計の二次的利用サービス：
http://www.nstac.go.jp/use/archives/ippan-microdata/

２．利用者アンケートについて

今後、一般用ミクロデータの利便性向上と利活用増進について検討を進めていくためには、利用者の皆様からのご意見が必要となりますので、ご意見・ご要望等ございましたら、アンケートにてお知らせください。

３．これから利用される方のために

一般用ミクロデータをご利用になって作成された教材や開発した分析用プログラムなど、支障のない範囲で公開していただきますと、これから一般用ミクロデータの利用を考えている方々への有益な情報となりますのでご協力をお願いします。

ご協力いただける場合は、「利用に関する質問」のフォームからご連絡いただきますようお願いします。

統計センターから改めてご連絡させていただき、お寄せいただいた資料を活用事例としてホームページなどで公開させていただきます。

４．全国消費実態調査に基づく一般用ミクロデータ

全国消費実態調査に基づく一般用ミクロデータは、既存結果表とのバランスを考慮した統計量（度数及び数量（平均、標準偏差等））に基づく集計表を作成、公表したのち、 その統計量に基づき乱数を発生させて作成した、ミクロデータ形式の擬似データです。利用に当たっては、その基となる統計調査の報告書などを参照し、調査方法、用語の定義、収支項目分類等についてご理解のうえお使いください。

(1) 基となる統計調査の名称

平成 21 年全国消費実態調査

平成 21 年全国消費実態調査に関する情報は調査実施機関である総務省統計局のホームページからご覧いただけます。

平成 21 年全国消費実態調査（総務省統計局）

http://www.stat.go.jp/data/zensho/2009/index.htm

(2) ファイルのフォーマット

一般用ミクロデータは、CSV 形式のファイルを用意しています。

ファイル名等は表１のとおりです。

表１ 各ファイルのファイル名及びファイルサイズ

データ名	データ形式	ファイル名	サイズ
全世帯 （十大費目）	CSV	ippan_2009zensho_z_dataset.csv	約 7MB
勤労者世帯 （十大費目）	CSV	ippan_2009zensho_k_dataset.csv	約 4MB
全世帯 （詳細品目）	CSV	ippan_2009zensho_s_dataset.csv	約 75MB

(3) ファイルの特徴

各ファイルとも、１レコードが１世帯を表します。各ファイルの特徴は表２のとおりです。

なお、各ファイルの先頭に注意事項及び項目名が収録されていますので、利用する前にご確認ください。

表２ 各ファイルの特徴

<table>
<tr><th rowspan="2">データ名</th><th rowspan="2">レコード数</th><th colspan="2">収録項目</th></tr>
<tr><th>世帯属性等</th><th>収支項目</th></tr>
<tr><td>全世帯
（十大費目）</td><td>45,811</td><td>７項目

