日本統計学会誌
第27巻，第3号，1997年
263頁～280頁

統計表における秘匿の補完法

稲葉由之*，岩崎　学**

Imputation Procedures for Nondisclosure Cells in Statistical Tables

Yoshiyuki Inaba* and Manabu Iwasaki**

値が公表されないセルである秘匿の存在は，各種統計表における解析上の重大な阻害要素となっている．本論文の目的は，秘匿の存在により不完全となっている統計表を疑似的に完全な統計表にすることにある．秘匿は値を公表しない形式をとっているため，本問題は欠測値問題と同様な対処方法を考えることが可能である．しかし一方，問題の発生原因である秘匿には通常の欠測値問題では得られないような情報が存在する．そこで，この秘匿値に関する情報を考慮した秘匿の補完法を3つ提示し，その適用結果と補完値の分布評価に関するシミュレーション結果から，各補完法の有効性を確かめる．なお，ここで提案する方法は，一般的な欠測値の補完問題に区間情報等を追加することにより，補完精度を高めるための工夫を加えたものである．

1. はじめに

値が公表されないセルである秘匿の存在は，各種統計表における解析上の重大な阻害要素となっている．本論文の目的は，秘匿の存在により不完全となっている統計表を疑似的に完全な統計表にすることにある．秘匿は値を公表しない形式をとっているため，本問題は欠測値問題と同様な対処方法を考えることが可能である．Little and Rubin [10] は，欠測値への対処法を以下の4方法，(1) *Procedures Based on Completely Recorded Units*, (2) *Imputation-Based Procedures*, (3) *Weighting Procedures*, (4) *Model-Based Procedures*, に分類している．本論文では，3つの秘匿補完法を提案するが，それらの方法は *Imputation-Based Procedures* と *Model-Based Procedures* に基づいた方法である．*Imputation-Based Procedures* とは，欠測値に関して補完値を代入して分析する方法であり，*Mean imputation ; Regression imputation ; Hot deck imputation* などが含まれる．一方，*Model-Based Procedures* は，*Imputation-Based Procedures* を発展させた方法であり，欠測値に関するモデルを作成して分析を行う．これら各種方法の理論や適用例に関しては，Little and Rubin [10], Madow *et al.* [11, 12, 13] に整理されている．ところで，ここで問題とする秘匿を含む統計表には，通常の欠測値問題では得ることのできない情報が存在する．本論文では，秘匿値に関して得られる情報として，秘匿値に関する部分合計と，秘匿値に関する区間情報という2種類の情報に注目する．これらの情報については2.3節において説明し，秘匿値に関する情報を考慮した3つの秘匿補完法を提案する．3つの補完法とは，(1) 回帰による予測値を修正する方法，(2) 制約条件付きの回帰による方法，(3) 秘匿値に関する区間情報を考慮したEMアルゴリズム，である．はじめの2つの方法は，*Imputation-Based Procedures* のなかでも *Regression imputation* の考え方に基づいた方法であり，3番目の

論文受付：1997年1月　受理：1997年9月
* 成蹊大学工学研究科 (Seikei University)
** 成蹊大学工学部 (Seikei University)

方法は，*Model-Based Procedures* に分類される方法である．また，これらの方法は，秘匿値に関する情報を追加することにより，一般的な欠測値補完方法よりも補完精度を高めるための工夫を加えている．なお，本論文では，これら3つの方法を実際の統計表に適用し，その適用結果とともに，補完値の分布を評価するために行ったシミュレーション結果から各補完法の有効性を確認する．適用例としては，秘匿値の実際値に関する情報がある程度判明している市区部の統計表（千葉市における商業統計表）と，そのような情報が存在しない町村部の統計表（神奈川県町村部における商業統計表）の2例を用いることにする．

本論文の構成は以下のとおりである．まず第2章において，統計表における秘匿と秘匿値に関して得られる情報を簡単に説明し，第3章で秘匿値に関する情報を考慮した3つの補完法を提示する．そして，第4章において3つの補完法に関する適用例を示し，第5章で各補完法に関する考察を行う．

2. 統計表における秘匿と秘匿値に関する情報

2.1 秘匿の目的と対象

我が国において，統計表の秘匿は商業統計表や工業統計表などに設定されている．このような統計表では，秘匿はプライバシー保護を目的として設定されており，設定された箇所における集計値を公表しない形式をとっている．このとき秘匿の対象となるのは，商店数や工場数などの事業所数以外のすべての項目である．つまり，事業所数に関しては完全な統計表が得られている．また，秘匿はその設定基準の違いにより，単純秘匿と関連秘匿の2種類にわけることができる．川崎 [7, 8, 9] は，統計表における秘匿一般に関して整理し，秘匿の設定基準についても例をあげて説明している．

2.2 秘匿の設定基準

秘匿の設定基準は各統計表により異なるものであるが，本論文では，適用例でとりあげる商業統計表における秘匿の設定基準に関して説明を行う．いま，秘匿を含む統計表 **A** を考える．統計表 **A** は $(m+1)\times(n+1)$ 行列とし，$(m+1)(n+1)$ 個の要素からなる．統計表 **A** の要素は，m 行 n 列の内部要素 $a_{ij}, i=1,\cdots,m, j=1,\cdots,n$；$(n+1)$ 番目の列に示される行和 $a_{i,n+1}(=\sum_{j=1}^{n} a_{ij})$；$(m+1)$ 番目の行に示される列和 $a_{m+1,j}(=\sum_{i=1}^{m} a_{ij})$；総合計 $a_{m+1,n+1}(=\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij})$ である．このような秘匿を含む統計表 **A** に関して，秘匿箇所と公表箇所を示す指標行列 $\mathbf{R}(=(r_{ij}, i=1,\cdots,m+1, j=1,\cdots,n+1))$ をつぎのように定義する．

$$r_{ij}=\begin{cases}1, & a_{ij}\text{が公表されている},\\ 0, & a_{ij}\text{が秘匿されている}.\end{cases}$$

