歯科疫学統計
－第9報　欠損値・異常値処理法－
－不完全データをどう扱うか－

瀧　口　　徹

A review of oral epidemiological statistics
－ Part IX ： A comparison of imputation techniques for handling missing values and abnormal values －

Toru Takiguchi

キーワード：欠損値（異常値）、リストワイズ除去法、ペアワイズ除去法、平均値代入法、多重代入法、欠完全情報最尤推定法（完全報知最尤法）

はじめに

長年データ処理に携わってきたお陰で今回本テーマを扱うこととなった。このテーマは統計学の演習において受講者や学生にとっては決して興味あるものではないように見受けられる。難しい検定や分析にチャレンジし結果が出たときは喜びがあるが、データ値に問題があるとは、いわばマイナスからのスタートでやる気を殺がれてしまうのかもしれない。それ以上に、社会通念上の感覚からして欠損値を含むデータを除去するならまだしも、実在しない値で補完するということは何かデータを改竄（かいざん）するような、いままでの学校や社会の教えに背くような後ろめたい気持ちになるのかもしれない。気持ちを入れ替えてここを乗り越えたとして、統計学者が推奨する欠損値や異常値の高度な処理法は難解でありビギナーは入口で完全に立ち往生しまうだろう。そこで本稿ではデータを不完全なまま分析することの統計的な問題点を説明し、その上で伝統的および最新の処理法の概念を掴んでいただくことに主眼を置いた。

【著者連絡先】
〒341-0003　埼玉県三郷市彦成3-86
深井保健科学研究所
主席研究員　瀧口　徹
TEL&FAX：048-957-3315
E-mail：taki8020@mth.biglobe.ne.jp
受理日：2012年11月30日

1．事例のプロフィール

伝統的な欠損値処理、異常値処理の例題として感覚的に掴みやすい身長、体重を事例の指標とし、対象を過食による肥満が問題になっている南太平洋の某島国のデータの一部を抽出し加工して用いた。

2．解説および共通事例による解析

下記の手順で行う。

＜解説＞
（1）欠損値処理、異常値処理を行う前に
（2）欠損値処理の長短と改竄（かいざん）との違い
（3）欠損値のメカニズム
（4）欠損値・異常値処理法の一覧と特徴比較
（5）除去法（非代入法）
　①ペアワイズ法、②リストワイズ法
（6）単一値代入法
　①平均値代入法、②Hot－deck法
　③最悪値代入法、④前回観測値代入法
　⑤回帰式推定法、⑥Cold－deck法

（7）多重代入法（MI）
（8）完全情報最尤推定法（FIML）
（9）MI法とFIML法の優位性および優劣について
（10）異常値処理法1
　①箱ひげ図による方法
　②Smirnov-Grubbs検定による方法
（11）異常値処理法2
　①人口動態調査等におけるベイズ推定法
　②地域集積性分析における扱い

（1）欠損値処理、異常値処理を行う前に

ａ）欠損値の発生原因は不可避か？

フィールド調査におけるアンケートなどの欠損値は、”知らない：Don't know”、“答えたくない：Refused,”、“理解できない；Unintelligible,”という3つの理由から主として成っている。Joseph Lら[1]は心理学の調査を例に上げ、これらの反応のどこに真実があるかを欠損値処理の前に（調査に先立って）考慮すべきとしている。その例としてアンケートで「あなたはマリファナを吸ったことがありますか？」と聞いた場合をあげている。この場合の答えは上記3つの理由による欠損値のみならず虚偽の答えを誘発するであろうことは想像に難くない。未成年の喫煙歴を調査する際も同様であるが、真実を知りたいと思ったらまず回答用紙を氏名未記入の封筒に入れ糊付けする、それをさらに別の大きな無記名の封筒に例えばクラス全員分入れて郵送／提出する等の個人を特定できないようにする調査を行わないと欠損値は増加するのみならず、データ値そのものの信頼性が著しく低下する。このように統計的な欠損値処理に入る前に「真実を回答できる」質問と環境を作り欠損値の発生を出来るだけ減らす必要がある。

ｂ）異常値の発生原因は不可避か？

同じく、異常値（但し、欠損値を異常値として扱わない場合）処理に入る前に（調査に先立って）、その発生原因と抑制を考える必要がある。異常値発生原因のひとつは①虚偽記載、である。例えば年収を何らかの意図で過小、もしくは過大に記載する場合である。そうした意図的で無い場合で②勘違いによる記載、がある。例えば収入を月収で30万円と答えるべきところを年収で500万円と答えた場合、新生児体重を3200（g）と答えるべきところ単位をkgと間違えて3.2gと答えた場合等である。また回答者ではなく入力者側のミスで生じる場合は初期値として設定してあった999とか0とかがそのままデータとして使われてしまった場合、入力時に体重を75といれるべきところを775と入れてしまった場合等がある。これら異常値を扱う際に特に注意すべきは欠損値と異なり、例えば体重が180kgの場合とか「真実の値」である可能性を排除しないこと、その逆に体重660kgと記載されたものを安易に66kgとか60kgに修正してしまわないことである。データ収集時に例えば体重が180kgについて特記事項にその旨を記載しておけばデータ値を異常値扱いされないので完璧である。後述する手法等を用いて異常値と判断したらデータが存在しないことと同じで欠損値と同義となるので、異常値を欠損値として扱う必要が出る。

（2）欠損値処理の長短と改竄（かいざん）との違い

Scheffer J[2]は欠損値処理を行うことの長所として、得られた貴重なデータを有効活用し、データの偏り（バイアス）を縮小できる。もしそうならないのであればデータを捨てなければならないとしている。一方、短所としては選択した欠損値代入法の影響を受け、ある偏りを持ち込んできており、代入したデータ値は真のデータではないので分散は不確実性を反映したものにならなければならない。しかし、特に単一代入法は分散を減少させ代入による不確実性の増大を反映できないのが問題だとしている。

新井ら[3]によれば,欠損値処理とデータの改竄との違いは、改竄は欠損値や実際値を「解釈するのに都合の良いデータで置き換えること」であり、一方、欠損値処理は「統計学的な理論に基づいた適切な置換」と説明している。このように両者の

差は明らかである。科学業績の分野において起きうる不正行為として改竄とデータ値の捏造（ねつぞう）があるが、これらの行為によって何らかの被害が出た場合は当然のことながら犯罪が成立し研究者としての社会的、道義的責任が厳しく問われることになるのは内外に例を待たない。

