1. 欠測データに伴う問題

統計調査においては、無回答や無記入により,調査客体又は調査項目の一部について情報を得られないことがある。このような統計調査から作成されるデータは、本来観測・記録されるべき値の一部が観測・記録されておらず、「不完全データ（incomplete data)」と呼ばれる。不完全データを用いて推定を行う場合、観測された値のみを用いて推定を行うことが最も直截な方法（「完全ケース分析」と呼ばれる。第 2.1 節参照）であり、実際に広く行われているが、このような推定には、「欠測バイアス」と「推定精度の低下」という２つの問題がある。

○欠測バイアス

不完全データでは、標本設計において意図された目標母集団の代表性が損なわれている可能性があるため、推定にバイアスを伴うおそれがある。データの一部が観測されないことによって、推定に生じるバイアスを「欠測バイアス」と呼ぶ。

例として、個人の所得額の平均値を推定するための統計調査を考える。仮に回答者の大半が学生や無業者であり、無回答者の大半が残業の多い高所得者である場合、当該調査から推定される所得額の平均値は、真の値である目標母集団の所得額の平均値を下回ることが考えられる。このとき下方の欠測バイアスが生じている。

平均値の推定における欠測バイアスは、次のように、より一般化して示すことができる。まず、統計調査の目的を、目標母集団Uの平均値μの推定とする。ここでは、目標母集団Uが、標本に含まれた場合には必ず回答する調査客体の集合U_Rと、必ず回答しない調査客体の集合U_Mに分割できるとする。目標母集団Uのなかで回答者集合U_Rに含まれる要素の割合をπ_Rとし、無回答者集合U_Mに含まれる要素の割合をπ_Mとする（$\pi_R + \pi_M = 1$)。目標母集団Uの平均値μは、回答者集合U_Rの部分母集団平均μ_Rと無回答者集合U_Mの部分母集団平均μ_Mとの部分集合構成比による加重平均に等しい（$\mu = \pi_R\mu_R + \pi_M\mu_M$)。標本抽出の結果得られる標本は、調査実施後に回答者と無回答者の部分標本に分割される。回答者の部分標本のみを用いて算出した平均値$\bar{y}_R$の推定バイアスは、$\mathrm{Bias}(\bar{y}_R) = \mu_R - \mu = (1 - \pi_R)(\mu_R - \mu_M)$である。数式から明らかなとおり、(1)目標母集団$U$における回答者の割合$\pi_R$が小さいほど、あるいは(2) 回答者部分母集団$U_R$と無回答者部分母集団$U_M$の（値$\mu_R - \mu_M$で測られる）異質性が大きいほど、欠測バイアスは大きくなる。第１の点は、欠測率が大きいほど欠測バイアスは大きいということであ

る。第２の点をより一般化して表現すると、「回答の成否と当該変数の間の相互依存性が強いほど欠測バイアスは大きくなる」といえるが、この点については、第 1.1.1 節で明らかにする。

○推定精度の低下

不完全データを用いた推定に伴う第２の問題は、不完全データでは本来得られるべき情報の一部が失われているために、推定の精度が低下することである。たとえば、平均値の推定における欠測による推定精度の低下については、標本サイズが縮小している分だけ、標本平均の標準誤差が増加するので、推定精度の低下の程度が分かる。

当然ながら、第１の問題（欠測バイアス）の方が、第２の問題（推定精度の低下）よりも重要であり、優先的に対処することが求められる。したがって、欠測データの統計的処理は、欠測バイアスの問題を解決することを第１の目標とし、この目的を果たす限りにおいて、第２の目標である推定精度の改善を目指す。ただし不完全データの統計的処理にはさまざまな手法があり、分析対象となる不完全データの性質や統計調査の目的に応じて適切な手法を用いることが重要である。

1.1 欠測データ処理方法の適性を決める諸条件

欠測を含むデータ、すなわち不完全データに基づく推定において、欠測バイアスを緩和、ないし除去する統計的手法にはいくつかあるが、それらの手法ごとの適性を決める条件を列挙すると次のとおりである。

(1) 欠測データメカニズム
(2) 補助的な変数の利用可能性
(3) 推定目標
    (3.1) 推定対象となる母集団特性値のモーメント次数
    (3.2) 点推定か区間推定かの別
(4) 欠測パターンと欠測率