また，秘匿の対象とならない商店数に関する統計表を **B** とする．このとき，行列 **B** は，行列 **A** と同じく $(m+1)\times(n+1)$ 行列であり，要素も同様な意味をもつ．なお，以降において秘匿値を表す際には，秘匿を含む統計表 **A** における要素の形式 (a_{ij}) で示すことにする．

2.2.1 単純秘匿

単純秘匿は秘匿の目的に沿って設定される秘匿であり，商店数が1または2のセルを秘匿として設定する．この措置は，セルの値を公表することにより，個々の商店の状況が明示されないようにするための措置である．このため，単純秘匿の設定基準は，商店数に関する統計表 **B** のみに依存する (If $b_{ij}\leq 2$, then $r_{ij}=0$.)．なお，商店数が3以上であっても，秘匿値が算出される恐れのあるものについては，単純秘匿の対象となる場合がある．

2.2.2 関連秘匿

関連秘匿は，統計表 **A** の行和 $a_{i,n+1}$，あるいは列和 $a_{m+1,j}$ と公表箇所の値を用いて，単純秘匿

の値を計算することができないように設定する秘匿である．関連秘匿を設定した結果，秘匿が存在する行や列には，単純秘匿と関連秘匿を合わせて少なくとも2箇所の秘匿が存在することになる．また，関連秘匿は，他の関連する集計表における秘匿設定に影響を受けて設定されることもある．なお，商業統計表において関連秘匿の設定基準は公表されていない．

例1：商業統計表における秘匿

単純秘匿ならびに関連秘匿の例として，実際の商業統計表を示す．表1，表2は，平成6年商業統計表産業編市区町村編における千葉市の産業中分類別統計表から作成したものである．行には千葉市の各区，列には産業中分類の項目があり，表1には年間販売額を，表2には商店数を表示している．このため，表1は秘匿を含む統計表**A**を，表2は商店数に関する統計表**B**を表すことになる．表1によると，統計表における内部要素となる36カ所のセルのうち7カ所が秘匿として設定されている．このとき，3カ所のセル(a_{21}, a_{41}, a_{51})が単純秘匿，4カ所のセル($a_{25}, a_{46}, a_{55}, a_{56}$)が関連秘匿にあたる．このことは，商店数を示した表2において，3カ所のセル(a_{21}, a_{41}, a_{51})に対応する要素が2以下であることから判断することができる．また，残り4カ所の関連秘匿は，これら単純秘匿と，他の関連する集計表における秘匿設定に影響を受けて設定されたものであり，その設定基準は公表されていない．

表1　千葉市における産業中分類別商業統計表（年間販売額：百万円）

	各種商店	衣服等	飲食料品	自動車	家具等	その他	小売業計
中央区	154781	61509	90653	43588	34218	99166	483915
花見川区	a_{21}	10058	55098	29747	a_{25}	31559	140024
稲毛区	22567	8028	50029	33990	12564	32618	159798
若葉区	a_{41}	6339	47078	16476	11318	a_{46}	113107
緑区	a_{51}	3924	24145	7658	a_{55}	a_{56}	61292
美浜区	31793	7583	34618	30283	5901	21711	131889
千葉市（合計）	223673	97441	301620	161742	76904	228644	1090024

資料：平成6年商業統計表産業編市区町村表

表2　千葉市における産業中分類別商業統計表（商店数）

	各種商店	衣服等	飲食料品	自動車	家具等	その他	小売業計
中央区	8	582	880	133	208	836	2647
花見川区	1	91	418	103	88	309	1010
稲毛区	4	172	452	76	78	353	1135
若葉区	2	114	425	71	85	348	1045
緑区	2	88	211	33	28	175	537
美浜区	4	94	231	52	27	185	593
千葉市（合計）	21	1141	2617	468	514	2206	6967

資料：平成6年商業統計表産業編市区町村表

2.3　秘匿値に関する情報

秘匿の設定基準が公表されていない統計表であっても，つぎに示す2つの情報を得ることは可能である．この情報の存在が，欠測メカニズムに関する情報を得ることができない欠測値問題と大きく異なる点である．

2.3.1　秘匿値に関する部分合計

秘匿を含む統計表 **A** において，行和 $a_{i,n+1}$，列和 $a_{m+1,j}$ として表示された値は秘匿値を含んだ合計値である．このため，行和 $a_{i,n+1}$，列和 $a_{m+1,j}$，そして，指標行列 $\mathbf{R}=(r_{ij})$ の定義から，つぎの（1），（2）式が成り立つ．（1）（2）式における左辺第1項は公表箇所における合計値を，第2項は秘匿箇所における合計値（以降，秘匿値に関する部分合計と呼ぶ）を示している．

$$(1)\qquad \sum_{j=1}^{n} a_{ij}r_{ij}+\sum_{j=1}^{n} a_{ij}(1-r_{ij})=a_{i,n+1},\quad i=1,\cdots,m.$$

$$(2)\qquad \sum_{i=1}^{m} a_{ij}r_{ij}+\sum_{i=1}^{m} a_{ij}(1-r_{ij})=a_{m+1,j},\quad j=1,\cdots,n.$$

このとき，公表箇所における合計値は計算可能であり，行和 $a_{i,n+1}$，列和 $a_{m+1,j}$ については公表されていることがほとんどであるため，（1）（2）式から，秘匿を含む行や列において，秘匿値に関する部分合計を計算することが可能となる．

例2：秘匿値に関する部分合計の計算

ここでは，例1において示した表1を用いて秘匿値に関する部分合計を計算する．いま，表1における「花見川区」に注目する．この行には，a_{21}, a_{25} という2箇所の秘匿が含まれており，（1）式を適用すると，つぎのようになる．

$$\begin{aligned} a_{21}+a_{25}&=140024-(10058+55098+29747+31559)\\ &=13562. \end{aligned}$$