(3) 欠損値のメカニズム

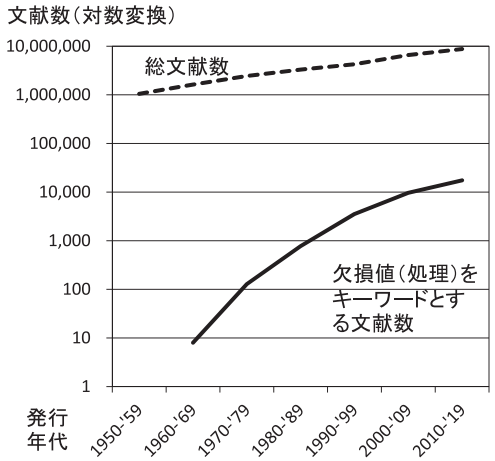
Rubin DB[4, 5, 6, Web01] が1976～2002年に掛けて共同研究者と共に、いわゆる欠損値メカニズムとその処理法を確立して以後、欠損値処理法は俄に脚光をあびコンピュータソフトの開発と相俟って世界的に多用されるに至っている。それを裏打ちする指標として関連論文の時代変遷をPubMedで確認し図1に示した。後述する従来の欠損値処理法とRubinらが開発した新しい欠損値処理法の優位性の比較研究を核として関連論文数は1960年以降PubMed掲載論文数の増加数が10倍なのに対し、欠損値処理をキーワードとする論文は100倍増加している。欠損値メカニズムとは図2に示すMCAR : missing completely at random, MAR : missing at random, およびMNAR : missing not at randomの3タイプを指している。図において、

Y：欠損値のある変数

R：Yの欠損値の有無を示す確率変数（例えば有：1、無：0）

X：分析しているデータに含まれるその他の変数

を示している。ここで上記の3分類の違いをわかりやすく説明するため著者自身が経験した事例を示す。著者がJICA専門家としてスリランカに2年間赴任していたことを契機に同国と共同研究を継続してきているが、老人ホームにおける入所者のADL：日常生活動作、と味覚との関連を調査した際、ADLが低くベットから立ち上がることに支障を来す入所者に関しては体重計測値が欠損し易いことが判明した。ベットに仰臥していても身長は何とか計測可能であるが体重は計測者が背負って体重計に乗って後で計測者の体重を引く手間がかかる等が理由と考えられた。ここで図１との対応では体重がY、RはYの計測が欠損になる場合が１、計測が可能な場合が0となる確率変数である。またADLがXで示される。本例ではADL（X）が低くなると体重計測値が欠損し易いのだからRはXに依存傾向がある。すなわち、欠損の３タイプのうちMAR : Missing At Random



注1）検索語：(1) imputation techniques for handling missing data, (2) imputation of missing data,(3) missing data (values), (4) full information maximum likelihood method, (5) multiple imputation method

注2）2010-2019の数値は3年間：2010/01/01-2012/12/31の実績（文献総数：2,629,336件、欠損値をキーワードとする文献：5,234件）を比例配分した予測値

注3）縦軸の文献数は対数軸のため増加比率が等価になっている

図1　欠損値（処理）を扱った文献数の推移（PubMed収載）

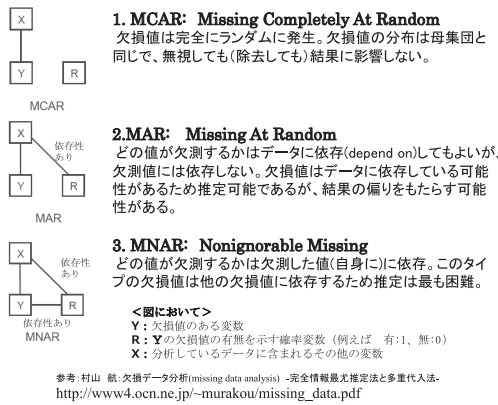


図2　欠測発生のメカニズム（欠損の３タイプ）

ということになる。このタイプの場合MCAR：Missing Completely At Randomの場合と異なり欠損値を含むデータレコードを除去する（欠損値を含む被験者を非対象とする）手段だと例数が減るディメリットはそう問題にならないがADLが高い対象者がデータから欠落し易い傾向が内在した分析結果となる。従って何らかの方法でデータ値を補完する必要があることになる。

また、体重を自記式でデータ化する場合、肥満傾向の強い被験者と反対に痩せ傾向の強い被験者が体重の自記式を嫌うような傾向が顕著な場合は欠損値が存在するYそのものの値にRが依存することからMARではなくMNAR：Nonignorable Missingとなる。

(4) 欠損値・異常値処理法の一覧と特徴比較

表1に15分類した欠損値・異常値処理法の一覧と特徴を示し、以下解説する。

(5) 除去法（非代入法）

①ペアワイズ法：重回帰分析等において説明変数の全ての組み合わせ（ペア）について相関を求める。その際、ペアとした説明変数の値がどちらか一方でも欠損している場合は除去する。