世帯主の年齢、住居の所有関係など</td><td rowspan="2">12 項目

年間収入、消費支出及び十大費目</td></tr>
<tr><td>勤労者世帯
（十大費目）</td><td>26,239</td><td>４項目

世帯主の年齢、産業、職業、企業規模</td></tr>
<tr><td>全世帯
（詳細品目）</td><td>45,811</td><td>７項目

世帯主の年齢、住居の所有関係など</td><td>12 項目

年間収入、消費支出、十大費目及び 410 品目分類</td></tr>
</table>

※　収録項目の詳細については、「符号表」をご確認ください。

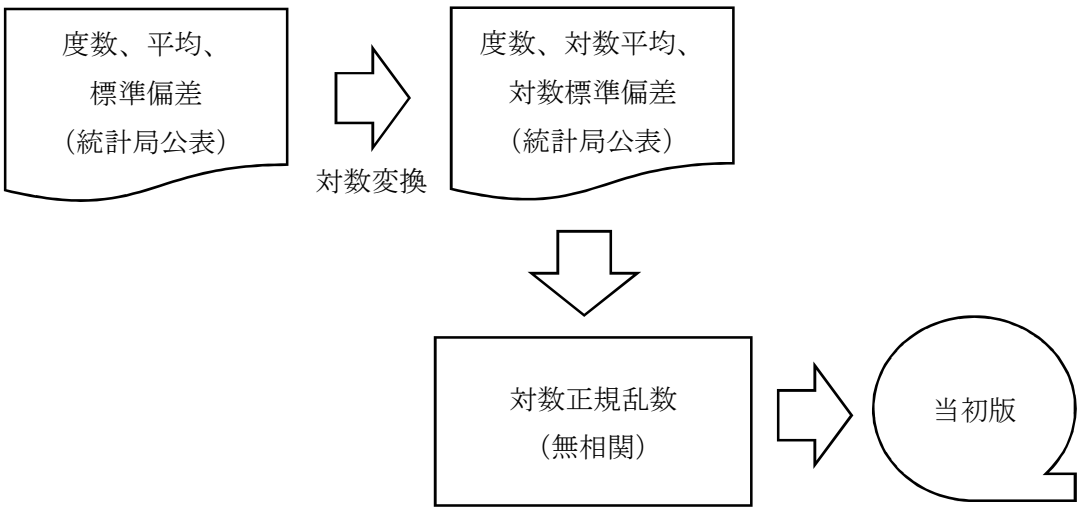
なお、共通の収録項目として、集計用乗率が各レコードに付与されています。

(4) 簡易データ当初版（十大費目）の作成手法

統計局公表の度数、平均、標準偏差を対数変換して対数正規乱数により作成

① クロス世帯属性別の 12 項目（年間収入、消費支出、十大費目）について度数、平均、標準偏差を集計（統計局公表）

② 12 項目（年間収入、消費支出、十大費目）について、クロス世帯属性別に平均及び標準偏差の対数変換を行い、度数、対数平均、対数標準偏差により生成した対数正規乱数により当初版を作成



対数変換の式

（n：標本数、μ：平均、σ：標準偏差）

対数標準偏差　$\sigma' = \sqrt{log((\sigma/\mu)^2 + 1)}$

対数平均　　$\mu' = log(\mu) - \sigma'^2 / 2$

Rによる対数正規乱数（rlnorm関数）

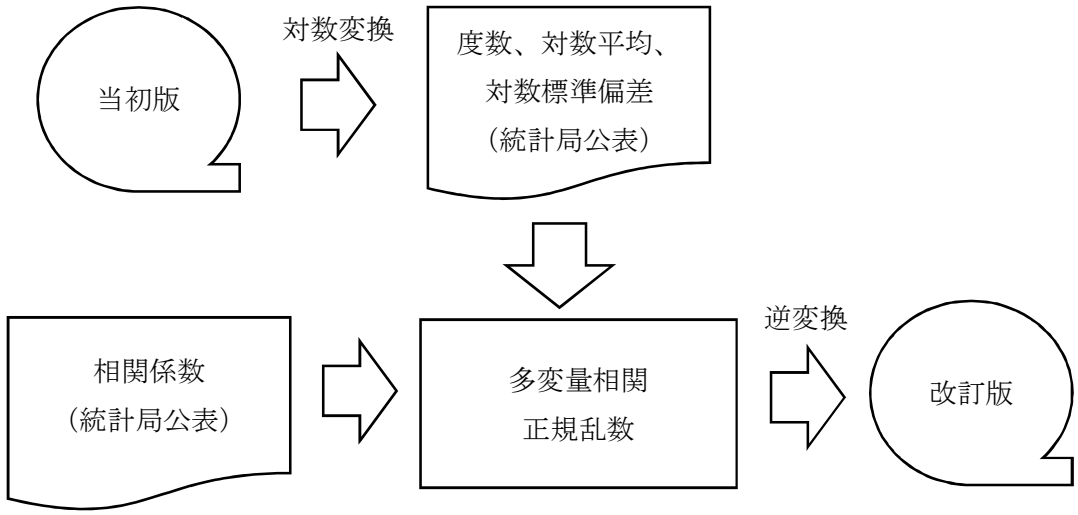
```
Sdlog <- sqrt(log((sd/mean)^2 + 1))
meanlog <- log(mean) – (sdlog^2) / 2
z <- rlnorm(n, meanlog, sdlog)
```