同様に，（1），（2）式を用いて，秘匿を含む行や列における秘匿値に関する部分合計を計算することができる．計算結果は表3に示すとおりである．

表3　秘匿値に関する部分合計

	部分合計
$a_{21}+a_{25}=$	13562
$a_{41}+a_{46}=$	31896
$a_{51}+a_{55}+a_{56}=$	25565
$a_{21}+a_{41}+a_{51}=$	14532
$a_{25}+a_{55}$	12903
$a_{46}+a_{56}=$	43590

2.3.2　秘匿値に関する区間情報

商業統計表では関連秘匿に関する設定基準は公表されていない．しかし，それを代替するような追加情報を公表値から計算することが可能である．この情報とは，秘匿箇所における上限値，下限値という区間情報であり，本論文では，対象とする統計表よりも，さらに細かな区分に分類された統計表の情報を用いて区間情報の計算を行う．計算方法としては，まずはじめに関連秘匿における下限値を計算し，計算した下限値と秘匿値に関する部分合計の情報を利用して，秘匿箇所における上限値ならびに下限値を反復的に求めていく．この計算に関して以下に説明する．

いま，秘匿を含む統計表 **A** において，列の要素 j をさらに細かく分類する区分 $k(k=1,\cdots,K_j)$ を考え，その要素を a_{ijk}, $i=1,\cdots,m+1$, $j=1,\cdots,n+1$, $k=1,\cdots,K_j$ と表すことにする．このとき，a_{ij} と a_{ijk} との関係は，$a_{ij}=\sum_{k=1}^{K_j} a_{ijk}$ となる．これは例えば，商業統計表において添字 j が産業中分類の区分を表すとき，添字 k は中分類をさらに細かく分類した産業小分類を表すこ

とになる．もし，ある産業中分類の項目が関連秘匿に設定されていたとしても，その箇所における商店数は2以下ではないため，その中分類を構成する小分類においていくつかの項目が公表されている可能性がある．いま，指標行列 $\mathbf{R}=(r_{ij})$ を統計表 $\mathbf{A}$ と同様に添字 k まで拡張し，$\mathbf{R}=(r_{ijk})$ とするとき，a_{ij} をつぎのように示すことができる．

$$(3)\qquad a_{ij}=\sum_{k=1}^{K_j}a_{ijk}r_{ijk}+\sum_{k=1}^{K_j}a_{ijk}(1-r_{ijk})\geq\sum_{k=1}^{K_j}a_{ijk}r_{ijk}.$$

このとき，(3)式における $\sum_{k=1}^{K_j}a_{ijk}r_{ijk}$ は公表箇所の合計値であり，この値を関連秘匿 a_{ij} の下限値として考えることができる．

例3：関連秘匿箇所における下限値の計算

表1において，関連秘匿は4カ所($a_{25}, a_{46}, a_{55}, a_{56}$)存在している．ここでは，この4カ所の関連秘匿のうち，a_{25} と a_{55} における下限値の計算を行う．表4に，家具等小売業の産業小分類区分における年間販売額(花見川区，緑区)を示す．なお，表中のセル a_{253}, a_{552} は産業小分類における秘匿箇所である．表4において(3)式を適用すると，つぎのように a_{25}, a_{55} の下限値を計算することができる．

$$\begin{aligned}a_{25}&=1569+1267+a_{253}+8097=10933+a_{253}\\&\geq 10933.\\a_{55}&=459+a_{552}+28+1195=1682+a_{552}\\&\geq 1682.\end{aligned}$$

表4　家具等小売業の産業小分類区分における年間販売額

	家　具	金　物	陶磁器	家庭用機　械	その他
花見川区	1569	1267	a_{253}	8097	0
緑区	459	a_{552}	28	1195	0

資料：平成6年商業統計表産業編市区町村表

多くの場合，例3のように関連秘匿箇所における下限値の計算を行うことは可能である．つぎに，この下限値に関する情報と秘匿値に関する部分合計を利用して，反復法により区間情報を計算する．いま，a_{ij} の上限値を $u_{ij}^{(t)}$，下限値を $l_{ij}^{(t)}$ とおき，(t) は反復計算における t 回目の値を示すものとする．まず，上限値，下限値の初期値 $u_{ij}^{(0)}, l_{ij}^{(0)}$ を設定する．下限値の初期値 $l_{ij}^{(0)}$ は，下限値に関する計算が可能な場合には(3)式を用いて計算した値とし，計算不可能な場合には0とする．また，上限値の初期値 $u_{ij}^{(0)}$ は，行や列において計算された部分合計のうち，より小さい方の値を初期値として設定する．つまり，つぎのように，上限値，下限値の初期値を公表箇所も含めて計算する．このとき，公表箇所における上限値と下限値の初期値は公表値に等しくなるように設定する．

$$(4)\qquad u_{ij}^{(0)}=\begin{cases}\min\{(a_{i,n+1}-\sum_{j=1}^{n}a_{ij}r_{ij}),(a_{m+1,j}-\sum_{i=1}^{m}a_{ij}r_{ij})\}, & \text{if } r_{ij}=0,\\ a_{ij}, & \text{if } r_{ij}=1.\end{cases}$$

$$(5)\qquad l_{ij}^{(0)}=\begin{cases}0, & \text{if } r_{ij}=0 \text{ and } \sum_{k=1}^{K_j}r_{ijk}=0,\\ \sum_{k=1}^{K_j}a_{ijk}r_{ijk}, & \text{if } r_{ij}=0 \text{ and } \sum_{k=1}^{K_j}r_{ijk}\geq 1,\\ a_{ij}, & \text{if } r_{ij}=1.\end{cases}$$

このように設定した初期値を用いて，つぎの上限値計算ステップ((6)式)と下限値計算ス

テップ（(7) 式）を値が変わらなくなるまで繰り返し，秘匿値に関する区間情報を計算する．また，最終的に得られた上限値を u_{ij}^{*}，下限値を l_{ij}^{*} とおくことにする．