②リストワイズ法：欠損値を含むレコードを削除する方法。

SPSSでは①、②および下記の(6)の①の平均値代入法は標準装備されており分析に先立って欠損値処理の条件をダイアログボックスで選択できるようになっている。

(6) 単一値代入法

表2および図3に下記①平均値代入法、⑤単回帰式推定法の実例を示した。表2－aにおいて灰

表1　欠損値・異常値処理法の一覧と特徴

No.	方法名 和	方法名 英	英略語[a]	代入法 (IMP: imputation)	欠損値処理法	採用上の注意点	統計ソフト (+ package) 例
	＜狭義の欠損値処理＞						
1	欠損値の放置	default in handling of missing data			欠損を放置（そのままモデル化）		
2	ペアワイズ除去法	Pairwise(Pair-wise) deletion	PD		重回帰分析等において説明変数の全ての組み合わせ（ペア）について相関を求める。その際、ペアとした説明変数の値がどちらか一方でも欠損している場合は除去する		spssでは多変量解析に際していずれかを選択
3	リストワイズ除去法	Listwise(List-wise) deletion (Complete-case analysis)	LD		欠損値を含むレコードを削除	selection biasの恐れ	
4	平均値代入法	Personal mean score	PMS	single IMP	欠損値以外の数値の平均値を代入	データの変動を過小評価する恐れ	
5	最悪値代入法	Personal worst score	PWS		欠損値以外の数値の「最悪値[b]」を代入	データの変動を過大評価する恐れ	
6	単回帰式推定法	Personal regression score	PRS		欠損値以外の数値からの回帰式[c]による推定値を代入		
7	近似値代入法	Hot-deck imputation	HD	single IMP	欠損値を含む人と属性の近似している人の値を代入		
8	重回帰式推定法	Cold-deck imputation	CD	single IMP	重回帰式などで欠損値を推定して代入		
9	前回観測値代入法	Last observation (value) carried forward	LOCF (LVCF)	single IMP	前回の観測値を代入		
10	多重代入法	Multiple imputation	MI	multiple IMP	多重代入法によりある一つの欠測値に対して複数回の補完を行う		spss21 R(mi, mice) stata11
11	完全情報最尤推定法（完全報知最尤法）	Full information maximum likelihood	FIML		異常値を除いた全てのデータ値を用いて最尤法の手法を用いて下記の2つの手法で推定値を求める。①CFA:確証（認）的因子分析（confirmatory factor analysis）②SEM:構造方程式モデル（structural equation models)=共分散構造分析		spss(Amos) R(lavaan,sem)
	＜特殊な欠損値処理＞ 異常値除去後の処理を含む						
12	外れ値（異常値）の欠損値扱い –箱ひげ図による	The abnormal data were excluded by box-(whisker-)plot			箱ひげ図で検出した外れ値（異常値）のうち補正適用外値を欠損値扱いする手法	予め棄却限界値を設定	エクセル統計 spss stata
13	外れ値（異常値）の欠損値扱い– Smirnov–Grubbs法による	handling of outliers (high or low grade data) as missing values by Smirnov-Grubbs test			Smirnov–Grubbs法で検出した外れ値（異常値）のうち欠補正適用外値損値扱いする手法。 箱ひげ図との違いは有意性検定値が判定指標となるが異常値がなくなるまで繰り返す必要がある。	予め棄却限界値を設定	エクセル統計 spss stata
14	人口動態統計におけるベイズ推定法を用いた異常値補正法	Bayesian estimation of indices related to municipalites' vital statististics			小規模町村で特に出生数や死亡数が少ない場合には、観測年で数値が大幅に上下し指標の信頼性を著しく低下させる。これを回避するため周囲の自治体情報を加味して推定値を求めるベイズ推定法を用いる		New R Package for BEST (Bayesian ESTimation supersedes the t
15	地域集積性データの欠損値の扱い	handling of missing values in regional clustering			地域集積性を検出するデータのうち欠損値を湖や内海扱いする手法。13のように周囲の情報を加味して近似させると地域集積性が人為的に高まることを回避する目的	補正により人為的に地域集積性を高める（低める）ことを回避する目的	OpenGeoda [d]

a) 英略語は各研究報告（論文）,参考書で一定でない。
b)「最悪値」: 研究仮説からみて最悪の想定値を代入。 例）検証したい仮説を否定する方向の最大値、または最小値を代入
c) 回帰式：予測したい値を従属変数とする単回帰による。例）性別に年齢と身長を独立変数とし体重を従属変数とする単回帰式を利用
d) 地域集積性の指標であるMoran's I, LISAを計算する米国GeoDaセンター提供のフリーソフ　https://geodacenter.asu.edu/

色の矩形の部分の７名は要介護度が高いためベットから立たせることが出来ず、体重の計測ができなかったと仮定した。その結果、表２－ｂに示すように欠損値の影響で男女とも平均値は増加し、標準偏差は男性が縮小した。このように欠損値がランダムに発生しない場合（図１のMCARで無い場合）は平均値、標準偏差等に無視できない偏りが生じる。これを単一値代入法で補完する場合どのようになるであろうか。

図３においてf1は欠損値発生前、f2は欠損対象７名の体重を０とした場合、f3は下記の①平均値代入法を適用した例、f4は⑤単回帰式推定法を適用した例である。

表２－ａ　欠損値が他の変数の影響を受ける場合の処理法

ID	性別	年齢	間食回数	要介護度	身長	体重	BMI
1	1	75	0	5	170	**45**	**15.6**
2	1	85	0	5	166	**55**	**20.0**
3	2	68	0	5	160	**45**	**17.6**
4	2	82	1	5	153	**33**	**14.1**
5	1	70	1	4	177	**65**	**20.7**
6	2	72	2	4	155	**60**	**25.0**
7	2	71	1	4	157	**45**	**18.3**
8	2	66	1	4	159	44	17.4
9	1	68	3	3	165	85	31.2
10	1	66	1	3	178	77	24.3
11	2	70	1	3	160	50	19.5
12	1	76	0	2	172	49	16.6
13	2	76	0	2	174	55	18.2
14	2	71	1	2	170	60	20.8
15	2	69	1	2	160	70	27.3
16	1	73	2	1	169	69	24.2
17	1	72	3	1	173	88	29.4
18	2	72	4	1	162	95	36.2
19	2	65	5	1	152	79	34.2
20	1	76	2	0	175	77	25.1
21	2	87	2	0	159	55	21.8
22	1	65	2	0	174	67	22.1
23	1	78	3	0	173	78	26.1
24	2	68	3	0	152	55	23.8
25	2	80	0	0	145	40	19.0

表２－ｂ　欠損の有無別体重の値

			欠損値 無し	欠損値 有り
男	n	人数	11	8
	mean	平均値	68.64	74.13
	sd	標準偏差	14.20	12.68
女	n	人数	14	10
	mean	平均値	56.14	60.30
	sd	標準偏差	16.38	16.72

注1）表1-aにおいて灰色の矩形の部分の7名は要介護度が高いためベットから立たせることが出来ず、体重の計測ができなかったと仮定
注2）表2-bは欠損値の影響で男女とも平均値は増加し、男性の標準偏差は縮小

単一値代入法は解説書等[7, Web02)]を総括すると下記６種類に分類される。

①平均値代入法：欠損値以外の数値の平均値を代入する方法である。f3がその例示であるがこの方法はサンプル平均を変化させないが図でわかるとおりサンプルの分散を減少させてしまうことがわかる。分散が縮小し、欠損値を置き換えたため例数（n）が減少しないため、有意差検定等では有意と成りやすくなることが問題である。