(5) 簡易データ改訂版（十大費目）の作成手法

対数正規乱数により作成した当初版に相関係数を反映して作成

① 12 項目（年間収入、消費支出、十大費目）について項目間の相関係数を集計（統計局公表）

② 12 項目（年間収入、消費支出、十大費目）について、クロス世帯属性別に平均及び標準偏差の対数変換を行い、度数、対数平均、対数標準偏差、相関係数により生成した多変量相関正規乱数を逆変換することにより改訂版を作成



対数変換（１を加算）

x’＝log(ｘ＋１)　ｘ：

当初版の値を代入

相関正規乱数（mvrnorm 関 数 ）

ｚ＝mvrnorm(ｎ，mu，Sigma，empirical=TRUE)

ｎ：変数の個数、mu：０値、Sigma：相関係数

逆対数変換（１を減算）

$z = e^{(z * \sigma' + \mu')} - 1$

μ'：対数平均、 σ'：対数標準偏差

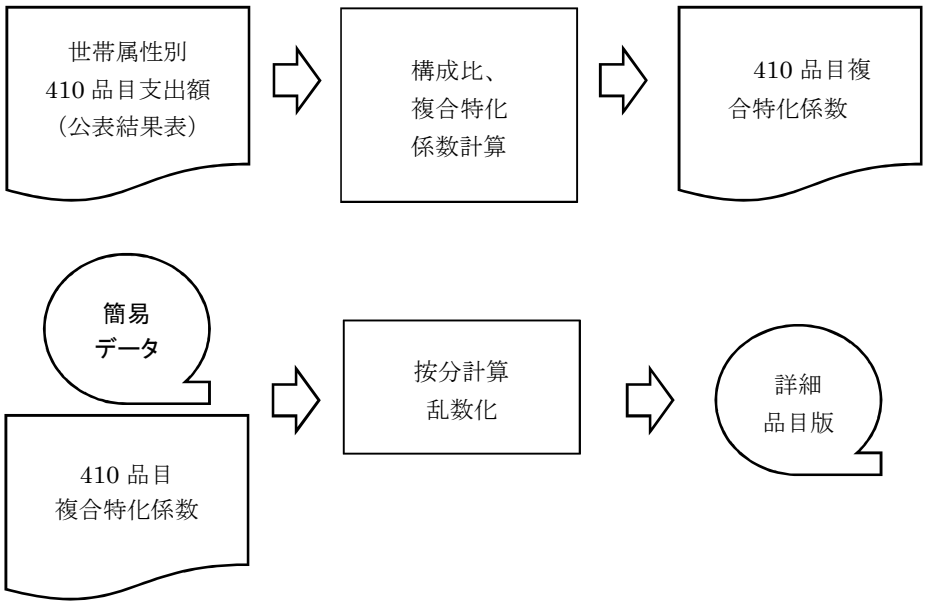
Ｒによる多変量相関正規乱数（mvrnorm 関数）

```
dfRan <- log1p(dfRan)
meanR <- colMeans(dfRan)
sdR <- apply(dfRan,2,sd)
z <- mvrnorm(n, mu, Sigma, empirical=TRUE)
z <- expm1(t(t(zz) * sdR + meanR))
```

(6) 詳細品目版の作成手法

平成 29 年６月に公開した一般用ミクロデータの詳細品目版については、相関を反映した改訂版を基に、世帯属性別の特化係数に着目することで 12 項目間（年間収入、消費支出、10 大費目）の相関を維持しつつ、410 品目分類間の相関について再現を図っています。具体的な作成手法としては、世帯属性別の特化係数を反映した 410 品目分類別の構成比を按分比率として、10 大費目に乗算して得られた按分値をさらに乱数化する方法を用いています。

① 公表結果表（品目編）から 410 品目分類の構成比を計算
② 世帯属性による複合特化係数を計算
③ 簡易データ（改訂版）の 10 大費目ごとに 410 品目分類の按分値を計算
④ 410 品目分類の按分値を乱数化



＜参考＞
一般用ミクロデータの試作データ(詳細版)作成手法
～平成 21 年全国消費実態調査に基づく擬似データ～
http://www.nstac.go.jp/services/society_paper/28_06_01.pdf

５．就業構造基本調査に基づく一般用ミクロデータ

就業構造基本調査に基づく一般用ミクロデータは、公表結果表に基づいて系統抽出と同様の方法により作成したミクロデータ形式の擬似標本データです。利用に当たっては、その基となる統計調査の報告書などを参照し、調査方法、用語の定義、調査項目等についてご理解のうえお使いください。

