SSM2015 の子ども情報の代表性[1]

余田翔平
（国立社会保障・人口問題研究所）

【論文要旨】

本稿では、SSM2015 の子ども情報の代表性を検証する。2015 年調査では調査対象者の子どもの諸属性が収集されていることが調査設計上の特徴のひとつである。しかし、この子どもの情報は、ある世代やコーホートを代表するものとしてはバイアスを伴っている可能性がある。その理由は 2 つある。第 1 に、調査対象者の出生コーホートにおいて調査時点までにすでに死亡した成員の子どもの情報は必ず欠測になる。第 2 に、調査年から過去にさかのぼった年次ほど、出生データは若齢出産に偏る。そこで、Human Mortality Database および人口動態統計を用いて、こうした潜在的なバイアスの大きさを評価した。分析結果を踏まえて、本稿の最後には SSM2015 の子ども情報を分析する際の注意点について議論する

キーワード：世代、コーホート死亡率、若齢出産

１．はじめに

過去の SSM 調査と比較したとき、2015 年 SSM 調査の特徴のひとつとして、調査対象者の子どもの情報が豊富に得られていることが挙げられる。SSM2015 では調査対象者の最大 4 人までの子どもについて、性別、出生年、回答者との続き柄、回答者との同別居、学歴を尋ねている。これは、調査対象者に複数の子どもの情報がネストした、典型的なマルチレベル構造のデータでもあり、過去の SSM 調査と比べて、分析の幅はかなり広がるといえる[2]。

しかし、調査対象者の子どもに関する情報を用いるうえで注意すべき点がある。それは、調査対象者としてサンプリングされているのはあくまで「親」の世代であるということである。そして、調査対象者の子どもの情報は、ある世代やコーホートを代表するものになっていない可能性がある[3]。

いま、調査対象者を潜在的な「親世代」と呼ぶことにする。「潜在的な」というのは、一部の調査対象者は子どもを持たないからである。親世代である調査対象者のデータはランダムサンプリングによって得られているため、そのデータは調査時点における母集団の静態を代

[1] 本研究は、JSPS 科研費 JP25000001 の助成を受けたものです。
[2] 例えば、調査対象者の子どもの教育達成を従属変数とした分析を行う場合、それぞれの子どもに共通する家族的要因を統制したうえで、出生順位などの影響を検証できる。
[3] なお、以下で議論する世代ないしコーホートの代表性の問題は、調査対象者とその親の地位とを比較する従来の世代間移動研究に対して指摘されてきた問題点（Duncan 1966; Song and Mare 2015）にも通ずるものである。

表するものになっている。

一方、調査対象者の子どもを「子世代」と呼ぼう。SSM2015 の設計では、子世代の情報はあくまで親世代である調査対象者を通じてしか得ることができない。言い換えれば、子世代は、あるコーホートや世代を母集団として設定し、そこからランダムサンプリングされているわけではない。それゆえ、子世代の情報があるコーホート・世代を代表するものになっているのかについては SSM2015 以外のデータから評価する必要がある。

そこで本稿では、人口動態事象に関する全数調査データを用いて、SSM2015 の子ども情報のカバレッジを検討する。次節では、子世代の情報を歪めうる原因として、(1) コーホート内の死亡と (2) 出生時の（母）親年齢があることについて説明する。3 節ではデータと方法について説明し、4 節で分析結果を示す。最後に 5 節では、SSM2015 の子ども情報を分析に用いるうえでの注意点について議論する。

2．子ども情報の（潜在的）バイアスの原因

子世代の情報を歪める可能性のある要因は少なくとも 2 つある。以下、それぞれについて説明する。

2.1　調査対象者の出生コーホート内の死亡

SSM2015 では調査対象者の年齢は 20〜79 歳となっており、過去の SSM 調査と比較すると上限が 10 歳上がっている。70 歳代の高齢者のデータが豊富に収集できていること自体は望ましいが、その子世代の情報には何らかの偏りが存在する可能性がある。なぜならば、調査対象に含まれる出生コーホートにおいて、調査時点以前に死亡した人々の子どもの情報は必ず欠測データとなるためである。