②Hot-deck法：無回答者と属性の似た他の回答者の回答を代用する

③最悪値代入法：「最悪値」：研究仮説からみて最悪の想定値を代入。例えば、検証したい仮説を否定する方向の最大値、または最小値を代入

④前回観測値代入法：前回の観測値を代入

⑤単回帰式推定法：回帰式：予測したい値を従属変数とする単回帰による。例えば、性別に年齢と身長を独立変数とし体重を従属変数とする単回帰式を利用する方法で、f4にその例を示す。（6）の①と異なり代入する体重値の妥当性が増加する利点があるが、残った値でみた相関係数が0.25と低いことから実際値はこの回帰直線の前後に相当程度ばらつくと考えられるので、その不確実性が加味されていないことは前述のScheffer J[2)]の指摘のとおりである。

⑥Cold-deck：無回答者の属性から重回帰式等多変量回帰式により推計する方法であり、単回帰式より相関係数の絶対値が高くなるので不確実性問題が改善される利点がある。

（7）多重代入法（MI）

シミュレーション研究でも上記（1）～（6）の単一値代入法の場合推定量のバラツキを過小評価[Web03)]する可能性があることが知られている。Rubin DB[4, 5, 6, Web02)]が1970年代に開発した多重代入法（MI）は欠損値を代入したデータセットを

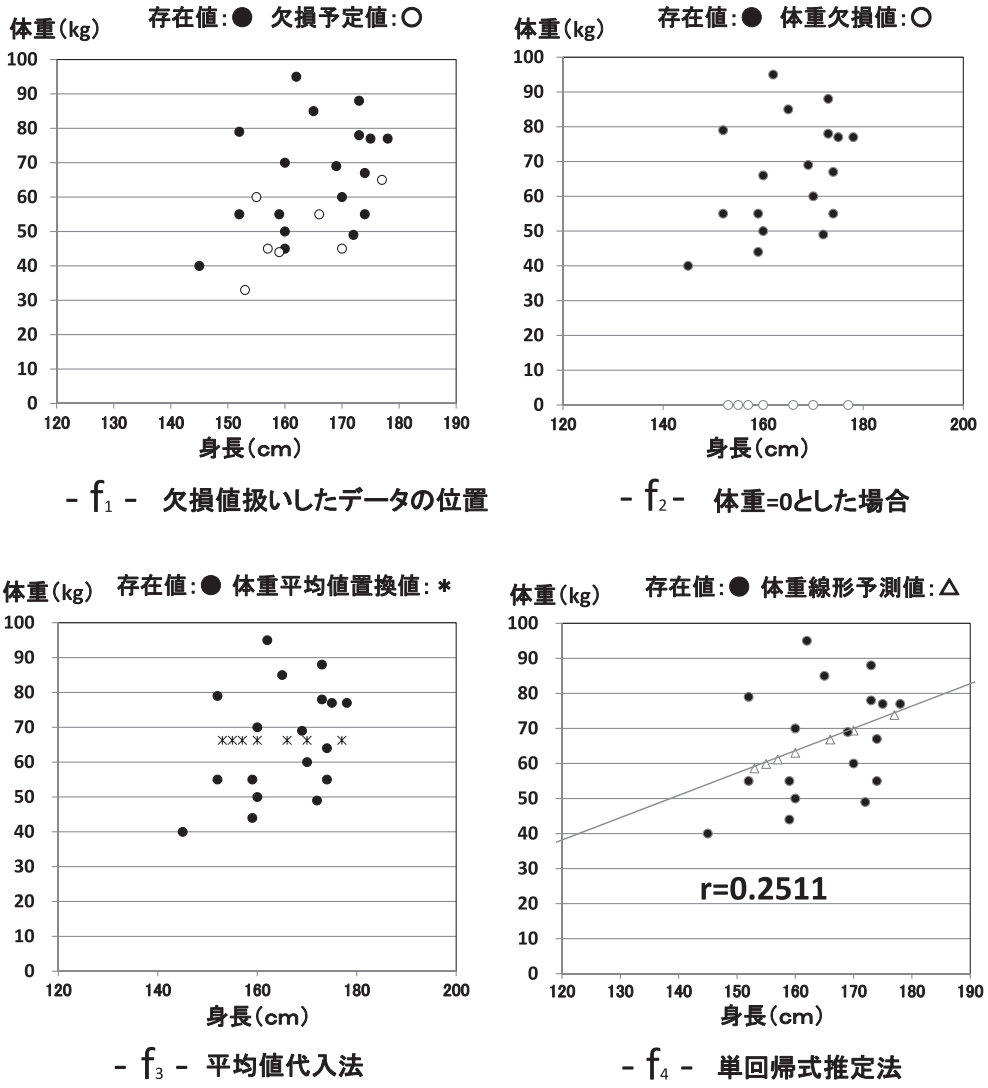


図3　欠損値が他の変数の影響を受ける場合の処理法

複数作成し，その結果をプールし統合することで欠損値データの統計的推測を行う方法である。多重代入法は、ある一つの欠測値に対して複数回の補完を行うことにより､この不確実性を欠損値補完値に反映させるのが最大の特徴である。図4は多重代入法の概念図である。本法の手順は次の3段階から構成される。

