第2章　調査における欠測の分類と対応

本調査研究においては、統計データの補完推計について調査を行うため、調査において発生する欠測の分類と対応について、まず整理を行うこととした。ここでは、欠測を含むデータを特徴づける要素として「欠測が生じる状況」、「欠測パターン」及び「欠測メカニズム」について整理を行い、次に欠測を含むデータに対する対応方法として、欠測を欠測として扱う「欠測を含むデータの取扱」、及び代入を伴う「欠測値の補完」について整理を行った。

1.　欠測が生じる状況

統計調査においては何らかの理由によって、データを得ることができない場合がある。欠測を含むデータが生じる状況には、主に以下の 4 つが挙げられる。

1)　脱落（Drop out）

脱落とは、同一の対象者に対して長期にわたる調査を行う場合に、ある時点のデータが欠測する状況のことである。パネル調査等、一定の期間を置いて調査を繰り返す際に、特定の調査回以降の回答が得られない場合が該当する。

2)　無回答

無回答とは、調査全体に対する拒否（ユニット非回答）や、一部質問に対する回答忘れ・拒否（項目非回答）等によって欠測が生じる状況のことである。

3)　打切り（Censoring）

打切りとは、ある変数が決められた範囲外の値を取ることに伴って、関連するデータが欠測する状況のことである。例えば、一定の調査期間が経過したところで調査を打ち切った場合、その時点以降のデータは欠測となる。また、ある一定の所得以下の世帯については一部の質問項目を調査していない、というような場合も打切りに該当する。

4)　切断（Truncation）

切断とは、上記の打切りに類似しているが、欠測となったデータの件数も欠測している状況のことである。例えば、特定の従業員規模以下の事業所については調査対象から除外したことで、そうした事業所の数も質問項目の回答もわからない、といった場合が該当する。

上記 4 つの状況のうち、本調査研究においては、1)脱落及び 2)無回答について考察を行っている。さらに 2)無回答については、統計調査の品質に大きく影響することからも、様々な対応がなされており、以降の第 3 章・第 4 章でもそうした対応について整理を行っている。

2.　欠測パターン

欠測を含むデータについて、欠測の生じ方を示すものが欠測パターンである。大きく分けると、全ての項目についてデータが欠測する「ユニット非回答」と、一部項目についてデータが欠測する「項目非回答」の 2 つがある。欠測パターンに応じた分析を行う必要がある。

1)　ユニット非回答（Unit Non Response）

ユニット非回答とは、調査対象者の不在や拒否、転居、対象把握誤り等の理由で、全ての項目のデータが欠測するものである。全ての項目が欠測するため、代わりとなる回答者を追加サンプリングすること、事前に多めにサンプリングすることや、ウエイトの調整等の対応が取られることもある。

2)　項目非回答（Item Non Response）

項目非回答とは、記入漏れや誤記入等の理由で、一部の変数（質問）に対するデータが欠測するものである。項目非回答は、さらに欠測パターンが単調か、単調でないか、によって分類される。

(1)　単調な欠測パターン

単調な欠測パターンとは、欠測を含むデータの変数と回答者（ケース）を以下のように入れ替えられるものである。

図表　単調な欠測パターンのイメージ（○：回答、×：欠測）

回答者＼変数	①	②	③	④	⑤	⑥	…
A	○	○	○	×	○	○	
B	×	○	×	×	○	×	
C	○	○	○	○	○	○	
D	×	○	×	×	○	×	
E	○	○	○	○	○	○	
F	×	×	×	×	○	×	
⋮							

変数の並び順を入れ替え
回答者の並び順を入れ替え

回答者＼変数	⑤	②	①	⑥	③	④	…
C	○	○	○	○	○	○	
E	○	○	○	○	○	○	
A	○	○	○	○	○	×	
B	○	○	×	×	×	×	
D	○	○	×	×	×	×	
F	○	×	×	×	×	×	
⋮							