こうした死亡による潜在的バイアスを避けるためには、調査時点で多くの成員が生存している出生コーホートに限定する必要がある。すなわち、調査対象者の年齢に上限を設け、分析対象を限定したうえでその子世代の情報を使えばよい。

2.2　出生時の（母）親年齢

ところが、ここでもうひとつの問題が浮上してくる。仮に、上記の死亡によるバイアスを回避するために、調査時点で 50 歳未満の調査対象者にサンプルを限定したとしよう。単純化のために、仮に調査が 2015 年 1 月 1 日午前 0 時に一斉に行われたとする。すると、1 年前の 2014 年 1 月 1 日はこのサンプルの年齢の上限は満 48 歳となる。さらに 1 年前の 2013 年における上限は満 47 歳となり、以下年次をさかのぼるごとにサンプルの年齢の上限は 1 歳ずつ若くなっていく（図 1)。

すると、図 1 の白い部分で生じた出生は補足できていないことがわかる。そして、その欠

測の程度は、調査年次から過去にさかのぼるほど深刻になる。例えば、2014 年に生起した出生のうち補足できていないのは「A」で示した部分で生じた出生のみであり、49 歳未満で発生した出生はすべて補足できているため、欠測の影響は小さいと考えられる。ところが、例えば 1982 年出生の子どもデータはすべて出生時の親年齢が 17 歳未満の出生に限られ、若齢出産にかなり限定されてしまう。

このような出生データの欠測の程度は、(1) 調査対象者の調査時年齢の上限と (2) 子どもの出生年、の 2 つの要因の組み合わせで決定される。まず、調査対象者の調査時年齢の上限が若いほど、出生データの欠測は大きくなる（しかし、その上限年齢を高く設定すると、上記で示したコーホート内の死亡による欠測が深刻になってくる)。一方、調査年から過去にさかのぼるほど SSM2015 から補足できない出生データの割合が高くなる。

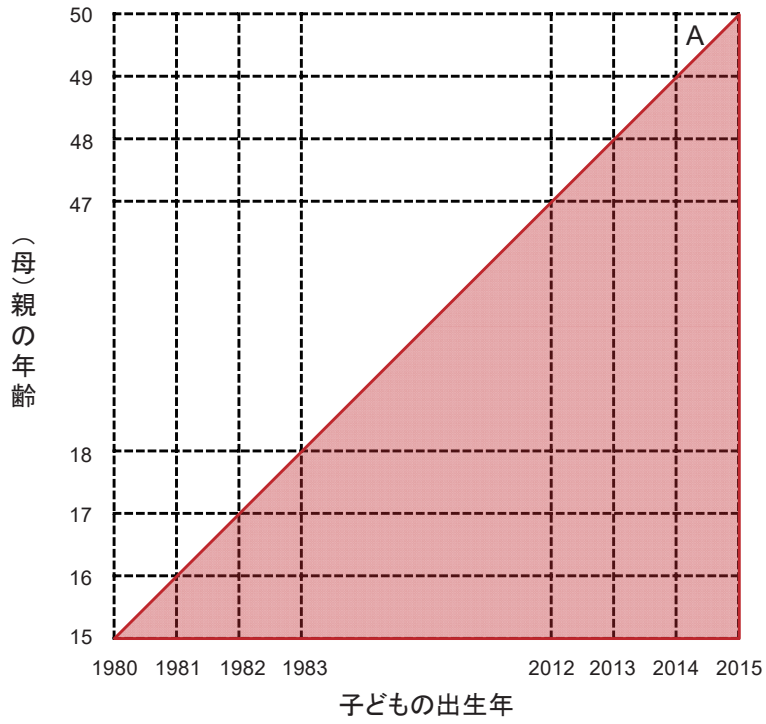


図 1　出生年と出生時親年齢

３．データと方法

前節で示した、子世代のデータの潜在的バイアスを評価するために、本稿では 2 つのデータを使用する。以下、それぞれについて説明する。

3.1　Human Mortality Database

第 1 のデータソースは、国際的な死亡データベースである Human Mortality Database（HMD 2017）である。HMD は人口学の専門家が死亡データの整備・加工法を手順書としてまとめた