(1) 基となる統計調査の名称

令和 4 年就業構造基本調査
平成 29 年就業構造基本調査
平成 24 年就業構造基本調査
平成 19 年就業構造基本調査
平成 14 年就業構造基本調査
平成９年就業構造基本調査
平成４年就業構造基本調査

就業構造基本調査に関する情報は調査実施機関である総務省統計局のホームページからご覧いただけます。

令和4 年就業構造基本調査（総務省統計局）
http://www.stat.go.jp/data/shugyou/2022/index.html

(2) ファイルのフォーマット

一般用ミクロデータは、CSV 形式のファイルを用意しています。
ファイル名等は表３のとおりです。

表３ 各ファイルのファイル名及びファイルサイズ

調査年	データ形式	ファイル名	サイズ	レコード数
令和４年	CSV	ippan_2022shugyou_dataset.csv	約 28MB	220,391
平成 29 年	CSV	ippan_2017shugyou_dataset.csv	約 29MB	221,953
平成 24 年	CSV	ippan_2012shugyou_dataset.csv	約 29MB	221,630
平成 19 年	CSV	ippan_2007shugyou_dataset.csv	約 28MB	220,603
平成 14 年	CSV	ippan_2002shugyou_dataset.csv	約 25MB	218,349
平成９年	CSV	ippan_1997shugyou_dataset.csv	約 23MB	213,306
平成４年	CSV	ippan_1992shugyou_dataset.csv	約 23MB	205,876

(3) ファイルの特徴

各ファイルとも、１レコードが１人を表します。各ファイルの収録項目は表４のとおりです。

なお、各ファイルの先頭に注意事項及び項目名が収録されていますので、利用する前にご確認ください。

表４ 各調査年の収録項目

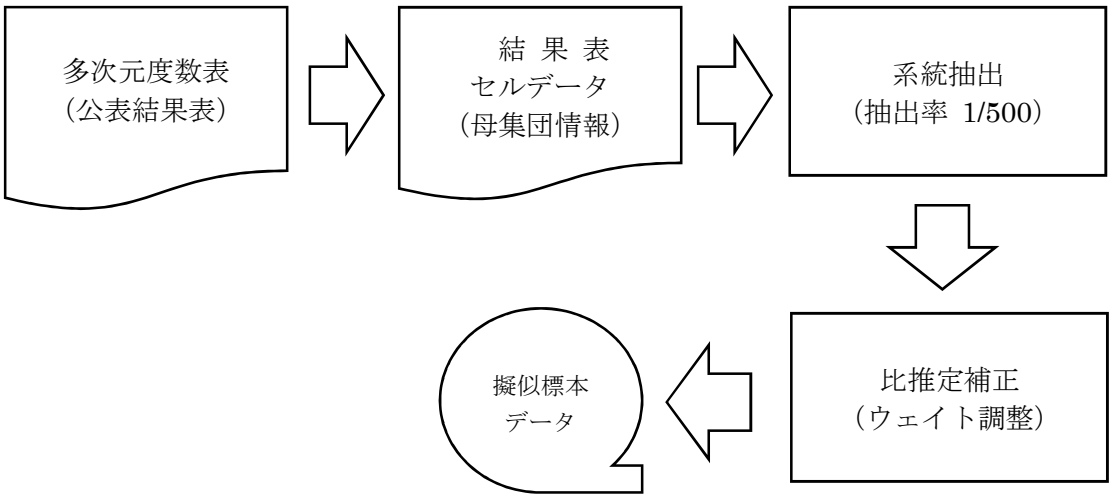
	収録項目（項目名）	調査年						
		2022	2017	2012	2007	2002	1997	1992
共通項目	データ名称(DataSource)	○	○	○	○	○	○	○
	調査年(Year)	○	○	○	○	○	○	○
	都道府県(Prefecture)	○	○	○	○	○	○	○
	政令指定市(City)	×	○	○	○	○	○	○
	政令指定市・県庁所在市・人口30万以上市(City)	○	×	×	×	×	×	×
	市部(Urban)	×	○	○	○	○	×	×
	性別(Gender)	○	○	○	○	○	○	○
	年齢(Age)	○	○	○	○	○	○	○
	就業状態(WorkStatus)	○	○	○	○	○	○	○
有業者	雇用者(WorkEmploy)	○	○	○	○	○	○	○
	正規就業者(WorkRegular)	○	○	○	○	×	×	×
	産業(WorkIndustry)	○	○	○	○	○	○	○
無業者	就業希望(NoworkWish)	○	○	○	○	○	○	○
	求職(NoworkApply)	○	○	○	○	○	○	○
	非就業者年齢(NoworkAge)	○	○	○	○	○	○	○
	配偶関係(NoworkMarriage)	○	○	○	○	×	○	○
符号ラベル（テキスト） T_Prefecture、T_City、T_Urban、T_Gender、T_Age、T_WorkStatus、 T_WorkEmploy 、 T_WorkRegular 、 T_WorkIndustry 、 T_NoworkWish、 T_NoworkApply、T_NoworkAge、T_NoworkMarriage								
復元ウェイト(Weight)								