上記のようにデータを表現したものは「データ行列」と呼ばれている。単調な欠測パターンとは、データ行列において、「ある回答者の変数が回答されていれば、上図のデータ行列の左あるいは上に位置する変数も回答されている状態」と捉えることが可能である。このパターンの場合、欠測値補完の計算が容易になる。

(2)　単調でない欠測パターン

単調でない欠測パターンとは、上記の単調な欠測パターンが成立しない状態である。このパターンの場合、欠測値補完にあたって反復計算が必要となり計算が煩雑になる。

3.　欠測メカニズム

欠測メカニズムとは、どのように欠測が発生しているのか、を示すものである。欠測値を含むデータを分析する際にバイアスが生じるか、また、どのようなバイアスが生じるかについて、欠測メカニズムを理解することが重要である。ここでは、欠測について検討する際に頻繁に用いられている 3 つの欠測メカニズム[1]を取り上げる。

1)　MCAR（Missing Completely At Random: 完全にランダムな欠測）

MCAR とは、ある値が欠測する確率が、その変数を含め他の変数に依存していない状態である。ある質問に対する回答を忘れる、データ集計の際の手違いによってデータの一部がランダムに欠測する、といった場合に MCAR となる。

項目非回答者と項目回答者とが類似している（欠測の有無によらずに両者が類似している）、という意味において、MCAR は最も単純な欠測メカニズムである。しかしながら、現実の調査において、MCAR が成立することは少ないと考えられる。

2)　MAR（Missing At Random: ランダムな欠測）

MAR とは、ある値が欠測する確率が、その変数以外の他の変数に依存している状態である。例えば、以下のような場合に MAR となる。

（例1）
- それぞれの年代層（例えば高年代層と若年層）の中ではある値が欠測する確率が、その変数を含め他の変数に依存していない MCAR の状態にある
- しかし、欠測メカニズムが高年代層と若年層の回答者の間では異なっている

（例2）
- 大規模企業及び小規模企業のグループの中では、MCAR の状態にある
- しかし、大規模企業と小規模企業のグループ間では異なる欠測メカニズムとなっている

MAR は MCAR に比べてより複雑な状態であるが、MCAR となる適切なグループ分けを行うことができれば、そのグループ内では MCAR を前提として欠測値を補完することも可能である。実務上では、欠測メカニズムは MAR と仮定されることが多い。

3)　NMAR （Not Missing At Random: ランダムでない欠測）

NMAR とは、ある値が欠測する確率が、その変数に依存している状態である。収入が高い人ほど収入を回答しない、売上高が欠測するか否かが売上高の多少に依存している、といった場合に NMAR となる。

NMAR は最も複雑な欠測メカニズムであり、NMAR の状態にあるデータにおいては、観測された値のみでは補完を行うことができないため、欠測している変数についての欠測メカニズムに何らかの仮定を置かなければならない。

[1] Rubin, D.B. (1987), Multiple Imputations for Non-Response in Surveys, John Wiley & Sons, New York.

（参考）Ignorable（無視可能）と Nonignorable（無視不可能）

MCAR、MAR、NMAR と類似した、欠測メカニズムの分類に「Ignorable（無視可能）」と「Nonignorable（無視不可能）」がある。この分類は、ある値が欠測する確率がその変数に依存するか否かに基づいたものである。

①Ignorable（無視可能）

Ignorable（無視可能）とは、欠測メカニズムが MCAR あるいは MAR の状態を指す。この分類では、欠測する確率がその変数には依存しないため、適切な手法を用いることで対応が可能であることから、「無視可能」と呼称されている。

②Nonignorable（無視不可能）

Nonignorable（無視不可能）とは、欠測メカニズムが NMAR の状態を指す。この分類では、欠測が生じていることによる影響を分析した上で欠測への対応を行うことが求められる。