1．代入ステップ：欠測箇所にM個の異なる値を代入

2．分析ステップ：M個擬似的なデータセット（擬似完全データセット）を生成

3．統合ステップ：M種類の分析結果を一つに統合

ここで1の代入ステップにおける代入値の計算アルゴリズムは下記のとおりである。

Sub-step1-1：欠損値を含まないデータから初期

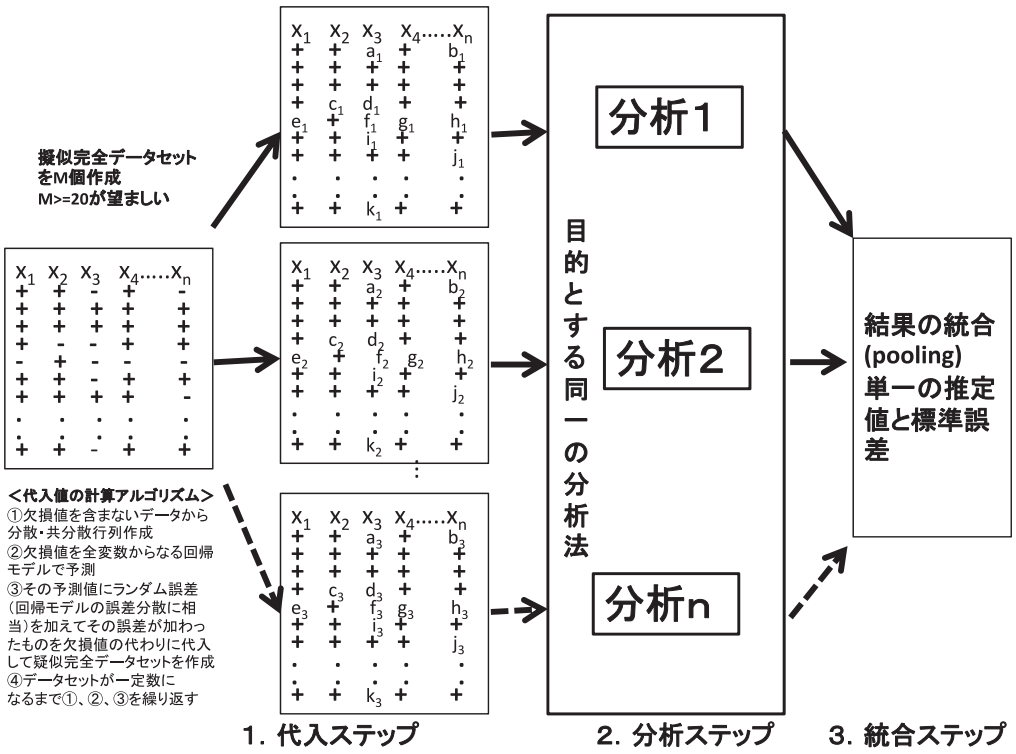


図4　多重代入法（MI：multiple imputation method）の概念図

値としての分散・共分散行列作成

Sub-step1-2：分散・共分散行列を用いた回帰モデルで欠損値の予測値を算出

Sub-step1-3：その予測値にランダム誤差（回帰モデルの誤差分散に相当）を加えてその誤差が加わったものを欠損値の代わりに代入して疑似完全データセットを作成

Sub-step1-4：データセットが一定数（M）になるまで1-1，1-2，1-3を繰り返す

擬似完全データセットの数（M）は小数だと不安定[8]になるためM>=20[Web02]が望ましいとされる。本法はspssや多くの統計ソフトでパッケージかしてきている。なお、Rubin DBはRCT法が実際的に不可能なケースで因果判定に効果を発揮するpropensity score法[Web04, Web05]の開発者としても有名であり、欠損値処理法とpropensity score法とは地下水脈で繋がっていることは興味深い。

(8) 完全情報最尤推定法（FIML）

著者は本統計シリーズ第2報[9]において確率（probability）と尤度（likelihood）の違いを論じた。例えば6面体の普通のサイコロをイメージして頂いて、3個を同時に振った時全て1が出る確率を計算するとサイコロの目は互いに独立事象であるから1/6 × 1/6 × 1/6=1/216 = 0.0046296 ≒ 1000回の試行のうち4回か5回となる、これが確率である。しかし、実際には市販されているものは1の目は溝が深く彫られていることで1が非常に出やすいものがある。あくまでも3つのサイコロいずれもが確率1/6で1の目が出ることを前提として各サイコロを無限に振った場合（すなわち母集団）の確率は1/6と決めつける方法が確率に基づく予測である。一方、例えば100回の試技で出た目からサイコロの1の目の出る確率は3つの間で一様に1/6ではなくそれぞれ1/5、1/5.5、1/6　である

のが尤もらしいという観察重視の立場に立つのが尤度（likelihood）である。そしてサイコロを繰り返し振ることにより、これらの尤度はある値に収束していく。この例で言えば1,000回試行した結果1/5.5、1/5.8、1/6.2となったとしたらこれらの値が最大推定尤度：MLE（maximum likelihood estimate）である。これもまた統計の祖、実験計画法、分散分析法の生みの親であるRA Fisherが確立した概念である。Fisherはこれも確率であるが確率とは違う扱いになることに配慮して（確率に敬意を表して）尤度という別名にしたという。

さてFIMLが（6）単一値代入法と異なる点はその名前－FIML：full information maximum likelihoodに表れている。すなわち、欠損値がもし存在していたら関連したであろう「全ての情報を総動員して最も尤もらしい推定値」を求める点にある。また多重代入法との違いは多変量回帰式で欠損値を予測するのではなく共分散構造分析法（analysis of covariance structures）[Web06]を用いて因子モデルで最大推定尤度（MLE）の条件下で欠損値を推定することである。例えば表1－aの例

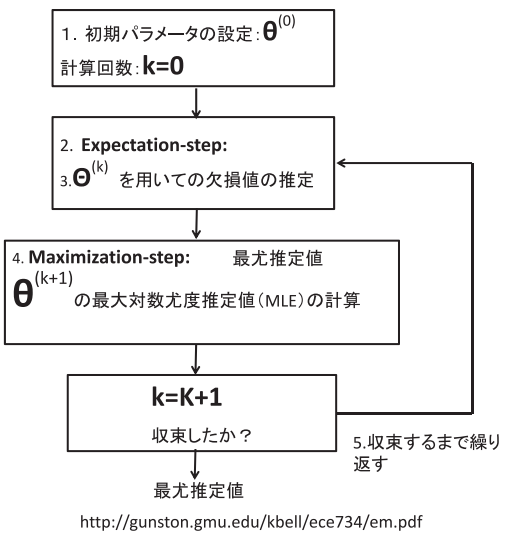


図5　EM algorithm：Eステップ（Expectation Step）とMステップ（Maximization step）の反復による最尤推定値の算出

を拡張して考えれば、体重に関わる変数は単に性別、年齢、身長、ADLだけではなく施設の食事状況、施設入所期間、経済状態、喫煙、飲酒、運動習慣、うつ等の精神状態、糖尿病等の基礎疾患の状態等があげられるだろう。これらの体重に関係する指標がほぼ十分に揃っていると仮定した場合、MIのようなそれぞれの指標を説明変数とする多変量回帰式ではなく、例えば性別、年齢、身長などの身体状況に関する因子、入所期間、経済状態などの生活自由度に関する因子、喫煙、飲酒、運動習慣などの生活習慣因子、うつ病、糖尿病などの疾病状況因子等の因子分析で求められた総合的で仮説的な因子との関連から欠損している体重を最も尤もらしく予測するのがFIMLである。具体的には期待値最大化アルゴリズム（Expectation-maximization algorithm）[10]等を用いて計算値を収束させて解を求める方法をとる。図5にFIMLの計算アルゴリズムを示す。