このなかで最も重要なのは、(1)欠測データメカニズムである。これについては、第 1.1.1 節で概要を示す。また、(2)補助的な変数の利用可能性は、(1)欠測データメカニズムと関連している。第 1.1.1 節では、その関連性も適宜指摘する。その他の条件(3)推定目標及び(4)欠測パターンと欠測率については、第 1.2 節で概説する。

1.1.1 欠測データメカニズムと欠測データ処理方法の適性

欠測データの統計的処理法の適性に影響する諸条件のなかで、最も重要なのは、「欠測データメカニズム」である。「欠測データメカニズム」は、簡単にいえば「欠測の生じるしくみ」である。欠測データメカニズムには３種類があり、欠測の生じる確率的メカニズムの違いによって区別されるが、ここではまず３種類のそれぞれを直感的に説明する。

〇完全にランダムな欠測（missing completely at random: MCAR）

変数の欠測する確率が、当該変数の値及び他の観測されている変数の値に依存しない場合のことである。

たとえば、調査対象者が硬貨を投げて、表が出るか裏が出るかに応じて、調査に協力するか否かを決めているとする。このとき、観測されたデータの標本は、目標母集団の縮図としての性格を失ってはいないとみることができるので、観測された値のみを用いた推定に欠測バイアスは生じない。

上述の例は現実的ではないにしても、それに近いことが実際に起こり得ないわけではない。たとえば、「調査対象者の身長」を調査項目とする標本調査で、無回答者の大部分が、調査票を送付してから回収締め切りまでの期間に住居を不在にしていた者であったとする。この場合、「調査対象者の身長」と住居長期不在の事象とは独立であると考えられる（ある一定期間に住居不在となる確率は、当該者の身長に依存しない）ため、長期不在を理由とする「調査対象者の身長」の欠測（無回答）は、MCAR に極めて近いと考えられる。

MCAR は強い仮定であり、現実的には MCAR が妥当であると考えられる事象は極めて少ない。

〇ランダムな欠測（missing at random: MAR）

変数の欠測する確率が、当該変数の観測された値及び他の観測されている変数の値には依存するが、欠測となった変数の値には依存しない場合のことである。

たとえば、目標母集団の平均値を推定する統計調査で、回答者の大半が学生や無業者であり、無回答者の大半が有業者である場合、所得という調査変数の値が欠測する確率は、調査対象者の就業状態という変数の値に依存している。この場合、所得が観測される標本は、学生や無業者に偏ってしまうため、目標母集団の平均所得の推定には、無業者側への下方バイアスが生じる。

ただしこの場合、就業状態がすべての調査対象について観測されていれば、所得が観測されている部分標本の学生・無業者側への偏りを補正することが可能

である。すなわち、「有業者は学生・無業者よりも所得が観測されにくい」という追加的な情報と就業状態の分布の情報を平均所得の推定に利用することで、欠測バイアスを緩和することができる。単純な例として、「学生・無業者は必ず所得額を回答するが、有業者は 50%の確率でしか所得額を回答しない」とすると、標本を学生・無業者と有業者とに分割すれば、それぞれの母集団の部分集合の縮図が再現される。部分標本ごとに観測された値のみを用いて標本平均を計算し、欠測を含む標本全体の就業状態構成比でそれらを加重平均すればバイアスのない推定となる。つまり、有業者の値に学生・無業者の値の 2 倍のウェイトを付けて加重平均を算出するという方法がバイアスのない推定方法のひとつとなる。

〇ランダムでない欠測（missing not at random: MNAR）

変数の欠測する確率が、その変数自体の値に依存する場合のことである。

たとえば、資産保有額の母集団平均を推定するための標本調査において、低中位資産額階級と比べて上位資産額階級は資産保有額の情報を秘匿する傾向が強いとする。このような場合、標本が低中位資産階級に偏る（欠測バイアスが生じる）。この場合は MAR と異なり、バイアスの問題を緩和するのは容易ではない。標本が低中位資産階級に偏っていること自体は分かっていても、資産保有額の情報が低中位資産階級の部分しか得られていないため、MAR の場合に示したような偏りの補正を実行することはできない。この場合は、欠測が生じるしくみをモデル化する必要がある。MNAR の下では、MCAR や MAR の場合と異なり、手持ちの情報だけではバイアスのない推定を行うことができないため、“モデルの力を借りる”必要がある。