注）収録項目については、調査年によって項目がない年次や、基となる公表結果表においてクロス集計されていないため収録されていない年次があります。

(4) 就業構造基本調査版の作成手法

就業構造基本調査に基づく一般用ミクロデータ（PUMESS: Public Use Microdata of Employment Status Survey）は、就業構造基本調査の公表結果表を母集団情報として用い、系統抽出と同様の方法により抽出率 500 分の 1 の擬似標本データを作成しています。このため、PUMESSデータには母集団の分布特性が反映されており、母集団フレームからの系統抽出による擬似的な無作為標本とみなすことで、収録項目の範囲内で人口分析や統計演習等に利用することが可能です。

＜作成フロー＞



PUMESS データには公表結果表をベンチマーク人口（就業状態、都道府県、男女、年齢別の人口）として比推定補正を行った復元ウェイトを収録しており、この復元ウェイトを用いて集計を行うことで集計結果の標本誤差を抑える工夫をしています。利用にあたっては、擬似標本の標本規模が実際の調査の 5 分の 1 程度で標本誤差が大きくなっており、集計結果は公表結果表と完全には一致しないこと、また、丸め誤差を含む結果表から作成しているため、標本誤差に加えて丸め誤差が含まれていることに留意する必要があります。

なお、データ作成の際に、政府統計総合窓口(e-Stat)の API 機能を使用していますが、データ内容は国によって保証されたものではありません。

＜参考＞
一般用ミクロデータ就業構造基本調査版の概要
～系統抽出による擬似標本データ～
http://www.nstac.go.jp/services/society_paper/30_06_03.pdf

6．国勢調査に基づく一般用ミクロデータ

国勢調査に基づく一般用ミクロデータは、公表結果表に基づいて系統抽出と同様の方法により作成したミクロデータ形式の擬似標本データです。利用に当たっては、その基となる統計調査の報告書などを参照し、調査方法、用語の定義、調査項目等についてご理解のうえお使いください。

(1) 基となる統計調査の名称

令和 2年国勢調査
平成 27 年国勢調査
平成 22 年国勢調査
平成 17 年国勢調査
平成 12 年国勢調査

国勢調査に関する情報は調査実施機関である総務省統計局のホームページからご覧いただけます。

令和 2 年国勢調査（総務省統計局）
https://www.stat.go.jp/data/kokusei/2020/index.html

(2) 収録内容

収録データの内容は、「15歳以上就業者」になります。
（非就業者及び15歳未満の人口は含まれません。）

(3) ファイルのフォーマット

一般用ミクロデータは、CSV 形式のファイルを用意しています。
ファイル名等は表５のとおりです。

表５ 各ファイルのファイル名及びファイルサイズ

調査年	データ形式	ファイル名	サイズ	レコード数
令和２年	CSV	ippan_2020kokusei_dataset.csv	約 20MB	192,245
平成 27 年	CSV	ippan_2015kokusei_dataset.csv	約 20MB	196,303
平成 22 年	CSV	ippan_2010kokusei_dataset.csv	約 20MB	198,692
平成 17 年	CSV	ippan_2005kokusei_dataset.csv	約 18MB	205,101
平成 12 年	CSV	ippan_2000kokusei_dataset.csv	約 19MB	210,108