（9）MI法とFIML法の優位性および優劣について

John W[11]らはモンテカルロ法を用いてMI法とFIML法の優劣をシミュレーション研究で検証した。その結果、従来言われていたように両者は精度に関して等価（equivalent）であることが確認された。Peyre Hraら[12]は2003年実施の10年毎の36項目からなるフランス健康標本調査結果（参加者数　23,018人）から300人と1,000人の標本を抽出し、3，6，および9％の3種類の人工的欠損値を発生させMCAR、MARおよびMNARの3タイプの欠損値メカニズムの人口データを作成した。その上で欠損値処理法として平均値代入法、近似値代入法（hot-deck法）、MI、FIML法の優位性を比較した。その結果、MI法とFIML法が標準法との結論を得た。Nicholas JHら[13]もまたMI法とFIML法が 単一代入法より優れていることを確認した。但し、MI法はサンプリング数（m）を従来言われていたより増やすことが推奨している。以上、MI法とFIML法はこれからのデータ管理（クリーニング）には必需品と言える。

（10）異常値処理法１

①箱ひげ図による方法：図6に生活習慣病対策が急務のＡ国の成人のBMI分布（模擬データ）を性別、年齢別にXYプロット図として示した。WHOの3段階基準からみて女性にobese Ⅲが多いことが見て取れる。この場合、箱ひげ図ではどのような形状になるのであろうか。女性が群を抜いてはずれ値（outlets）が多いことになるのであろうか。図7に箱ひげ図：box（whisker）plot[Web07]を示す。図6の分布から見て取れる異常値情報と図7から見て取れる異常値とは本質的に別物であることがわかる。すなわち、前者は健康指標の基準からみてはずれる値を示す割合が多いことを示し、後者は統計的な分布上の異常値を表している。

箱ひげ図の使用に際してのポイントを図中に示したので細かい使用法の説明は省略する。また比較的よく使われる図なので改めて説明は不用かも知れないが使用にあたっての注意点は下記である。

(ⅰ)項目によって必ずしも異常値でない場合がある。例えば体重

(ⅱ)はずれ値を幾つか削除すると箱ひげ図のイメージが大きく変化するので確認は１回で終わらないで複数回箱ひげ図で確認した方がいい

(ⅲ)箱ひげ図の描画方法は複数存在するため、学術論文では方法を明記する

②Smirnov-Grubbs検定による方法：スミルノフ・グラブス検定[Web08]は外れ値検定と呼ばれ、データから外れ値を除くための手法である。平均値からのずれを標準偏差で割った値を元に、外れが大きいものから順に外れ値を除いていくので1回の検定で1つの外れ値を除き、外れ値がなくなるまで検定を繰り返す必要がある。また箱ひげ図と異なりはずれ値と判定するのに主観を除ける利点がある。しかし、機械的に異常値を判定しないようにすることが肝要である。