＜上限値計算ステップ＞

行（あるいは列）に含まれる他の下限値の合計を，行和（あるいは列和）から引いた数値のうち，小さい値を上限値とする．

$$(6)\qquad u_{ij}^{(t)}=\min\left\{u_{ij}^{(t-1)},\left(a_{i,n+1}-\sum_{j=1}^{n}l_{ij}^{(t-1)}+l_{ij}^{(t-1)}\right),\left(a_{m+1,j}-\sum_{i=1}^{m}l_{ij}^{(t-1)}+l_{ij}^{(t-1)}\right)\right\}$$

＜下限値計算ステップ＞

行（あるいは列）に含まれる他の上限値の合計を，行和（あるいは列和）から引いた数値のうち，大きい値を下限値とする．

$$(7)\qquad l_{ij}^{(t)}=\max\left\{l_{ij}^{(t-1)},\left(a_{i,n+1}-\sum_{j=1}^{n}u_{ij}^{(t-1)}+u_{ij}^{(t-1)}\right),\left(a_{m+1,j}-\sum_{i=1}^{m}u_{ij}^{(t-1)}+u_{ij}^{(t-1)}\right)\right\}$$

例 4：秘匿値に関する区間情報の計算

上記の計算過程に従い，表 1 における秘匿値の区間情報を計算すると，表 5 に示すとおりとなる．この計算では，2 回目の計算ステップにおいて値が変わらなくなったため，秘匿値に関する区間情報は，$u_{ij}^{*}=u_{ij}^{(2)}$，$l_{ij}^{*}=l_{ij}^{(2)}$，となる．

表 5　秘匿値に関する区間情報の計算

秘匿箇所	$t=0$		$t=1$		$t=2$	
	$u_{ij}^{(0)}$	$l_{ij}^{(0)}$	$u_{ij}^{(1)}$	$l_{ij}^{(1)}$	$u_{ij}^{(2)}$	$l_{ij}^{(2)}$
a_{21}	13562	0	2629	2341	2629	2341
a_{25}	12903	10933	11221	10933	11221	10933
a_{41}	14532	0	3753	0	3753	0
a_{46}	31896	28143	31896	28143	31896	28143
a_{51}	14532	0	13146	8150	12189	8150
a_{55}	12903	1682	1970	1682	1970	1682
a_{56}	25565	10737	15447	11694	15447	11694

3. 秘匿値に関する情報を考慮した補完法

3.1 補完法において用いるモデル

本論文において提案する 3 つの補完法は，基本的に *Regression imputation* に基づいているため，秘匿値を推定するモデルが必要となる．本論文では，結果の比較を容易にするため，3 つの補完法はすべて同じモデルを使用する．秘匿値を推定するモデルは，秘匿を含む統計表 **A** の内部要素 $a_{ij},(i=1,\cdots,m,j=1,\cdots,n)$ を求めるために構成し，統計表における行の効果 (β_i) と列の効果 (γ_j) によるつぎのモデルを採用する．

$$(8)\qquad \begin{gathered} y_{ij}=\mu+\beta_i+\gamma_j+\varepsilon_{ij}\quad \varepsilon_{ij}\sim N(0,\sigma^2)\quad i=1,\cdots,m,j=1,\cdots,n,\\ y_{ij}=\log(a_{ij}/b_{ij}),\\ \textstyle\sum_{i=1}^{m}\beta_i=0,\quad \sum_{j=1}^{n}\gamma_j=0. \end{gathered}$$

このとき，a_{ij} は年間販売額，b_{ij} は商店数とするため，y_{ij} は第 i 行第 j 列のセルにおける1商店あたりの年間販売額の対数となる．また，モデル（(8) 式）のパラメータを $\boldsymbol{\theta}$ とおく．

$$\boldsymbol{\theta}=(\mu, \beta_1, \cdots, \beta_m, \gamma_1, \cdots, \gamma_n) \tag{9}$$

目的変数をこのような形式にしたのは2つの理由があり，一つは単純秘匿の設定が商店数のみに依存しているため，商店数が与えられた下で秘匿の設定をランダムであるとみなすことができるという点である．もう一つの理由は，適用例に対する予備解析に基づく理由であり，公表値をデータとしてパラメータ推定を実施すると，本モデル（(8) 式）の当てはまりが比較的良いとともに，誤差項における正規性の仮定を保つことができるという点である．

3.2 3つの補完法

本節では，第2章において説明した秘匿値に関する情報を考慮した3つの補完法を提案する．3つの補完法とは，(1) 回帰による予測値を修正する方法，(2) 制約条件付きの回帰による方法，(3) 秘匿値に関する区間情報を考慮した EM アルゴリズム，である．はじめの2つの方法は，秘匿値に関する部分合計の情報のみを利用した方法であり，3番目の方法は，秘匿値に関する区間情報も合わせて用いた方法である．このように，各補完法は，秘匿値に関する情報を追加することにより，一般的な欠測値補完方法よりも補完精度を高める工夫を加えた方法である．

3.2.1 回帰による予測値を修正する方法

回帰による予測値を修正する方法は2つの過程から成る．この方法では，まずはじめに秘匿箇所における予測値を計算し，つぎに，それを用いて秘匿箇所における補完値が秘匿値に関する部分合計を満たすように修正を加える．

第1の過程は以下に示すとおりであり，Healy and Westmacott [4] の提案した方法と同じ計算過程をとる．この過程のみであると，基本的な *Regression imputation* と同じ方法となる．まず，公表値のデータを用いて，モデル（(8) 式）のパラメータ $\boldsymbol{\theta}$（(9) 式）に関する推定を実施する．すなわち，次式を最小とするパラメータ $\boldsymbol{\theta}$ を求める．

$$S=\sum_{i=1}^{m}\sum_{j=1}^{n}\{y_{ij}-(\mu+\beta_i+\gamma_j)\}^2 r_{ij}. \tag{10}$$