4.　欠測を含むデータの取扱

欠測値を含むデータを分析する際に当たっての考え方について整理を行う。大きく分けると、代入等の補完を行わずに、欠測を欠測のままとして扱う考え方と、欠測値に値を代入することで完全なデータとして扱う考え方の 2 つに分かれる。ここでは、まず前者の考え方について整理を行う。

1)　完全ケースに基づく分析（Complete Case Analysis）

完全ケースに基づく分析とは、一部の変数でも欠測を含むケース（回答者）は全て取り除いて分析を行うことである。最も簡単な方法ではあるものの、欠測を含むケースが多い場合には適さない、欠測メカニズムが MCAR 以外であると結果に影響が生じる、といった欠点がある。

2)　利用可能なケースに基づく分析（Available Case Analysis）

利用可能なケースに基づく分析とは、ある変数について分析する場合に、その変数が得られているケース全てを対象として分析を行うことである。欠測を欠測のままとして扱う考え方としては、利用できる情報は可能な限り用いることができるものの、完全ケースに基づく分析と同様に、欠測メカニズムが MCAR 以外であると結果に影響が生じる。

5.　欠測値の補完

次に欠測値を含むデータについて、その欠測値を補完、特に代入を行うことで対応する手法について概観する。

データには「量的データ」（間隔尺度、比率尺度）、「質的データ」（名義尺度、順序尺度）といった種類や、複数年データ、パネルデータ等の時系列データといった調査形態による差異も存在する。通常、欠測値の補完では、上記のようなデータ種類や調査形態に応じた欠測値補完がなされるため、複数の補完手法を用いることも多いが、補完手法における概念は下記の 2 つに大きく分けることが可能である。

1 つ目は、欠測値を 1 つの値で代入する”Single Imputation”（単一代入法）である。これは、データ内の欠測値を 1 つの値で代入し、完全データセットとすることを目指すものである。

2 つ目は、複数の値を代入する”Multiple Imputation”（多重代入法）である。こちらは、1 つの欠測値に対して複数の値をシミュレーションするものである。

一般的に、単一代入法では推定値の分散を過小評価する傾向が強く、全体分布が歪められてしまう可能性があることが指摘されるが、多重代入法では複数の完全データセットを作成することでその課題への対応が図られている。

まず、単一代入法において用いられる具体的な補完手法について概観する。

1)　単一代入法

(1)　平均値による補完

回答を得られているケースの平均値を、欠測値の部分に代入するもの。平均値には通常、算術平均が用いられるが、中央値や最頻値を用いることも可能である。年収や売上高、消費金額等の数量値に対して行われることが多い。

長所:	－ 補完作業が容易である
	－ 集計値（平均値）に影響を及ぼさない
短所:	－ 欠測値を含むケースが多い場合には不適切
	－ 回答者の属性が反映されない

平均値を用いた補完における長所として、補完作業が容易という点が挙げられる。算術平均値の算出は、作業自体も容易で、再現性も備えている。加えて、平均値算出のベースが全体なのか、特定のセグメント（20 代や 30 代といった年齢セグメント、関東地方や近畿地方といった地域セグメント等）なのか、といった条件に影響されるものの、集計値として広く用いられる平均値の集計には影響を及ぼさない、という点も長所として挙げられる。

一方で、短所としては、欠測値を含むケースが多く、平均値の算出が少数の回答ベースとなることで、平均値自体の信頼性が劣る可能性が存在する点が挙げられる。加えて、補完しようとする欠測値が回答者の属性に依存する場合（例えば、年収は回答者の就業形態や業種に影響されることが一般的である）には、そうした属性の影響を加味しない形で補完が行われるということがある。

なお、極端に小さい又は大きい値（外れ値）の影響を除くために、これらを除外した平均を使用する場合もある（刈り込み平均）。

(2) Hot Deck

回答を得られているケースから、背景データが類似したケース(ドナー)を探し出し、ドナーの値を欠測値の代わりとして代入するもの。数量値、カテゴリカルデータのいずれにおいても適用することが可能である。