MNAR と MAR の違いは欠測確率が欠測する変数の値に依存するか否かという点であるが、これは関連する他の変数の利用可能性に関係している。ここで、説明の便宜上、上位資産階級は北部地域に住む住民が大部分を占め、低中位資産階級は南部地域に住む住民によって構成されているという仮想的な経済を考える。上述の資産保有額に関する標本調査の例では、調査客体の居住地域の情報は調査項目として収集されていないと考えている。仮に標本に含まれるすべての調査客体について居住地域に関する情報が得られていれば、上述の説明とは状況が違ってくる。すなわち、上位資産階級が低中位資産階級と比べて資産保有額の欠測を生じやすいということの裏返しとして、北部地域に住む住民は南部地域に住む住民よりも資産保有額の欠測を生じやすいといえる。居住地域という関連する変数が利用できない場合は、保有資産額の欠測する確率が保有資産額自体の値に依存している（すなわち MNAR である）と言わざるを得ないのに対して、居住地域という関連する変数が利用できる場合は、保有資産額の欠測する

確率が居住地域に依存しており、とりわけ、居住地域で条件付ければ、保有資産額の欠測する確率が保有資産額自体の値に依存しない（すなわち MAR である）ということができる。欠測確率と欠測する変数自体の値との間に相関があっても、条件付けることで、欠測する変数自体の値に対する欠測確率の依存性を消去できるような補助変数が、すべての調査客体について観測されていれば、それは MAR であるといえる。逆にそのような補助変数が理念的に存在しても、すべての調査客体について観測されていなければ（すなわちデータとして利用可能でなければ）、それは MNAR と異ならない。

1.1.2 図による解説

欠測データの統計的処理は、MCAR、MAR、MNAR の順に難しくなる。このことを図１－１～１－３に基づいて説明する。具体的なイメージをつかみやすくするために、ここでは世帯が保有する金融資産の額を対象とする。

○MCAR

図１－１は、MCAR の場合を示したものである。図１－１（イ）は、正しく設計された標本抽出に従って得られた標本で、仮に金融資産保有額 Y の値がすべての調査客体について観測されるとした場合の、金融資産保有額 Y のヒストグラムである。現実には無回答により、一部の調査客体について金融資産保有額 Y の値が観測されない。図１－１（ロ）は、図１－１（イ）の標本で実際に無回答による欠測が発生した場合の、回答者と無回答者とを区別した金融資産保有額 Y の合成ヒストグラムである。灰色部分が欠測値、白色部分が観測値をそれぞれ表す。図１－１（ロ）における回答者と無回答者とを分けて、それぞれについての金融資産保有額 Y のヒストグラムを示したものが図１－１（ハ）及び（ニ）である。図１－１（イ）および（ハ）に示された点線は、それぞれの観測されたヒストグラムから推定される金融資産保有額 Y の分布を表す。図１－１（イ）では、標本設計が正しい限り、真の分布を偏りなく推定できる。

図１－１（ロ）をみると、金融資産保有額 Y の値による階級区分ごとの回答率が等しいことが分かる。この特徴が本例題における MCAR の条件を反映している。この場合、回答率が金融資産保有額 Y の値に依存していない。そのため、図１－１（ハ）及び（ニ）のヒストグラムはいずれも真の姿（図１－１（イ））と比べて左右に偏ることなく、図１－１（イ）のヒストグラムを縦軸方向に定率で縮小したものとなっている。そして、図１－１（ハ）の点線に示すとおり、観測された値のみを用いた分布の推定は、欠測がなければ正しく推定される分布（図１－１（イ）の点線）と互いに縦方向に定率倍した関係となる。つまり MCAR