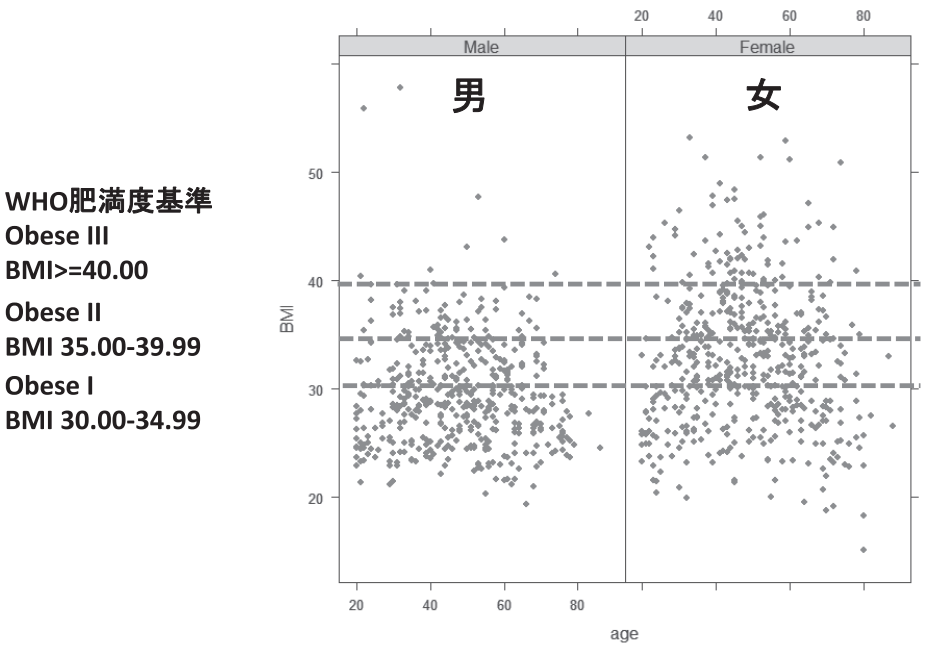


図6　生活習慣病対策が急務のＡ国の成人のBMI分布（模擬データ）－性別、年齢別XYプロット－

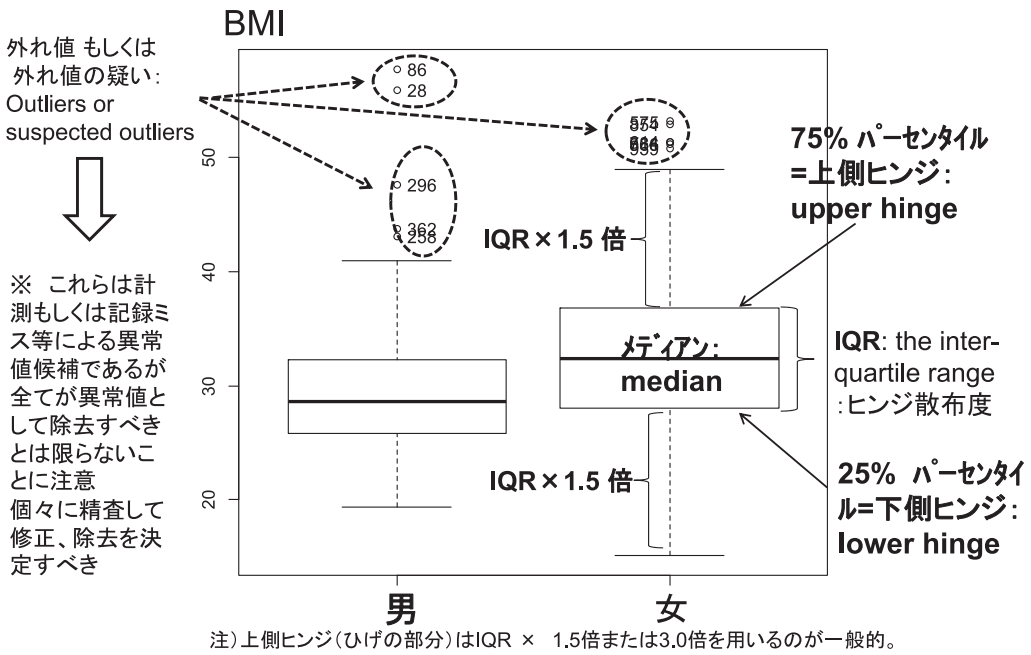


図7　－箱ひげ図：box-(whisker-) plot －による異常値の処理
－模擬データ：生活習慣病対策が急務のA国の成人の男女別BMI分布－

（11）異常値処理法２

①人口動態調査等におけるベイズ推定法[14, Web09)]

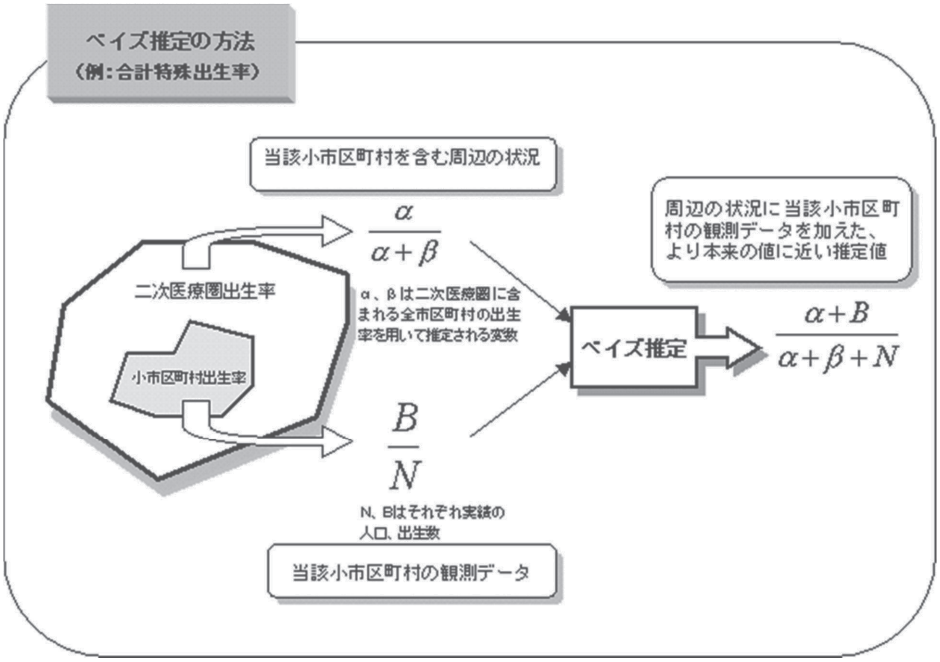
市区町村別生命表では、市区町村死亡率の推定にあたり、ベイズ統計学を用いて人口規模が小さいことに起因して時に観測死亡率データ値が著しく突出することがある。具体的には、図8に示すように当該市区町村を含むより広い地域である二次医療圏のグループの出生、死亡の状況を情報として活用し、これと各市区町村固有の出生、死亡数等の観測データとを総合化して当該市区町村の合計特殊出生率、標準化死亡比を推定するという形で「ベイズ推定」を適用し、数値を算出している。この例のように人口小規模地区特有の人口動態、疾病統計データ値の不安定性を緩和し、安定的な率推定を行うことが可能となっている。

②域集積性分析における扱い

本シリーズの第8報[15)]でこの問題に言及した。地域集積性を検証するデータにおいては欠損値があった場合、隣接する地域（市区町村等）のデータの平均値による置き換えると、その作業そのものが地域集積性を人為的にプラスの方向で評価するように働き、いわばデータを改竄するという由々しきことになる。この対策として地域集積性計算ソフトを提供している米国GeoDaセンター[Web10)]に確認したところ内海、湖、潟と同じような扱いが望ましい旨の回答を得た。このようにある地域が欠損値の場合は地理情報ファイル（shapefile）から除去するのが望ましい。すなわち内海、湖、潟と同じ扱いすることにする。しかしながら、例えば環境汚染、特に水質汚染などの情報等の場合は湖、潟は重要な意味を持ち、特定のデータ値を持つので除去してはならない。

おわりに

「歯科疫学統計」シリーズの第9報では欠損値・異常値処理法を扱った。統計シリーズの話の順序からすればもっと早い段階がよかったのだがテーマとしての独立性に難があったのでこれまで



厚生労働省：平成10年～平成14年　人口動態保健所・市区町村別統計の概況人口動態統計特殊報告
http://www.mhlw.go.jp/toukei/saikin/hw/jinkou/tokusyu/hoken04/5.html

図8　人口動態統計におけるベイズ推定法を用いた異常値補正法

遅れてしまった。しかしながら、Lubin DBの多重代入法（MI）の開発以来、内外の研究者の関心の高まりとPC機能の向上等によって本テーマは欠損値の数学的・統計的研究に専門の立場でない我々ユーザにおいても独立して議論すべきテーマと成りつつある。また臨床疫学手法の一人横綱として君臨しているRCT法の適用が事実上不可能な治療、手の予後、食習慣等の健康への影響等の研究においてpropensity score法が注目されてきており、その開発者も同じLubin DBであることからわかるように、欠損値処理とpropensity score法は地下水脈で繋がっている。こうしたことから今回あえて本テーマを独立させて解釈した。研究者各位が行っている研究のデータ管理の一環として本報が役立てば幸いである。