つぎに，得られたパラメータ推定値 $\hat{\boldsymbol{\theta}}(=(\hat{\mu}, \hat{\beta}_1, \cdots, \hat{\beta}_m, \hat{\gamma}_1, \cdots, \hat{\gamma}_n))$ を用いて，つぎのように秘匿箇所における予測値を計算する．

$$\hat{a}_{ij}=b_{ij}\exp\{\hat{y}_{ij}\}=b_{ij}\exp\{\hat{\mu}+\hat{\beta}_i+\hat{\gamma}_j\}, \quad i=1, \cdots, m, j=1, \cdots, n. \tag{11}$$

第2の過程では，計算した予測値（(11) 式）を基にして，秘匿値に関する部分合計を満たすような補完値を求める．この過程では，秘匿値に関する部分合計を制約条件として，第1の過程で計算した予測値 $\hat{a}_{ij}$ との間のカイ二乗統計量を最小にするような補完値 a_{ij}^* を求める．すなわち，つぎに示す (13)，(14)，(15) 式の下で (12) 式を最小とする補完値 a_{ij}^* を求める．

$$\min \quad \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(a_{ij}^*-\hat{a}_{ij})^2}{a_{ij}^*}(1-r_{ij}) \tag{12}$$

$$\text{subject to} \quad \sum_{j=1}^{n}a_{ij}^*(1-r_{ij})=a_{i,n+1}-\sum_{j=1}^{n}a_{ij}r_{ij}, \quad i=1, \cdots, m. \tag{13}$$

$$\sum_{i=1}^{m}a_{ij}^*(1-r_{ij})=a_{m+1,j}-\sum_{i=1}^{m}a_{ij}r_{ij}, \quad j=1, \cdots, n. \tag{14}$$

(15) $$a_{ij}^{*}\geq 0,\quad i=1,\cdots,m, j=1,\cdots,n.$$

この計算の基準は，Deming and Stephan [1] の提案した方法と同様な基準を用いている．Deming and Stephan [1] の方法は，母集団における周辺分布が既知，同時分布が未知であるとき，母集団の周辺分布を制約条件として，同時分布に関する標本値を修正し，母集団の同時分布を推定しようとする方法である．このとき，例えば問題が2変量の場合には，推定した同時分布は母集団の状況を表すクロス集計表のセルの値を示し，集計表における行和列和は，既知である周辺分布に従うことになる．本補完法では，既知である母集団の周辺分布を秘匿箇所に関する部分合計に，同時分布に関する標本値を第1の過程で計算した予測値に置き換え，母集団の同時分布に相当する補完値を計算する．なお，本補完法における最適化計算は，秘匿箇所に限定して計算を行うという点で，Deming and Stephan [1] の方法とは異なっている．また，母集団における同時分布を推定する考え方とその方法についてはいくつかの研究があり，それらは Smith [16]によって整理されている．計算の基準としては，Deming and Stephan [1] がカイ二乗統計量を計算の基準としたのに対して，Ireland and Kullback [6]は，discrimination information を基準とすることを提案し，その統計的性質に関する考察を行っている．

3.2.2 制約条件付きの回帰による方法

先に示した補完法は2段階の過程による方法であった．これに対して，制約条件付きの回帰による方法は，計算を2段階にわけずに，得られた予測値をそのまま補完値として採用することができる方法である．

補完値の計算としては，予測値が秘匿値に関する部分合計という条件を満たす下でパラメータ $\boldsymbol{\theta}$ ((9) 式) を推定し，得られたパラメータ推定値 $\hat{\boldsymbol{\theta}}$ から秘匿箇所における予測値を計算する．この計算過程では，まず，(16)，(17)，(18) 式の下で (10) 式を最小とするパラメータ $\boldsymbol{\theta}$ ((9) 式) を求める．

(16) $$\sum_{j=1}^{n} b_{ij}\exp\{\mu+\beta_i+\gamma_j\}(1-r_{ij})=a_{i,n+1}-\sum_{j=1}^{n}a_{ij}r_{ij},\quad i=1,\cdots,m.$$

(17) $$\sum_{i=1}^{m} b_{ij}\exp\{\mu+\beta_i+\gamma_j\}(1-r_{ij})=a_{m+1,j}-\sum_{i=1}^{m}a_{ij}r_{ij},\quad j=1,\cdots,n.$$

(18) $$b_{ij}\exp\{\mu+\beta_i+\gamma_j\}\geq 0,\quad i=1,\cdots,m, j=1,\cdots,n.$$

つぎに，得られたパラメータ推定値 $\hat{\boldsymbol{\theta}}$ を用いて，(11)式のように秘匿箇所における予測値 $\hat{a}_{ij}$ を計算する．得られた予測値 $\hat{a}_{ij}$ は，秘匿値に関する部分合計の条件を満たしているため，そのまま補完値 a_{ij}^{*} として採用する．

3.2.3 秘匿値に関する区間情報を考慮した EM アルゴリズム

3番目に示す補完法は，秘匿値に関する部分合計という情報に加えて，秘匿値に関する区間情報も合わせて考慮した方法である．この方法では，EM アルゴリズム (Dempster *et al.* [2], Little and Rubin [10] を参照) を用いる．EM アルゴリズムは，欠測値に関する期待値を計算するステップ(Expectation の頭文字から E ステップと呼ぶ)と，得られた期待値をデータとみなして最尤推定を実施するステップ (Maximization の頭文字から M ステップと呼ぶ) の2ステップの反復によるアルゴリズムである．本方法では，秘匿値に関する区間情報を用いて EM アルゴリズムにおける E ステップを改良する．なお，付加的情報を用いて E ステップを改良した研究に，Hasslblad, Stead and Galke [3], Schmee and Hahn [15] がある．また，稲葉，岩崎 [5] では，E ステップにおいて秘匿値に関する部分合計を満たす方法を提案した．これに対して，本方法は秘匿値に関する区間情報を用いて期待値を修正する．つまり，E ステップにおい