上でデータセットを作成し、データベースの形式に整備したものである。このように標準化された方法に従ってデータが加工されているため、各国間や時系列での比較が可能になっている[4]。

本稿では、SSM2015 の調査時における、それぞれの出生コーホートの生存者割合を追っていく[5]。これは生命表（life table）の表記に従えば、x 歳到達時の生存者割合であるl_xに着目することに他ならない。生命表とは、ある出生コーホートが加齢にしたがって死亡していく過程をいくつかの年齢の関数で表したものである。なお、生命表には期間生命表とコーホート生命表があるが、本稿ではいずれもコーホート生命表を想定する。期間生命表とは、ある年次における年齢別死亡率を生涯にわたって経験したとする仮想的なコーホート（仮設コーホート）について作成する生命表である。一方、コーホート生命表は実際にある年次に生まれたコーホートに対して作成する生命表である。

HMD では、x 歳到達時の生存者割合（l_x）はコーホートについては直接公表されていないものの、(1) 各年次の 1 月 1 日時点の各歳別人口と (2) レキシストライアングルごとの死亡数から算出できる。はじめに、レキシストライアングルの概念について簡潔に説明しておく。横軸に時間、縦軸に年齢を取ったレキシス図において、暦年t、満x歳で発生した死亡は図 2 の太枠で示された正方形領域に相当する。この正方形領域は、45 度線を境にして、2 つの異なる出生コーホートによる死亡から構成されている。「L」で示された直角二等辺三角形領域は、$t-x$年生まれコーホートが暦年tに満x歳で経験した死亡を意味する。他方で、「U」で示された直角二等辺三角形領域は、同様に暦年t、満x歳の死亡であるが、$t-x-1$年生まれコーホートによるものである。レキシストライアングルとは、このように正方形領域内の事象を生年によって区別するための概念であり、「U」と「L」にあたる直角二等辺三角形をそれぞれ、「上方トライアングル（upper triangle）」、「下方トライアングル（lower triangle）」と呼ぶ。

つづいて、各コーホートの x 歳到達時の生存者割合（l_x）の算出方法について説明する。$t-x$年コーホートの年齢別死亡確率（q_x）は、「x～$x+1$歳に発生した死亡数」を「x歳到達時の人口」で除せば得られる。前者は図 3 に示した 2 つのレキシストライアングルの和で表される平方四辺形領域で発生した死亡に対応する。他方で後者は、封鎖人口のもとでは、$t+1$年時点の満x歳人口に「$t-x+1$年に$t-x$年コーホートで発生した死亡数」を加えれば得られる。したがって、

---

[4] 比較可能性を阻害しない範囲で日本の死亡状況により適用させるための変更を施した死亡データベースとして「日本版死亡データベース（JMD）」がある。HMD および JMD については石井（2015）に詳しい。

[5] 2015 年 SSM 調査は 3 回に分けて 2015 年 1 月 31 日から 7 月 26 日にかけて実施されているが、ここでは単純化のため、2015 年 1 月 1 日時点の生存者割合を見ていくことにする。そのため、2015 年 1 月 1 日から実査までの死亡は無視する。