の場合、標本サイズが縮小する（推定の精度が落ちる）だけで欠測バイアスは生じない。

○MNAR

一方、図１－２は、MNAR の場合に生じる推定上の問題を同様に示したものである。図１－２（ロ）をみると、金融資産保有額 Y の値による階級区分ごとの回答率が互いに大きく異なっていることが分かる。この特徴が本例題における MNAR の条件を反映している。金融資産保有額 Y の値が大きい階層ほど回答率が低いため、図１－２（ハ）のヒストグラムは左側に、図１－２（ニ）のヒストグラムは右側にそれぞれ偏る。そして、図１－２（ハ）の点線に示すとおり、観測された値のみを用いた分布の推定は、欠測がなければ正しく推定される分布（図１－２（イ）の点線）と比べて、低位資産階層側に偏ることになる。分布のこの偏りが、あらゆる推定量の欠測バイアスの源泉である。

○MAR

図１－３は、MAR の場合を示したものである。図１－３（イ）及び（ロ）は、MNAR の場合の図１－２（イ）及び（ロ）と全く同じである。MAR が MNAR と異なるのは、推定の欠測バイアスを緩和するために活用できる補助的な変数が観測されているという点である。図１－３は MAR のもとでの欠測バイアス問題への対処を示す。MAR の場合は、他の観測された情報を用いて、金融資産保有額 Y の値に基づいて分けられた（図では７つの）階層をさらに細分化できる。たとえば、世帯主の就業状態を有業か無業かの２値変数 X として、その値に基づいて各資産階層を２つに分け、それぞれのグループで観測値と欠測値を区別したものが図１－３（ニ）である。欠測が生じなかった場合の金融資産保有額 Y と、就業状態 X の同時分布の情報を図１－３（ハ）に示す。つまり図１－３（ニ）は、図１－３（ロ）と（ハ）の情報を統合したものである。図１－３（ハ）及び（ニ）では、７つの資産階層が、就業状態に基づいてそれぞれ左右に分かれており、ヒストグラムの各棒（各保有資産階層）の左側を無業世帯主、右側を有業世帯主とする。図１－３（ハ）によると、金融資産保有額が低い階層では、無業者世帯の割合が高く、金融資産保有額が高い階層では、有業者世帯の割合が高くなっている。また、図１－３（ニ）によると、無業者世帯の方が有業者世帯よりも回答率が高くなっている。

ここでの設定としては、図１－３（ロ）から（ニ）へ変形しなければ、すなわち、保有金融資産階層の世帯主就業状態による細分化を行わなければ、保有金融資産階層ごとの回答率の分布は、MNAR を表す図１－２（ロ）のヒストグラムと異なるところがない、という点が重要である。つまり、欠測の起こり方は、一

見するとMNARと同様に、金融資産保有額Yの値が大きい階層ほど回答率が低い。しかしこの場合、それは見せかけの関係であり、世帯主就業状態Xの値で条件付けることにより、金融資産保有額の観測確率が、金融資産保有額の値自体に依存しない部分を取り出すことができる。そのことを示したのが、図１－３（ホ）及び（ヘ）である。図１－３（ホ）及び（ヘ）は、図１－３（ニ）を世帯主の就業所帯に応じて分割したものである。図１－３（ホ）及び（ヘ）は、順に無業者世帯及び有業者世帯それぞれの回答者と、無回答者とを区別した金融資産保有額Yの合成ヒストグラムである。図１－３（ホ）及び（ヘ）のそれぞれでは、回答率が金融資産保有額Yの値に依存していない。つまり世帯主の就業状態で条件付けたとき、MCARと同様の欠測状況が出現している。

図１－３（ト）及び（チ）は、順に図１－３（ホ）及び（ヘ）それぞれの観測データに基づいて金融資産保有額Yの条件付分布を推定した結果である。図１－１のMCARの場合と同様の理由で、世帯主の就業状態による条件付分布は偏りなく推定できる。最後に、図１－３（ト）無業者世帯の金融資産保有額分布の推定結果及び（チ）有業者世帯の金融資産保有額分布の推定結果に、図１－３（ハ）金融資産保有額Yと世帯主就業状態Xの同時分布の情報を合わせることができれば、全体の金融資産保有額Yの分布を偏りなく推定できる。そのことを表したのが図１－３（リ）である。

MNARに似た欠測の発生状況でありながら、適当な補助変数で条件付けることにより、MCARに似た状況を部分的に取り出すことができる、換言すれば、観測された情報によって適当に層化することで、層ごとに欠測の起こりやすさが欠測した変数の値に依存しないようにできるのが、MARである。