て，ある定められた区間内の値をとるように期待値計算を実施する方法をとる．このような計算方法を用いるため，最終的に得られる秘匿箇所における予測値は秘匿値に関する部分合計を満たすことにはならない．したがって，本方法は，はじめに示した回帰による予測値を修正する方法と同様に，計算した予測値を基にして，秘匿値に関する部分合計を満たすような補完値を求める過程も加えるものとする．ゆえに，本方法は回帰による予測値を修正する方法に，秘匿値に関する区間情報を加えて改良した方法であると考えることができる．

改良する第1の過程は，以下に示すEステップとMステップからなるEMアルゴリズムをパラメータ推定値が収束するまで繰り返し，最終的に得られたパラメータ推定値を用いて予測値を計算する過程である．これは，Hasslblad, Stead and Galke [3] の用いた計算過程と同様な過程である．

<Eステップ>

本方法におけるEステップは，通常の方法とは異なり，秘匿値がある区間 $(l_{ij}^{*}\leq a_{ij}\leq u_{ij}^{*})$ 内に存在するという情報に基づいて期待値を修正する．つまり，つぎのように，$y_{ij}^{(t)}$, $y_{ij}^{2(t)}$ を計算する．このとき，添字 (t) は t ステップ目の計算値であることを示し，パラメータ $\boldsymbol{\theta}^{(t)}$ は，$\boldsymbol{\theta}^{(t)}=(\mu^{(t)}, \beta_1^{(t)}, \cdots, \beta_m^{(t)}, \gamma_1^{(t)}, \cdots, \gamma_n^{(t)})$ である．

(19)
$$y_{ij}^{(t)}=E(y_{ij}|\log(l_{ij}^{*}/b_{ij})\leq y_{ij}\leq\log(u_{ij}^{*}/b_{ij}), \boldsymbol{\theta}^{(t)})$$
$$=\begin{cases}(\mu^{(t)}+\beta_i^{(t)}+\gamma_j^{(t)})+\sigma^{(t)}\delta_{ij}^{(t)}, & \text{if } r_{ij}=0,\\ y_{ij}, & \text{if } r_{ij}=1.\end{cases}$$

(20)
$$y_{ij}^{2(t)}=E(y_{ij}^{2}|\log(l_{ij}^{*}/b_{ij})\leq y_{ij}\leq\log(u_{ij}^{*}/b_{ij}), \boldsymbol{\theta}^{(t)})$$
$$=\begin{cases}\{(\mu^{(t)}+\beta_i^{(t)}+\gamma_j^{(t)})+\sigma^{(t)}\delta_{ij}^{(t)}\}^2+\sigma^{2(t)}(1-\xi_{ij}^{(t)}), & \text{if } r_{ij}=0,\\ y_{ij}^{2}, & \text{if } r_{ij}=1.\end{cases}$$

このとき，$\delta_{ij}^{(t)}$, $\xi_{ij}^{(t)}$ は以下に示すとおりであり，関数 f は標準正規密度関数，関数 F は標準正規分布関数を示す．

$$\delta_{ij}^{(t)}=-\frac{f(d_{ij}^{(t)})-f(c_{ij}^{(t)})}{F(d_{ij}^{(t)})-F(c_{ij}^{(t)})},$$
$$\xi_{ij}^{(t)}=\delta_{ij}^{(t)2}+\frac{d_{ij}^{(t)}f(d_{ij}^{(t)})-c_{ij}^{(t)}f(c_{ij}^{(t)})}{F(d_{ij}^{(t)})-F(c_{ij}^{(t)})},$$
$$c_{ij}^{(t)}=\frac{\log(l_{ij}^{*}/b_{ij})-(\mu^{(t)}+\beta_i^{(t)}+\gamma_j^{(t)})}{\sigma^{(t)}},$$
$$d_{ij}^{(t)}=\frac{\log(u_{ij}^{*}/b_{ij})-(\mu^{(t)}+\beta_i^{(t)}+\gamma_j^{(t)})}{\sigma^{(t)}}.$$

なお，Eステップにおける計算式の導出に関しては **Appendix** に示すものとする．

<Mステップ>

Mステップでは，Eステップにおいて計算した $y_{ij}^{(t)}$ を用いて，(21)式を最小とするパラメータ $\boldsymbol{\theta}^{(t+1)}$ を求める．

(21)
$$S=\sum_{i=1}^{m}\sum_{j=1}^{n}\{y_{ij}^{(t)}-(\mu^{(t+1)}+\beta_i^{(t+1)}+\gamma_j^{(t+1)})\}^2.$$

また，パラメータ $\sigma^{2(t+1)}$ は，$y_{ij}^{2(t)}$ における計算を利用して，

(22)
$$\sigma^{2(t+1)}=(mn-(m+n+1))^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\{\{y_{ij}^{(t)}-(\mu^{(t)}+\beta_i^{(t)}+\gamma_j^{(t)})\}^2 r_{ij}$$
$$+\sigma^{2(t)}(1-\xi_{ij}^{(t)})(1-r_{ij})\}$$

と計算する．なお，本方法の適用に際しては，データ数(mn)に比べてパラメータ数($m+n-1$)が大きくなるため，Mステップにおける分散評価式（(22) 式）は分母を(mn)ではなく，($mn-(m+n-1)$)として計算している．以上のEステップ，Mステップの反復により，最終的に得られたパラメータ推定値を $\hat{\boldsymbol{\theta}}$ とおき，(11)式のように秘匿箇所における予測値 $\hat{a}_{ij}$ を計算する．

第2の過程は，回帰による予測値を修正する方法と同様に，計算した予測値 $\hat{a}_{ij}$ を基にして，秘匿値に関する部分合計を満たすような補完値 a_{ij}^{*} を求める過程である．本方法の場合，秘匿値に関する区間情報も利用しているため，つぎの (23) 式も制約条件として加えるものとする．すなわち，(13)，(14)，(15)式，及び(23)式の下で(12)式を最小とする補完値 a_{ij}^{*} を求める．

$$l_{ij}^{*} \leq a_{ij}^{*} \leq u_{ij}^{*}, \quad i=1,\cdots,m, j=1,\cdots,n. \tag{23}$$