$$q_x = \frac{D_L(x,t) + D_U(x,t+1)}{P(x,t+1) + D_L(x,t)}$$

となる。年齢別生存確率（p_x）は、

$$p_x = 1 - q_x$$

により、ここで$l_0 = 1$とすれば、

$$l_x = l_0 \prod_{i=0}^{x-1} p_i = \prod_{i=0}^{x-1} p_i$$

が得られ、これが x 歳到達時の生存者割合となる。

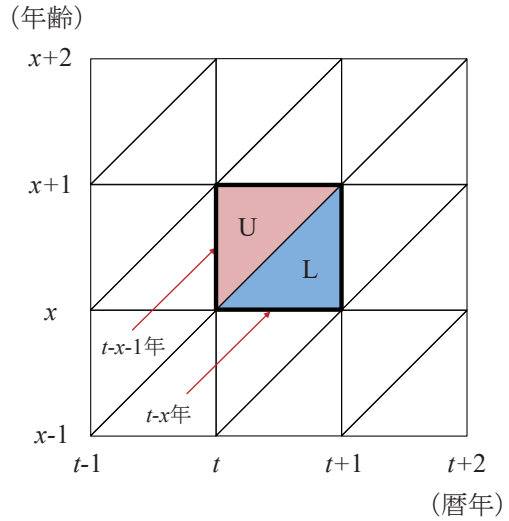
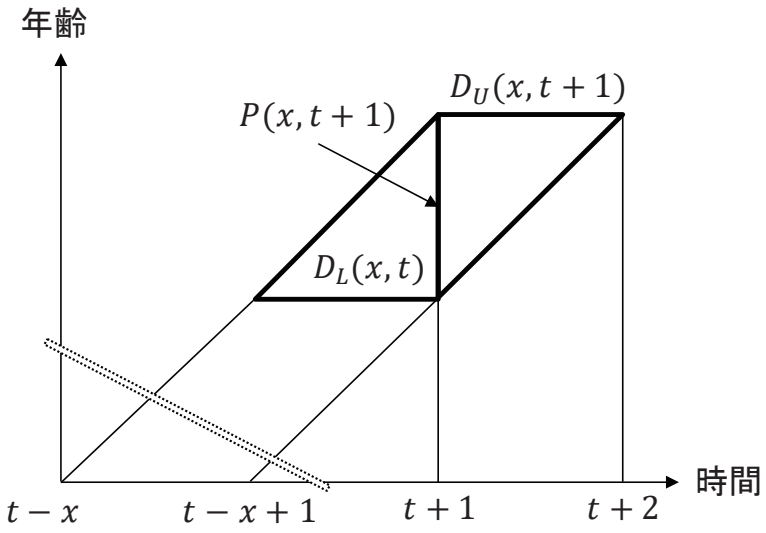


図 2　レキシストライアングル



出典：Methods Protocol for the Human Mortality Database（2017:33）

図 3　コーホート年齢別死亡確率（q_x）に必要なデータ

3.2　人口動態統計

人口動態統計の出生票では、年次別母年齢別の出生数が表章されている。ここから、調査対象者の調査時年齢に上限を設けたときに、どれほどの割合の出生数が SSM2015 データから把握できないかを算出できる。

なお、人口動態統計の出生は日本国籍児が集計対象となっており、日本人女性が生んだ子どもだけではなく、父日本人・母外国人の組み合わせからなる夫婦が生んだ子どもも含まれている。SSM2015 の調査対象者は日本国籍を持つ 20～79 歳男女であるため、厳密には日本人女性の出生に限定することが望ましいが、以下では e-stat からも容易に入手できる人口動態統計の定義にもとづいた出生データを用いる。

4．分析結果

4.1　コーホート別生存者割合

はじめに、調査時点において各出生コーホートの成員がどれほど生存しているかを見ていく[6]。図 4 は、コーホート年齢別死亡率から x 歳時生存者割合（l_x）を算出してプロットしたものである。

例えば、HMD からデータが得られる最も古いコーホートの 1947 年コーホートに着目してみよう。この出生コーホートは 2015 年 1 月 1 日時点で満 67 歳に相当するコーホートであるが、連続年齢（exact age）で 67 歳到達時に女性では約 20%、男性では 30%近くがすでに死亡していることがうかがえる。これより古いコーホートのデータは HMD からは得られないが、SSM2015 の調査時に 70 歳代のコーホートでは死亡による子世代データの欠測がより深刻になることは明らかである。

ただし、本稿の目的はあくまで子世代データの代表性を検討することであるため、再生産年齢以前に発生した死亡はここでは考慮する必要がない。なぜなら、15 歳未満人口の大半は無子であるため、そこで生じた死亡は子世代データの代表性にほぼ影響しないためである。