◇まとめ

まとめると、第１に、MCARのもとでは、観測された情報のみを用いた推定に欠測バイアスは生じない。第２に、MARのもとでは、観測された情報に基づいて条件付ける（層化する）ことで、条件（層）ごとに欠測バイアスを緩和できる。第３に、MNARのもとでは、欠測バイアスを緩和できるような条件付け（層化）は、観測された情報の中には存在しない。第2.6節でみるように、MNARに対しては、欠測データメカニズムのモデル化を行うことで、欠測バイアスの問題に対処する。

図１－１　MCAR

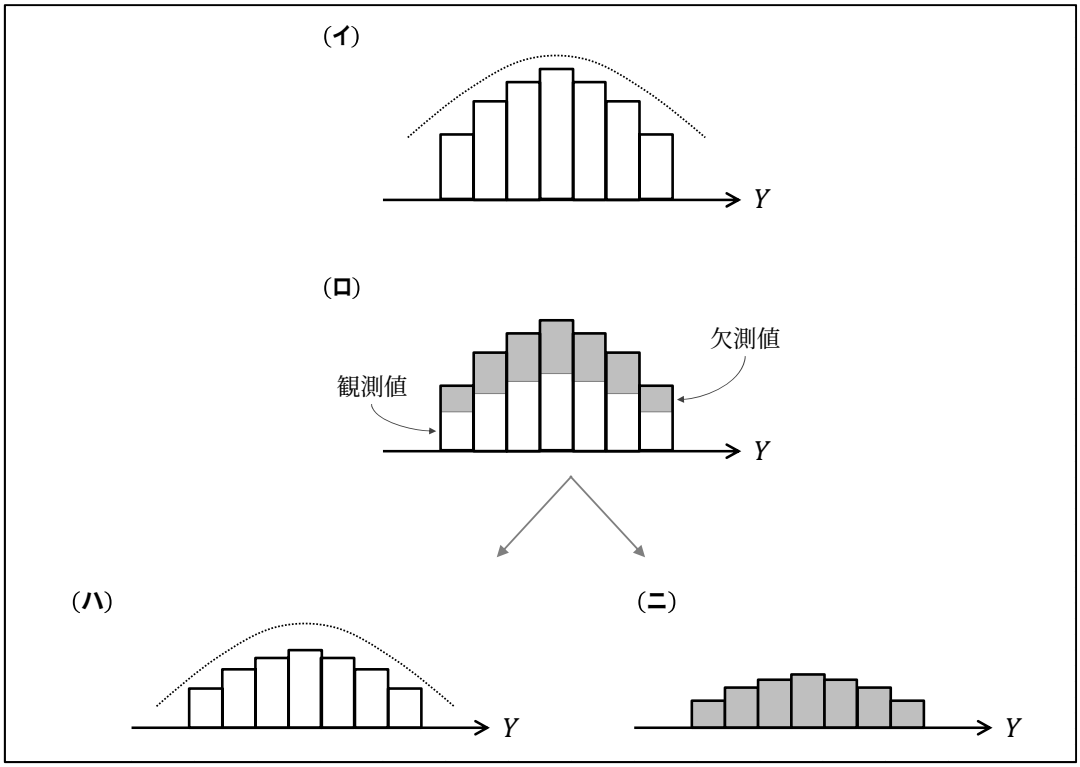


図１－２　MNAR

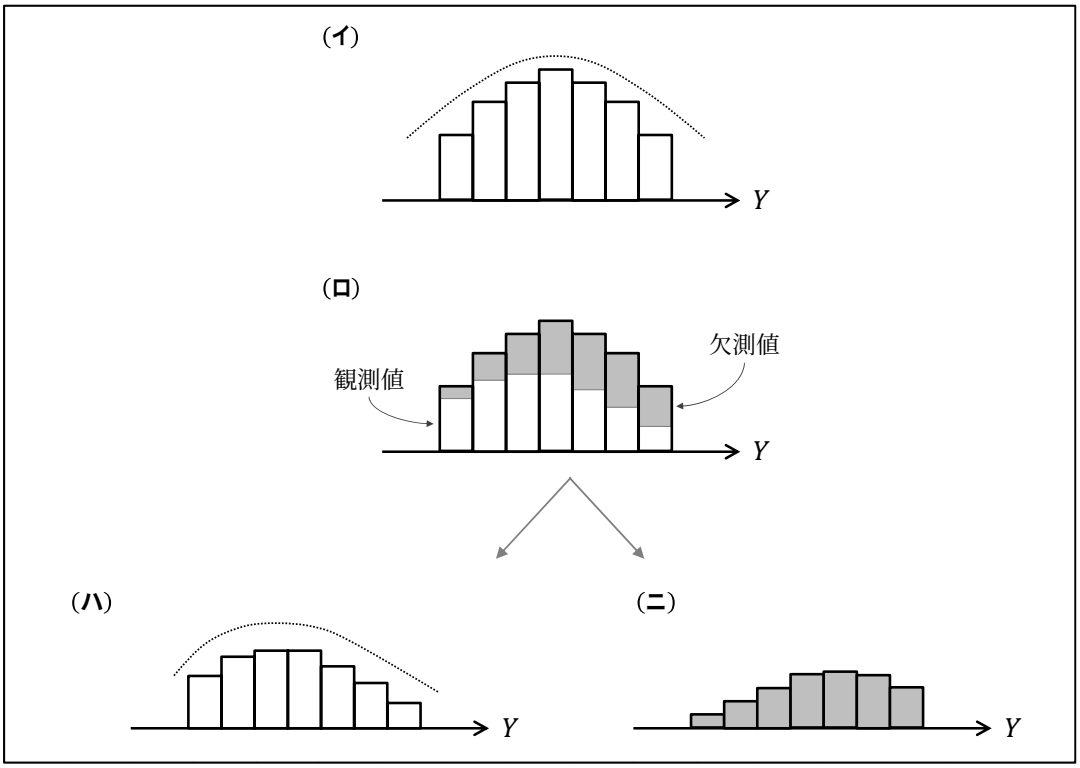
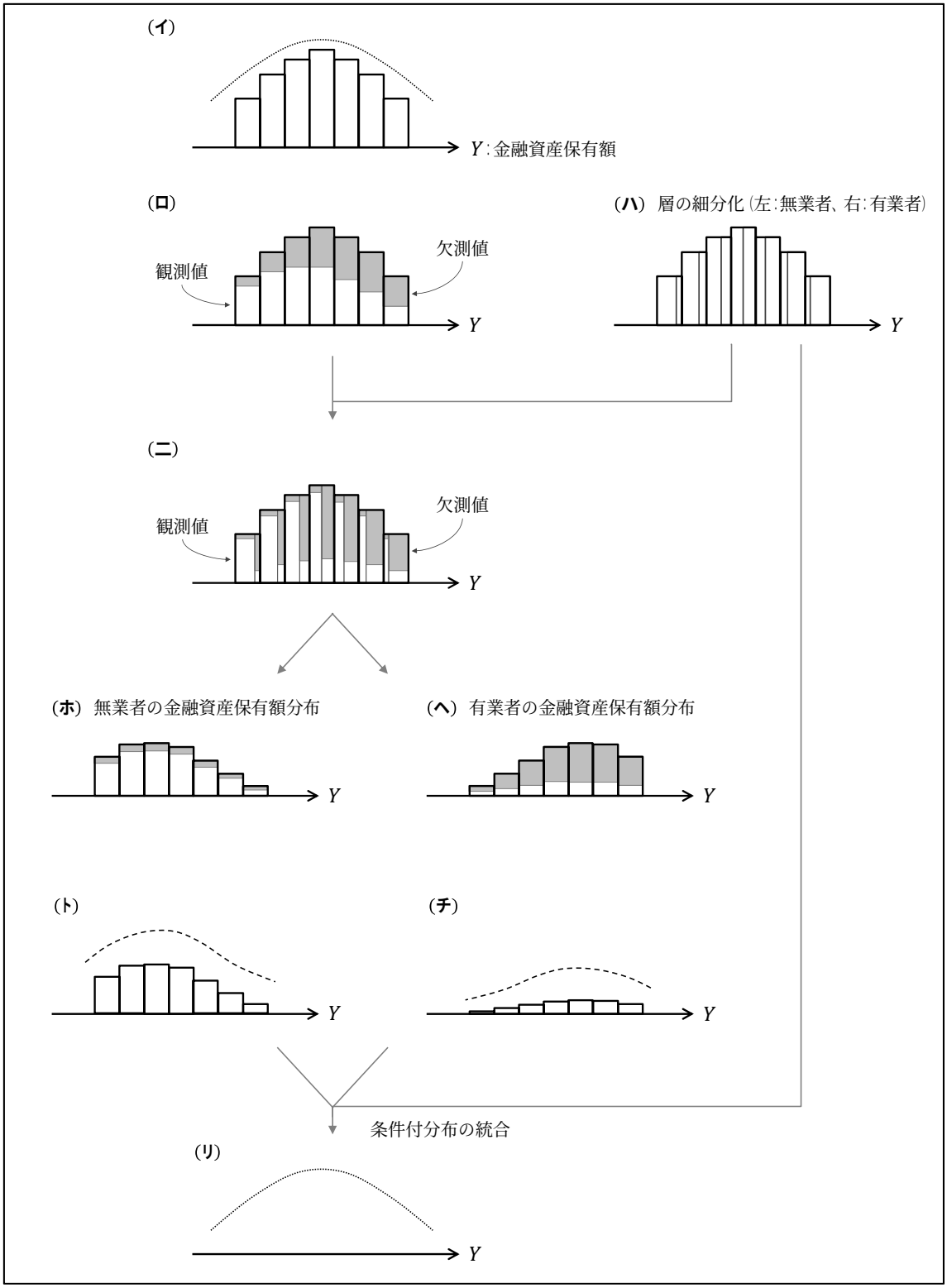