4. 適　用　例

本章では，提案した3つの補完法を実際の統計表に対して適用し，その有効性を確かめる．以降，各補完法はそれぞれ「予測値を修正する方法」，「制約付きの回帰による方法」，「区間情報によるEMアルゴリズム」と呼ぶことにする．

適用例としては，秘匿値の実際値に関する情報がある程度判明している市区部の統計表（千葉市における商業統計表）と，そのような情報のない町村部の統計表（神奈川県村部における商業統計表）の2例を用いる．秘匿値の実際値に関する情報とは，適用例に用いる産業編の統計表とは異なる区分で集計した品目編における統計表から得られた情報である．この情報は産業編と品目編における百貨店の集計に関する定義が等しく，品目編で百貨店が延べ商店数として集計されるため，単純秘匿の対象とならないセルが存在することを利用して計算した．なお，百貨店は各種商店小売業に含まれる産業小分類の項目である．しかし，このような計算が実施可能な統計表は，市区部における限られた統計表であり，本例における千葉市の統計表はそれにあたる．また，町村部における統計表では，このような秘匿値の実際値に関する情報を計算することは不可能であり，産業小分類の統計表も公表されていない．

4.1　多くの情報を利用可能な統計表に対する適用

4.1.1　千葉市区部における商業統計表

まず，秘匿値の実際値に関する情報が他の集計結果から計算可能となる統計表を例にあげる．用いる統計表は，例1において示した千葉市区部における商業統計表（表1参照）であり，表中における内部要素である36カ所のセルのうち7カ所が秘匿として設定されている．このときモデルは，区の効果(β_i)と産業分類の効果(γ_j)によるモデル（(8) 式参照）を用い，目的変数 y_{ij} は第 i 行第 j 列のセルにおける1商店当たりの年間販売額の対数となる．

4.1.2　秘匿値に関する情報

(1)　秘匿値に関する部分合計

秘匿値に関する部分合計は，例2において計算した表3に示したとおりである．

(2)　秘匿値に関する区間情報

表1に示した統計表は産業中分類による区分であり，市区部の統計表であることからさらに細かい分類である産業小分類の情報を区間情報の計算に用いることが可能となる．そこで，例3及び例4において示した過程を経て，表5に示したような秘匿値に関する区間情報を得ることができる．なお，表5における $t=2$ の数値が秘匿値に関する区間情報となる．

4.1.3　補完結果とシミュレーションによる補完値分布に関する評価

(1)　補完値の状況

各補完法を適用した結果得られた補完値は，表6に示すとおりである．表6に表示した「品

表6　各補完法による補完値の状況（千葉市区部）

秘匿箇所	(1) 回帰の予測値を修正する方法		(2) 制約条件付きの回帰による方法	(3)秘匿値に関する区間情報を考慮したEMアルゴリズム		品目編から得られた秘匿値の実際値に関する情報
	補完値	予測値	補完値	補完値	予測値	
a_{21}	3427	9604	3475	2454	2495	2581
a_{25}	10135	16029	10087	11108	11076	10981
a_{41}	4207	13737	4959	2788	3272	3679～3716
a_{46}	27689	26785	26937	29108	29900	28180～28217
a_{51}	6898	12808	6096	9290	9678	8234～8271
a_{55}	2768	3401	2816	1795	1833	1922
a_{56}	15899	12559	16653	14480	13393	15372～15409

目編から得られた秘匿値の実際値に関する情報」と補完値とを比較することにより，補完結果の有効性を確認することができる．実際値と補完値の状況を比較すると，各補完法における補完値は，おおよそ実際値に近い値をとっていることがわかる．予測値を修正する方法では，予測値においてはかなり離れた値となっているが，第2の過程で修正した補完値は実際値に近づいている．また，制約付きの回帰による方法は，予測値を修正する方法と同様な補完結果となる．これらに対して，区間情報を考慮したEMアルゴリズムでは結果の傾向が異なるが，これは他の2つの方法では考慮しなかった区間情報を用いたことによるものである．

(2)　シミュレーションによる補完値分布に関する評価

表6において示した補完値は，秘匿値に関する期待値としての意味をもち，補完値の分布を示したものではない．そこでここでは，補完値の分布としての評価を実施する．*Regression imputation* などでは，回帰による条件付き期待値と誤差分散により，代入値の分布を評価することができる．しかし，提案した補完法では，秘匿値に関する部分合計と補完値との整合性をとる過程も加えているため，その評価を理論的に実施することが困難である．そこで，シミュレーションから補完値の分布を評価する．また，このシミュレーションは，疑似データを多数セット用意する *Multiple imputation* (Rubin [14]) の可能性を探る意味もある．

評価する補完法は，秘匿値に関する部分合計と補完値との整合性をとる過程を含んだ，予測値を修正する方法と区間情報を考慮したEMアルゴリズムの2方法である．シミュレーションの方法としては，第1の過程で計算した予測値を，その分布から乱数を用いて生成し，生成した値を用いて最適化計算により補完値を求めるという方法である．このとき，予測値の分布はそれぞれ独立な正規分布とし，平均を予測値，分散をモデルの誤差分散とする．また，区間情報を考慮したEMアルゴリズムでは，区間情報が得られていることを前提とするため，棄却法を用いて，生成する予測値が区間内の値をとるようにする．

シミュレーションは100回実施し，得られた補完値の最大値，最小値を表7に示す．これによると，予測値を修正する方法では，最大値，最小値による区間がほぼ区間情報を含む結果となった．また，区間情報を考慮したEMアルゴリズムでは，最大値，最小値による区間が区間情報内に含まれている．この結果における注意点としては，区間情報を考慮したEMアルゴリズムによる最大値，最小値の区間が，4つの秘匿箇所（$a_{41}, a_{46}, a_{51}, a_{56}$）において実際値（表6参照）を含んでいない点にある．これは，実際値が区間情報の下限値あるいは上限値に近いことから生じているものと考えられる．このように，区間情報を考慮したEMアルゴリズムでは，