そこで、再生産年齢の起点である 15 歳に達した人口のうち、x 歳時点で生存している割合、

$$l_x \ / \ l_{15} \ \ (x \geq 15)$$

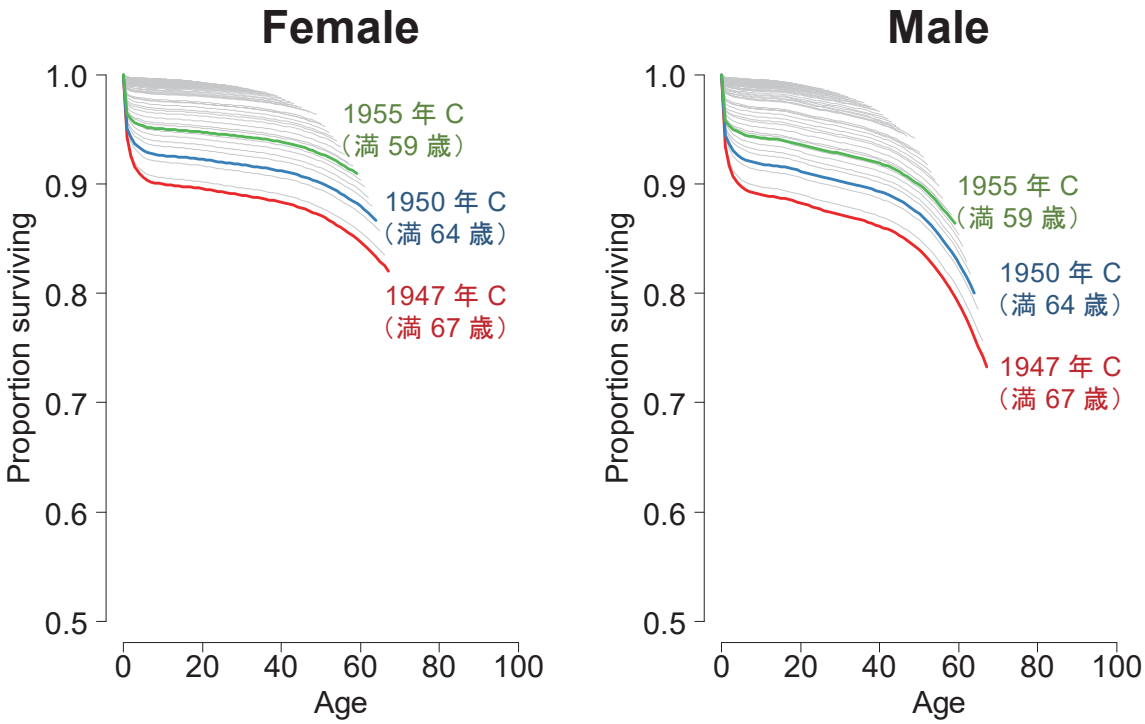
に着目する。これを示したのが図 5 である。1947 年コーホートで 15 歳まで生存した人口のうち 67 歳までに死亡した割合は、女性では 10%未満となっているが、男性では 15%を超えている。図 4 の場合と同様、調査時点で 70 歳代のコーホートでは、これら以上に死亡による潜在的バイアスが大きくなる。

---

[6] 厳密には、以下で検討していくのは 2015-x年出生コーホートの x-1 歳時生存者割合であり、調査時点における当該コーホートの生存者割合とは若干の乖離がある。後者は、