(4) ファイルの特徴

各ファイルとも、１レコードが１人を表します。各ファイルの収録項目は表６のとおりです。

なお、各ファイルの先頭に注意事項及び項目名が収録されていますので、利用する前にご確認ください。

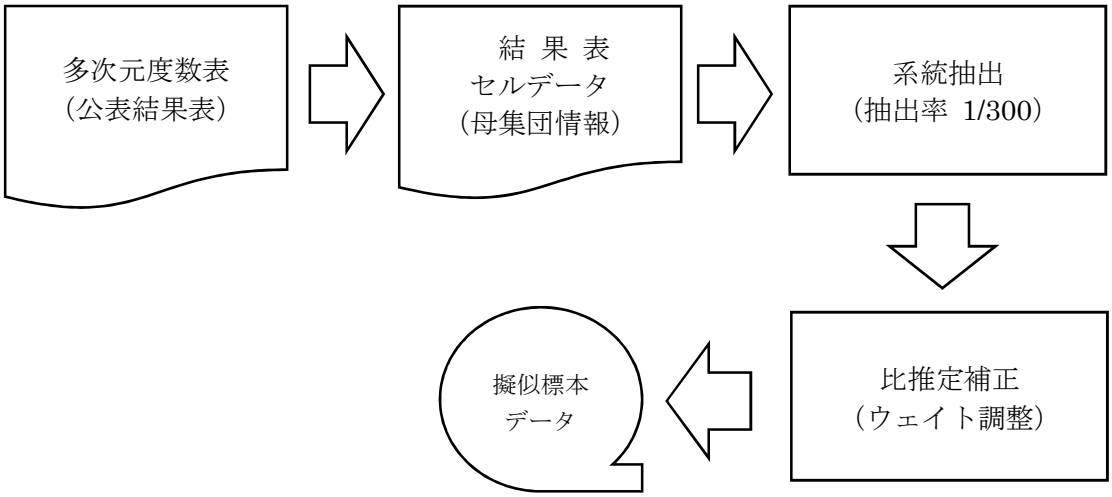
表６ 各調査年の収録項目

収録項目（項目名）	調査年				
	2020	2015	2010	2005	2000
データ名称(DataSource)	○	○	○	○	○
調査年(Year)	○	○	○	○	○
都道府県(Prefecture)	○	○	○	○	○
人口50万以上の市(City)	○	○	○	○	○
性別(Gender)	○	○	○	○	○
年齢(Age)	○	○	○	○	○
配偶関係(Marriage)	○	○	○	○	○
従業上の地位(Employment)	○	○	○	○	○
職業(Occupation)	○	○	○	○	○
符号ラベル（テキスト） T_Prefecture、T_City、T_Gender、T_Age、T_Marriage、T_Employment 、 T_Occupation					
復元ウェイト(Weight)					

(5) 国勢調査版の作成手法

国勢調査に基づく一般用ミクロデータ（PUMPC: Public Use Microdata of Population Census）は、就業構造基本調査版と同様に国勢調査の公表結果表（抽出詳細集計）を母集団情報として用い、系統抽出と同様の方法により抽出率 300 分の１の擬似標本データを作成しています。このため、PUMPCデータには母集団の分布特性が反映されており、母集団フレームからの系統抽出による擬似的な無作為標本とみなすことで、収録項目の範囲内で人口分析や統計演習等に利用することが可能です。

＜作成フロー＞



PUMPC データには公表結果表（抽出詳細集計）をベンチマーク人口（都道府県、男女、年齢別の人口）として比推定補正を行った復元ウェイトを収録しており、この復元ウェイトを用いて集計を行うことで集計結果の標本誤差を抑える工夫をしています。利用にあたっては、擬似標本の標本規模が実際の調査の30分の１程度で標本誤差が大きくなっており、集計結果は公表結果表と完全には一致しないこと、また、丸め誤差を含む結果表から作成しているため、標本誤差に加えて丸め誤差が含まれていることに留意する必要があります。

なお、データ作成の際に、政府統計総合窓口(e-Stat)の API 機能を使用していますが、データ内容は国によって保証されたものではありません。