長所:	− 回答者の属性等を反映することができる
	− 欠測値の発生要因に依存しない
短所:	− 補完作業が煩雑となる
	− 欠測値の生じるケースが多い場合には不適切

Hot Deck 法の長所としては、属性等が類似したケースを探し出すため、就業形態に応じた年収が代入値として得られる等、属性を反映させた結果となる(納得感のある結果となる)ことが挙げられる。加えて、欠測値の発生要因が MCAR(Missing Completely At Random:欠測は完全にランダムに発生)、MAR(Missing At Random:欠測はランダムに発生)、NMAR(Not Missing At Random:欠測はランダムでなく発生)のいずれの場合であっても適用可能である。

一方で、類似したケースを探し出すための作業が必要となるため、補完作業は煩雑となってしまう。特に、どのような定義をもって「類似している」と判断するかを慎重に決定する必要があるため、試行錯誤の繰り返しが発生することも考えられる。加えて、類似ケースを探し出すためのベースとなる「欠測値を含まないケース」が少ない場合には、結果に対する信頼性も劣る可能性がある。

なお、回答者の属性等を反映することができる Hot Deck 法の概念を基本として、類似したケースの平均値を代入する等、前述の手法との組合せも可能である。

(3) Cold Deck

背景データが類似したケース(ドナー)を他のデータセットから探し出し、そのケースの値を欠測値の部分に代入するもの。前述の Hot Deck では同一データセット内から類似ケースを探し出したが、Cold Deck では他のデータセットから探し出す点が異なる。例えば、企業売上高を調査ではなく、財務諸表等の情報から補完することも Cold Deck の 1 つと見なせるであろう。

長所:	− 回答者の属性等を反映することができる
短所:	− 補完作業が煩雑となる
	− 他のデータセットの利用可能性に制約される

Cold Deck 法では、Hot Deck 法同様に、属性等が類似したケースを探し出すため、属性を反映させた結果を得られることが多いことが長所として挙げられる。

一方で、類似したケースを特定するために、補完作業は煩雑となってしまう。また、他のデータセットが利用できない場合には Cold Deck 自体を行うことができない。

(4)　回答データの据置き（横置き）による補完

据置き補完（横置き補完）（Carry Forward）とは、特定のケースについて、過去の回答結果をそのまま代入するもの。同一客体を継続的に追跡する調査（企業統計調査、パネル調査等の時系列データ）における適用が想定されるものである。

長所：	－ 補完作業が比較的容易である
短所：	－ 外部環境変化（景気動向等）やケース属性の変化（従業員数増減等）が反映されない
	－ 長期にわたって回答していないケースには適用できない

同一客体を継続的に追跡する調査においては、据置き補完（横置き補完）の作業は比較的容易である点が長所として挙げられる。

短所としては、景気動向による影響や従業員数増減等、ケース属性の変化が反映されない点が挙げられる。特に、季節変動が存在する場合やリーマンショックのような大きな変化が生じた場合には、その影響を加味することができない。また、少なくとも前回調査における回答は必須であり、前回からさらに遡って回答が欠測する場合には適用できない（代入するもとのデータが古いデータとなる）点も短所として挙げられる。

なお、据置き補完（横置き補完）を行う場合には、今回調査と前回調査において同一の質問項目であることが前提となり、新たに追加した質問に対しては、据置き補完（横置き補完）を適用することはできない。

(5)　伸び幅又は伸び率を用いた補完

回答を得られているケースを基に、前期からの変動幅又は変動率（伸び率）を算出し、特定のケースの過去の回答結果にその値を乗じた値を代入するもの。

長所：	－ 外部環境変化（景気動向等）を反映することができる
	－ 補完作業が比較的容易である
短所：	－ ケース属性の変化（従業員数増減等）が反映されない
	－ 長期にわたって回答していないケースには適用できない