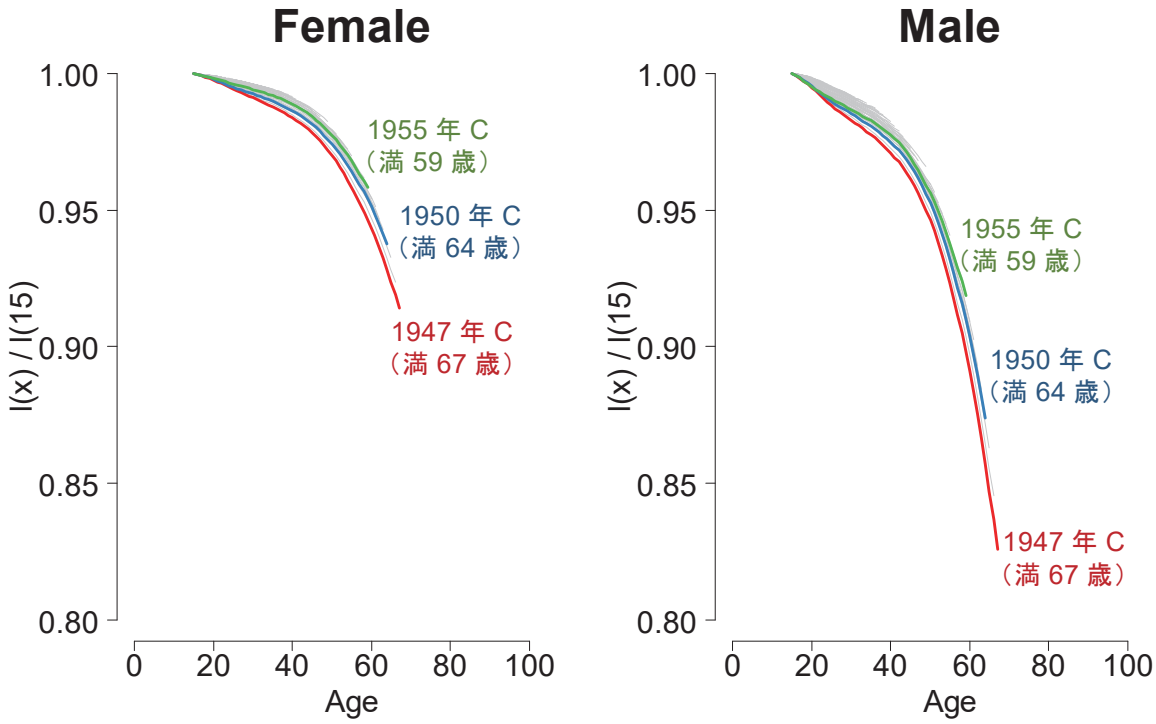
$$\frac{P(x,2015)}{P(0,2015-x+1)+D_L(0,2015-x)}$$

で表される。分母は封鎖人口のもとでは 2015-x年コーホートの出生数に等しい。



Source: Human Mortality Database

図 4　出生コーホート別にみた、x 歳時生存者割合



Source: Human Mortality Database

図 5　出生コーホート別にみた、15 歳時生存人口の x 歳時生存者割合（$x \geqq 15$）

4.2　（母）親年齢と子どもの出生年

つづいて、親世代の調査時年齢に限定をかけた際に、SSM2015 では補足できなくなる出生データがどれほど存在するのかを、人口動態統計を用いて見ていこう。図 6 には、女性回答者の調査時年齢の上限を 50 歳と 65 歳にそれぞれ設定した場合の、欠測データになる出生数の割合を示したものである。仮に分析対象を調査時年齢が 50 歳未満の女性回答者に限定した場合、2005 年あたりまでは大半の出生をカバーできているが、それ以降は欠測の割合が急増することがわかる。一方、分析対象を 65 歳未満の女性回答者にした場合、1985 年付近までは全体の出生の大半が SSM2015 のカバレッジに入っていることがうかがえる。

ただし、こうした欠測の程度は出生順位にも依存する。図 7 は出生順位別に欠測データになる出生数の割合をプロットしたものである。高順位の出生ほど欠測割合が上昇し始める年次が早いことがわかる。これは、出生順位が高いほど母親の年齢が高いためである。