謝　辞

本総説を執筆するにあたり各段階でご指導いただいた歯科疫学研究会の顧問である蓑輪眞澄、高江洲義矩、境脩、伊藤学而、佐々木英忠、平田幸夫、山本龍生の各先生、ならびに統計解析に関して貴重な示唆をいただいた新潟医療福祉大学の森脇健介先生に深謝申し上げます。さらに貴重なご助言をいただいた深井穫博同研究会会長をはじめとする幹事の方々に感謝申し上げます。

参照 Web-site（2013/6/21 現在）

Web01）Additional examples of Missingness Mechanisms – Follow up to SON Brown Bag Presentation – 3/20/13（C Thompson）– Missing Data part 1 http://nursing.jhu.edu/faculty_research/research/documents/BBL_Presentation_March2013_Followup.pdf（2012年6月20日アクセス）

Web02）村山　航：欠損データ分析（missing data analysis）－完全情報最尤推定法と多重代入法－ http://www4.ocn.ne.jp/~murakou/missing_data.pdf（2012年10月1日アクセス）

Web03）Missing Data Mechanisms
http://www.psych.yorku.ca/lab/psy6140/lectures/missing614-2x2.pdf（2012年6月20日アクセス）
Web04）Propensity score matching,
http://en.wikipedia.org/wiki/Propensity_score_matching（2012年6月20日アクセス）
Web05）Melissa Humphries Population Research Center: Missing Data & How to Deal: An overview of missing data
http://www.utexas.edu/cola/centers/prc/_files/cs/Missing-Data.pdf（2012年6月20日アクセス）
Web06）The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models
http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1065&context=edpsychpapers
Web07）エクセル統計2012 箱ひげ図
http://software.ssri.co.jp/statweb2/sample/example_29.html（2013年6月20日アクセス）
Web08）エクセル統計2012　解析手法一覧外れ値検定スミルノフ・グラブス（Smirnov-Grubbs）検定
http://software.ssri.co.jp/ex/function/smirnov.html（2013年6月20日アクセス）
Web09）厚生労働省　ベイズ統計
http://www.mhlw.go.jp/toukei/saikin/hw/jinkou/tokusyu/hoken04/5.html（2012年10月1日アクセス）
Web10）GeoDa Center for geospatial analysis and computation
https://geodacenter.asu.edu/（2012年10月1日アクセス）

文　献

1）Joseph L. Schafer and John W. Graham: Missing Data: Our View of the State of the Art, Psychological Methods, vol. 7：147-177，2002.
2）Scheffer J : Dealing with Missing Data, Res. Lett. Inf. Math. Sci.vol 3：153-160，2002.
3）新井宏嘉，太田　亨：組成データ解析における0値および欠損値の扱いについて，地質学雑誌，vol 112：439-451，2006.
4）Rubin DB. Inference and missing data. Biometrika. 63：581–590，1976.
5）Rubin DB. Multiple Imputation for Nonresponse in Surveys. Wiley；1987.
6）Rubin, DB : Multiple imputation after 18+ years (with discussion). Journal of the American Statistical Association, 91：473-489，1996
7）岩崎　学：統計学大系シリーズ　不完全データの統計解析，第2刷，エコノミスト社，東京，2011．263-267頁.
8）Graham JW, Olchowski AE, Gilreath TD. : How many imputations are really needed? Some practical clarifications of multiple imputation theory, Prev. Sci. 8：206-213，2007.
9）瀧口　徹：歯科疫学統計　－第2報　一般化線形モデルの意義と潮流－，Health Science and Health Care, vol 4：29-37，2004.
10）Chuong B Do, Serafim Batzoglou : What is the expectation maximization algorithm?, Nature Biotechnology 26：897-899，2008.
11）John W. Graham & Allison E. Olchowski & Tamika D. Gilreath : How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory, Prev Sci, 8：206–213，2007.
12）Peyre H, Leplège A, Coste J. : Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey, Qual Life Res. 20：287-300，2011.
13）Nicholas JH, Ken PK:d Much ado about nothing : A comparison of missing data methods and software to fit incomplete data regression models, Am Stat., 61：79–90，2007.
14）涌井良幸，涌井貞美：Excelでスッキリわかるベイズ統計入門　Bayesian Statistics　日本実業出版社，第1版　第2刷，2011，148-152頁.
15）瀧口　徹：歯科疫学統計　－第8報　空間（地理）疫学の基礎　その2－地域差をとらえる指標の相互関係－ Health Science and Health Care, No 1 vol 10：4-19，2010.

A review of oral epidemiological statistics
－ Part IX ： A comparison of imputation techniques for handling missing values and abnormal values －

Toru Takiguchi
（Fukai Institute of Health Science）

Key Words : missing data (outlier), list-wise deletion, pair-wise deletion, personal mean score, multiple imputation (MI), full information maximum likelihood (FIML)

Recent progress in dealing with missing data (values) has come from an understanding of the reasons why data may be missing. These can be described under three categories:
The probability that a data value is missing :
・ MCAR (missing completely at random) is unrelated to the observed or missing values. Like tossing a coin.
・ MAR (missing at random) may depends on other observed variables, but not on the variable which is missing. Example: Respondents in service occupations less likely to report income..
・ MNAR (missing NOT at random) depends on missing value itself.
Example: Respondents with high income less likely to report income..
Both MI and FIML are most suitable methods for MCAR or MAR type data accoding to previous studies. On the other hand, the most logality is not established in the case of MNAR.

Following methods of dealing with missing data and were discussed.
Dealing with missing data in the narrow sense
1 default in handling of missing data
Deletion (non-imputation)
2 Pairwise deletion
3 Listwise deletion (Complete-case analysis)
Simple imputation
4 Personal mean (average) score
5 Personal worst score
6 Personal regression score
7 Hot-deck imputation (imputation of approximate value)
8 Cold-deck imputation (by multiple regression analysis)
9 Last observation (value)　carried forward
MI and FIML
Both methods are equivalent and recommended methods for handling missing data.
10 MI : multiple imputation
11 FIML : Full information maximum likelihood
Dealing with outliers

a）box-(whisker-) plot

b）Smirnov-Grubbs test for outliers

c）Bayesian estimation for data outliers of population dynamics (e.g. mortality rate in small population municipality)

d）handling of missing values or outliers in the analysis of regional clustering

Health Science and Health Care 12（2）：104－117，2012