伸び幅又は伸び率を用いた補完の長所としては、回答を得られているケースからの情報に制約されるものの、景気動向等の外部環境変化を一定程度反映することが可能な点である。

一方で、短所としては代入する値が必ずしもケース属性の変化を反映していない点が挙げられる。具体的には、年数が経過するに従って従業員数が増加したにも関わらず、その影響を加味しない形で代入が行われる例が考えられる。また、長期にわたって回答がなされていないケースでは、元となるデータが存在しないあるいは古いものであるために、信頼性が劣る危険性も短所として挙げられる。さらに、伸び率を用いる場合には、前期が0であると計算できないということがある。

(6)　重回帰式モデルによる補完

回答を得られているケースを基に、複数の要因を加味した重回帰式モデルを作成し、欠測値を予測するもの。

長所:	- ケース属性や季節変動等複数の要因を加味することができる
	- 決定係数、調整済み決定係数等によるモデルの当てはまりを評価できる
短所:	- 重回帰式に投入する変数の選定が困難
	- 再現性が低い
	- 分散を過小評価する

長所としては、重回帰式には複数の変数を投入することになるため、複数の要因を加味することができる点が挙げられる。また、作成された重回帰式モデルについても、決定係数や調整済み決定係数を基に、当てはまりを評価することができる点も長所として挙げられる。

一方で短所としては、重回帰式に投入する変数の選定が困難である点が挙げられる。また、多重共線性の問題や係数の符号(正負)の整合性を考慮しなければならない点も短所として挙げられる。こうした点は、いったん重回帰式モデルを作成しても、そのモデルが次回調査には当てはまらない可能性がある、という再現性の低下につながる。同時に、都度こうしたモデルを作成する必要が生じる等、補完作業が極めて煩雑となる。また、重回帰式モデルの特性上、分散を過小評価する危険性も含んでいる。

2)　多重代入法

多重代入法では、1 つの欠測値に対して複数個の値を代入するが、一般的にそのステップは以下の 3 ステップである。

ステップ①:欠測値に対して複数個の値を代入し、複数個の完全データセットを作成する
ステップ②:複数個の完全データセットのそれぞれについて、集計・分析を行う
ステップ③:ステップ②で得られた複数個の結果を統合する

多重代入法の基本的な考え方は、Rubin によって 1980 年代に提唱されたものである。複数の完全データセットを作成する際には、例えば重回帰式モデルによる補完結果に、ランダムに発生させた誤差を加えた結果を代入する等、単一代入法によるデータセットを複数個作成するものが例として挙げられる。

以下では、多重代入法の概念に沿った主な手法について概観する。多重代入法の長所としては、複数のデータセットを作成することで、欠測値が生じることの不確実性を考慮し、分散の過少推計を避けることができる点が挙げられる。短所としては、変数の種類(尺度変数、名義変数等)や結果分析の目的に応じて適切な多重代入を行わないと結果にバイアス(標準誤差の過少評価)が生じることがある点が挙げられる。

(1)　EM アルゴリズムによる補完

EM アルゴリズムにより、最尤推定値を予測するもの。欠測値を含むデータの尤度を最大化することで、“もっともらしい”結果を得ようとするものである。E ステップ（estimation：期待値ステップ）と M ステップ（maximization：最大化ステップ）から構成される。

E ステップ　：尤度関数の期待値を求める
M ステップ　：E ステップにおける尤度の期待値の最大化を図る

(2)　マルコフ連鎖モンテカルロ法

マルコフチェーン・モンテカルロ法（MCMC）とも呼ばれ、I ステップ（Imputation:代入ステップ）と P ステップ（posterior：事後ステップ）の 2 ステップから構成される。それぞれのステップを繰り返し行うことで、代入値を算出するもの。