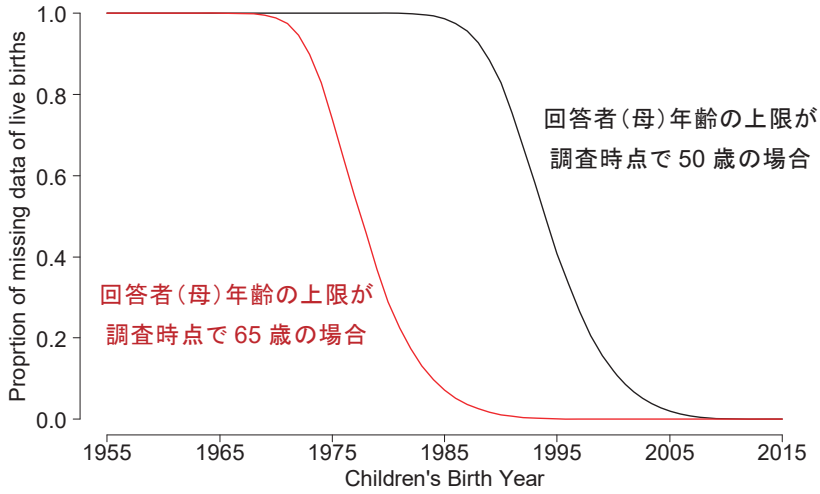


図 6　欠測データになる出生数の割合（全出生）

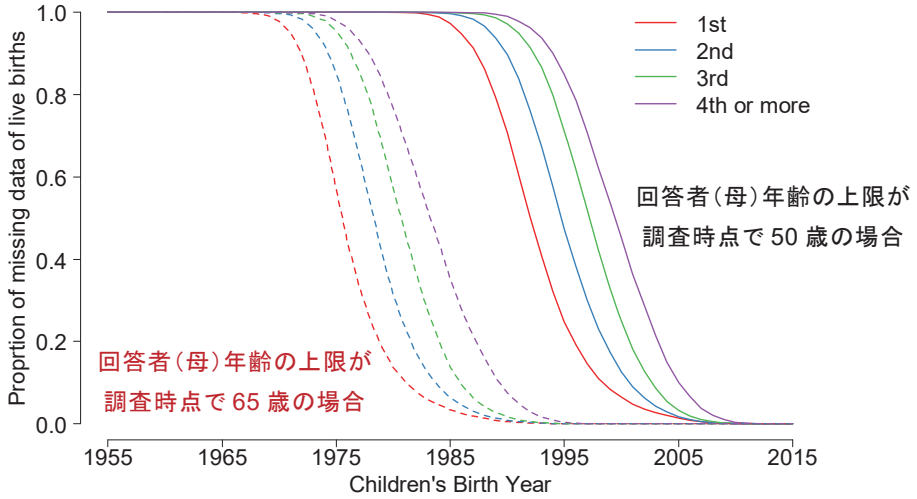


図 7　欠測データになる出生数の割合（出生順位別）

５．まとめ

調査対象者の子どもの情報が豊富に収集されていることは SSM2015 の特徴のひとつである。しかし、調査対象者の出生コーホートにおいて調査時点までに死亡した成員の子どもの情報は必ず欠測になる。さらに、調査年から過去にさかのぼった年次ほど子どもの情報は若齢出産によるものに限定される。そこで本稿では、SSM2015 の子世代データの代表性について、Human Mortality Database ならびに人口動態統計を用いて評価してきた。得られた知見を要約すると以下のとおりである。

HMD から死亡データを把握できる最も古い 1947 年コーホートの調査時点の生存者割合は女性では約 80%、男性では約 70%であった。調査時点で 70 歳代のコーホートについては直接データが得られなかったが、これらよりも生存者割合が低いことは確実である。

ところがこのような死亡によるデータ欠測を避けるために親世代の調査時年齢に限定を設けた際、一方で問題になるのは、SSM2015 で補足できる親出産時年齢別の出生データの範囲が限定されることである。仮に親世代の調査時年齢の上限を 50 歳と 65 歳に設定した場合、それぞれ 2005 年と 1985 年あたりまでは、SSM2015 の子世代データでも親出生時年齢別の出生データに深刻な偏りはないと推測される。

以上を踏まえると、SSM2015 の子世代データを用いる際に、どのような注意が必要であろうか。これは上記の 2 つの要因から生じる欠測データの大きさをどのように評価するかに依存するが、ここではひとつの方針を提言したい。第 1 に、SSM2015 の調査時点で少なくとも 70 歳代の調査対象者の子世代データは使用しないか、使用する際には 70 歳代の調査対象者を分析対象に含めた場合とそうでない場合とで分析結果を比較するなどの感度分析等を行うことが望ましい。先述のとおり、これらのコーホートでは調査時までの死亡者割合が女性では 20%、男性では 30%を上回ることは確実であり、これだけのコーホート成員の子どもの情報が全く得られないことは無視できないバイアスを生みかねない。第 2 に、第 1 の点も勘案すると、調査時の親年齢を 65 歳未満に限定し、かつ 1985 年以降に生まれた子どもの情報のみを使うことが考えられる。65 歳というカットポイントはやや恣意的であるが、60 歳代後半では加齢に伴う死亡リスクの上昇も加速していくこと、また調査時年齢の上限を 60 歳まで下げると子どもの出生年にかけるべき限定がより厳しくなってくる（出生年のレンジを小さくしなくてはならない）こと、を考慮した。