図１－３　MAR


(イ)
Y:金融資産保有額
(ロ)
観測値
欠測値
Y
(ハ) 層の細分化(左:無業者、右:有業者)
Y
(ニ)
観測値
欠測値
Y
(ホ) 無業者の金融資産保有額分布
Y
(ヘ) 有業者の金融資産保有額分布
Y
(ト)
Y
(チ)
Y
条件付分布の統合
(リ)
Y

1.2 統計調査ごとの目的・性質と欠測データ処理方法の適性

第 1.1 節冒頭で列挙したとおり、欠測データメカニズム以外にも欠測データ処理法の適性に影響を与える条件がある。それらは個別統計調査ごとの目的及び性質にかかわるものである。

公的統計において推定対象となるのは、多くの場合、母集団平均、母集団総計、母集団割合等の１次モーメントである。(一般的に、確率変数Yに対して期待値$E(Y^h)$を「確率変数Yのh次モーメント」と呼ぶ。通常１次モーメントと２次モーメントが興味の対象となることが多く、１次モーメントは平均値、２次モーメントのうち、平均値からの乖離は分散とよばれる。平均値は当該変数の「代表的な値」、分散は当該変数の「ばらつきの程度」の尺度である。）推定目標となる母集団特性のモーメント次数は、欠測データ処理法の適性を決める条件のひとつである。欠測に伴う分布のゆがみの補正結果は、手法ごとに異なるため、特に推定目標が１次モーメントであるか、１次より大きいモーメントであるかが手法の適性を大きく左右する。推定目標が１次モーメントである場合、補正後の分布が真の分布と対称性に関して同等となるような手法は、すべて推定に欠測バイアスをもたらさないといえる。

次に、推定目的が点推定にとどまるものか、区間推定ないし統計的仮説検定にも及ぶものであるかということも、欠測データ処理法の適性を決める。当然ながら、点推定のみを目的とする分析の方が、適切な処理法の選択肢の範囲が広い。欠測に伴う分布のゆがみの補正結果が、欠測バイアスを除去ないし緩和するものであっても、補正後の分布のばらつきが真の分布よりも小さくなる場合が多い。このような場合は、推定値の標準誤差が過小評価されるため、当該手法は区間推定ないし統計的仮説検定には適さないといえる。社会的な認識の大勢としては、公的統計の目的は点推定にとどまると考えられ、区間推定ないし統計的仮説検定には適さない手法の多くが従来用いられてきた。