I ステップ　：欠測値を含むデータセットに対する条件付き分布に基づき、完全データセットを作成
P ステップ　：得られた完全データセットから母集団情報を推定

（参考）傾向スコア

類似したケース（ドナー）を探し出す際の基準のひとつとして用いることができるものに、各ケースの傾向スコア（propensity score）がある。傾向スコアとは各ケースが回答する確率として表現されるものである。傾向スコアが同一のケースをドナーとして使用する等、ドナー選定の基準に用いられることがある。

基本的な考え方は、事後確率を求めるベイズ統計に沿ったものといえる。

（参考）ユニット非回答に対応する手法　キャリブレーション（Calibration）

キャリブレーション（Calibration）とは、回収標本から母集団を推定する際に、母集団の値を補助変量として用いることでウエイトを調整する手法である。ウエイトを調整するものであるため、項目非回答ではなくユニット非回答に対して用いられる手法である。「日本人の国民性調査」（実施：統計数理研究所）における調査不能バイアスの補正に関する研究において用いられた（本報告書第3章）。

回収標本から母集団を推定する際のウエイト調整には、以下の 2 段階のプロセスがある。

- 回収標本から抽出標本の推定
- 抽出標本から母集団の推定

前者には回答確率（回収率、傾向スコア、等）の逆数を用いた推定（IPW 推定）が用いられ、後者にはキャリブレーションが使用される場合が多い。回答確率の逆数を用いた推定においては、回答確率の低い回答者のウエイトが大きくなるため、推定量は不偏であっても分散が過大評価される、という欠点がある。

キャリブレーション先（母集団）の情報があらかじめ分かっていなければならない（「ある趣味をもった人に対する調査」のように、母集団に関する情報が未知の場合には使用できない）といった限界もあるものの、キャリブレーションは定まった母集団の値を使用することで、その分散を抑えることが可能である。

6.　調査における欠測への対応

調査における欠測については、海外（アメリカ・カナダ）においては主にセンサス（国勢調査）における非回答に関して、どのように扱うのか、という議論が行われてきた。カナダ統計局においては、1950 年のセンサスまでは、欠測値に対してデミング法（Deming’s method）と呼ばれる方法で補完を行っていた。これは、前回実施のセンサスにおける頻度分布を用いて、手作業で欠測値に代入を行うものである。さらに 1953 年には、アメリカセンサス局の統計家Morris H. Hansenらによって、1948年に実施された小売シェア調査（Survey of Retail Shares）の欠測の扱いについての議論が出版[2]された。先月の売上高・賃金等の情報に基づいて「補完（代入）」を行う、と統計調査における「補完」に初めて言及したものとされている。

その後、アメリカにおけるセンサスでは1960 年に実施されたものが、カナダにおけるセンサスでは1961 年に実施されたものが、コンピューターを利用した補完が行われた初めての調査とされている。補完に関する技術の進歩は、コンピューターの進歩とともにあったが、初期においてはホット・デック法が主な補完手法であった。

その後、様々なプログラムがアメリカセンサス局やカナダ統計局によって開発されてきた。

アメリカセンサス局

- SPEER（Structured Programs for Economic Editing and Referrals）
  1984 年に開発。経済データにおける比率の補完に用いられる。

カナダ統計局

- CANEDIT
  1976 年に開発。Fell センサスに利用される。
- GEIS（Generalized Edit and Imputation System）
  ※本報告書4章にて GEIS の後継プログラム Banff について記載
- CANCEIS/NIM
  （Canadian Census Edit & Imputation System / Nearest-neighbor Imputation Methodology）
  最近隣法によるホット・デック法を行う。

本調査研究において調査を行った海外事例（アメリカ・カナダ）では、上記のような経緯から、ホット・デック法が多く用いられている。多くの補完手法が利用可能となった現在においても、ホット・デック法が主要な補完手法として用いられ続けている。

[2] Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory*, Vol. I and II. New York: John Wiley and Sons.