ただし、子世代データに対してこのような限定が必要か否かは、当然研究目的にも依存する。例えば、高齢者の子どもとの同別居の状況など、調査時点における母集団の静態を把握することが目的の場合は、ここまで議論してきたような限定をデータにかける必要はない。一方で、子どもの出生年別にある変数の集計を行う場合は、本稿で論じてきた 2 つの要因によるデータ欠測に注意を払う必要がある。

最後に、データ欠測が必ずバイアスに直結するわけではないことも付言しておきたい。第

1 に、子世代データのある変数の分布自体に関心がある場合、例えば、出生コーホート別の学歴分布を算出するとしよう。この場合、学歴分布を正しく推定できないのは、データ欠測が学歴水準に依存する場合である。言い換えれば、学歴とは独立に欠測が生じている場合は当然のことながら学歴分布に影響はない。

第 2 に、変数の分布ではなく変数間の連関に関心がある場合はデータ欠測に対してもう少し楽観的な立場を取れるかもしれない。例えば、親世代と子世代との間で学歴の連関を探る場合を想定しよう。この場合、仮にデータ欠測が親世代あるいは子世代の学歴と相関しているとしても、親世代と子世代の学歴の「組み合わせ」と相関していなければ、周辺度数を統制した親子間の学歴の連関の指標（例えばオッズ比）は影響を受けない。言い換えれば、データ欠測を従属変数とした場合、親学歴あるいは子学歴の主効果が存在しても、親学歴×子学歴の交互作用効果が存在しなければ、親子間の学歴の連関は偏りなく推定できる。

しかしながら、どのような欠測メカニズムが働いているか特定できない以上、データ欠測による潜在的なバイアスに対して常に慎重な立場を取るべきであろう。SSM2015 では調査対象者を中心にその「親」と「子」にまでまたがった三世代のデータが取れているが、それぞれの世代のデータが何を測っているかについて内省的になる必要がある。しかし、それは裏を返せば、それぞれの世代のデータの特性を理解したうえで分析すれば、過去の SSM 調査では難しかった、格差・不平等の多世代分析（Mare 2011）を可能にするものと思われる。

【文献】

Duncan, Otis D., 1966, “Methodological Issues in the Analysis of Social Mobility.” S.M. Lipset and N. J. Smelser (Eds.), *Social structure and mobility in economic development,* Aldine: 51-97.

Human Mortality Database (HMD), 2017, University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (data downloaded on 01/08/2017).

石井太, 2015,「日本版死亡データベースの構築に関する研究」『人口問題研究』71(1): 3-27.

Mare, Robert D., 2011, “A Multigenerational View of Inequality.” *Demography* 48(1): 1-23.

Song, Xi and Robert D. Mare, 2015, “Prospective Versus Retrospective Approaches to the Study of Intergenerational Social Mobility.” *Sociological Methods and Research* 44(4): 555-584.

Analysis of the Quality of Data for Offspring Generation

Shohei Yoda
(National Institute of Population and Social Security Research)

Abstract

This paper assesses the quality of data for offspring generation of SSM2015 respondents. In the 2015 survey, respondents were asked questions about their offspring's characteristics. Data on respondents' offspring, however, could be distorted and not representative of their generation or birth cohort. The distortion mainly arises from two factors. The first is cohort mortality: cohort members who died before the survey date were necessarily omitted from the survey population. Data on their children would also be uniformly missing because children can be reported only if they have surviving parents. Second, children in older birth cohorts are likely to have a younger mother, resulting in a lack of representativeness in terms of the parent's age at giving birth. This study used data from the Human Mortality Database and Vital Statistics to evaluate the potential distortions due to the missing data mentioned above. Several strategies for analyzing the offspring generation data of the SSM2015 are discussed in the last section.

Keywords: generation, cohort mortality, births at early age