公的統計においては、統計調査の目的よりもデータの性質の方が、欠測データ処理法の適性により大きく影響すると考えられる。データの性質としては主に、(1)どのような補助変数（欠測が生じるしくみをモデル化する際に説明変数となり得る変数）が利用可能か、(2)目的となる変数を適当に加工したときにパラメトリックな分布（正規分布やポワソン分布など）で近似できるか、(3)調査項目の欠測可能性について先験的な知見があるか、という点が重要である。

第１の点については、公的統計における補助変数は、多くの場合、フレーム（標本抽出を行うための調査客体リストすなわち母集団データベース）に情報として含まれる調査客体属性である。統計調査の結果として不完全データが与えられたとき、適当な補助変数で標本を層化し、層ごとの回答率を確認することは有

益である。層ごとの回答率に顕著な傾向がみられる場合、MAR を仮定した推定において、当該補助変数を利用することができる。第 1.1.2 節図１－３の例では、世帯ごとの金融資産保有額の欠測に関して、世帯主の就業状態が有用な補助変数となっている。このような補助変数が利用可能か否かによって、データが MAR か MNAR のどちらであるかが決まると考えることもできる。

第２の点については、欠測データ処理の手法の中で、多重代入法や尤度法のようなモデル依存性の高い方法による場合、目的となる変数を加工して定型的な分布に可能な限り近づけることで、推定の一致性を保証する適切なモデルの特定化が可能となる。たとえば、個人の所得や企業の規模のように裾の長い分布を示す変数は対数変換することで、正規分布に近い分布を示す変数を得ることができる。

第３の点については、変数の欠測可能性に関する先験的な知見が、データからは検証することのできない分析の前提を裏付けるものとして、重要な役割を果たす。特に、欠測可能性に関する先験的知見は、有意な補助変数の選択に役立つだけでなく、欠測データメカニズムのモデルの定式化にも示唆を与える。たとえば調査協力に対する謝礼は、調査対象主体が回答と無回答を選択する意思決定における誘因となり、回答群の選択原理として作用する。一定額の謝礼から得られる効用が、調査対象主体ごとに異なるためである。たとえば、謝礼から得られる追加的な効用は、通常、所得の高い主体にとっては小さいが、所得の低い主体にとっては大きいと考えられ、所得の小さい主体ほど回答確率が大きい。この場合、所得額が欠測に有意な補助変数であることが分かるだけでなく、行動原理に関する知見から調査対象主体の意思決定モデルまでもが導かれる。尤度法（第 2.7 節）では、こうした先験的な知見に基づいて調査客体の意思決定をモデル化することで欠測バイアスの問題に対処することができる。

1.3 欠測データ処理の限界

第 1.1.1 節で示すとおり、「欠測データメカニズム」は、MCAR（missing completely at random: 完全にランダムな欠測）、MAR（missing at random: ランダムな欠測）、MNAR（missing not at random: ランダムでない欠測）の３種類に分類される。欠測データの統計的処理においては、MCAR は例外的であり、実践的には MAR と MNAR を想定しなければならない。

欠測データの統計的処理は、「欠測データ処理の適性は、欠測データメカニズムに応じて決まるが、欠測データメカニズム自体はデータによって検証できない」という事実によって限界づけられる。具体的には、不完全データに含まれる情報だけでは、MAR と MNAR のどちらの条件が成立しているかを検証できな

いのである。第 1.1.2 節図１－２及び１－３において、適当な補助変数を用いてパネル（ロ）からパネル（ニ）を作成できれば MAR であり、作成できなければ MNAR であるが、データに含まれる情報だけでは、図１－２の MNAR ではパネル（ロ）のヒストグラムを描くことができないし、図１－３の MAR では、パネル（ハ）の情報が得られていても、パネル（ロ）ひいては（ニ）のヒストグラムを描くことができないのである。MNAR ではないという仮定の下では、MCAR と MAR を比較する検定は可能であるが、MNAR ではないという仮定自体が検証できないという限界は残る。

この限界の下で最大限可能なことは、欠測が生じるしくみに関するあらゆる事態を網羅的に想定して、それらの想定ごとに適切な分析を実行し、結果を比較することである。これは「感度分析」と呼ばれるものである。不完全データが与えられたとき、その欠測データメカニズムが先験的に明らかでない限りは、いくつかの想定を組み合わせて感度分析をすることが望ましい。