表7　シミュレーションによる補完値の分布（千葉市区部）

秘匿箇所	(1) 回帰の予測値を修正する方法		(3)秘匿値に関する区間情報を考慮したEMアルゴリズム		秘匿値に関する区間情報	
	最大値	最小値	最大値	最小値	上限値	下限値
a_{21}	5620	1809	2703	2241	2629	2341
a_{25}	11753	7942	11321	10859	11221	10933
a_{41}	9133	0	3628	1615	3753	0
a_{46}	31896	22763	30281	28268	31896	28143
a_{51}	11118	2429	10309	8384	12189	8150
a_{55}	4961	1150	2044	1582	1970	1682
a_{56}	20825	11692	15320	13307	15447	11694

補完値の分布として精度の高い分布が得られているようにみえるが，実際の値を含んでいない可能性があることが問題である．

4.2　利用情報に制限のある統計表に対する適用

4.2.1　神奈川県町村部における商業統計表

つぎに，秘匿値の実際値に関する情報を計算できない統計表に対する適用例にあげる．用いる統計表は，神奈川県町村部における年間販売額の商業統計表（表8参照）であり，表中にお

表8　神奈川県町村部における産業中分類別商業統計表（年間販売額：百万円）

	各種商店	衣服等	飲食料品	自動車	家具等	その他	小売業計
葉山町	0	467	10187	2103	906	4544	18208
寒川町	0	1571	20121	5395	1651	8691	37429
大磯町	0	1566	9086	1546	582	6274	19054
二宮町	7279	965	9210	1825	1123	4303	24705
中井町	0	a_{52}	2420	1477	a_{55}	1716	5940
大井町	0	1343	8490	6881	2738	4595	24047
松田町	a_{71}	1210	4745	a_{73}	722	3051	9806
山北町	0	360	3693	1135	124	3276	8586
開成町	a_{91}	463	4050	3991	a_{94}	5400	17680
箱根町	0	580	11470	360	218	15776	28404
真鶴町	0	246	3935	$a_{11,4}$	$a_{11,5}$	2007	6828
湯河原町	0	1319	18850	1007	1656	7365	30199
愛川町	$a_{13,1}$	1355	13778	$a_{13,4}$	1733	8016	31563
清川町	0	$a_{14,2}$	690	0	$a_{14,5}$	313	1103
城山町	0	500	6474	3383	1206	4582	16145
津久井町	0	608	8359	1734	1329	6331	18362
相模湖町	0	126	3877	465	472	1836	6775
藤野町	0	69	2307	$a_{18,4}$	$a_{18,5}$	1541	4099
町村部（合計）	13394	12861	141739	34998	16319	89620	308930

資料：平成6年商業統計表産業編市区町村表

表9　神奈川県町村部における産業中分類別商業統計表（商店数）

	各種商店	衣服等	飲食料品	自動車	家具等	その他	小売業計
葉山町	0	21	94	15	28	64	222
寒川町	0	32	153	36	29	99	349
大磯町	0	28	154	18	24	92	316
二宮町	3	39	145	12	23	84	306
中井町	0	3	39	9	6	22	79
大井町	0	10	43	23	12	36	124
松田町	1	20	75	3	16	60	175
山北町	0	17	79	10	9	42	157
開成町	1	13	64	18	13	38	147
箱根町	0	19	111	5	12	151	298
真鶴町	0	11	63	1	8	41	124
湯河原町	0	59	185	13	44	97	398
愛川町	1	29	141	17	20	94	302
清川町	0	1	11	0	4	4	20
城山町	0	8	53	11	13	43	128
津久井町	0	18	110	12	29	69	238
相模湖町	0	7	53	5	16	32	113
藤野町	0	5	57	2	2	20	86
町村部（合計）	6	340	1630	210	308	1088	3582

資料：平成6年商業統計表産業編市区町村表

ける内部要素である108カ所のセルのうち14カ所が秘匿として設定されている．また，商店数の統計表は表9に示すとおりである．モデルは，先程の例と同様に，区の効果（β_i）と産業分類の効果（γ_j）によるモデル（（8）式参照）を用い，商店が存在しない15カ所のセルはデータから除外して考えることにした．

4.2.2　秘匿値に関する情報

(1)　秘匿値に関する部分合計

秘匿値に関する部分合計の計算結果は，表10に示すとおりである．

(2)　秘匿値に関する区間情報

神奈川県町村部においては，産業中分類よりも細かい分類である産業小分類の統計表が公表されていない．そこで，下限値の初期値 $l_{ij}^{(0)}$ はすべて0とおき，表11に示すような計算過程により秘匿値に関する区間情報を得ることができる．なお，表11における $t=3$ の数値が秘匿値に関する区間情報となる．

4.2.3　補完結果とシミュレーションによる補完値分布に関する評価

(1)　補完値の状況

各補完法を適用した結果得られた補完値は，表12に示すとおりである．本例では，実際値を計算することができないため，秘匿値に関する区間情報との比較を行う．これによると，各補完法における補完値は，a_{71} や a_{74} という一部の秘匿箇所を除いて，同様な傾向の値をとっていることがわかる．

表10　秘匿値に関する部分合計

	部分合計
$a_{52}+a_{55}=$	327
$a_{71}+a_{74}=$	78
$a_{91}+a_{95}=$	3776
$a_{11,4}+a_{11,5}=$	640
$a_{13,1}+a_{13,4}=$	6681
$a_{14,2}+a_{14,5}=$	100
$a_{18,4}+a_{18,5}=$	182
$a_{71}+a_{91}+a_{13,1}=$	6115
$a_{52}+a_{14,2}=$	113
$a_{74}+a_{11,4}+a_{13,4}+a_{18,4}=$	3696
$a_{55}+a_{95}+a_{11,5}+a_{14,5}+a_{18,5}=$	1859

表11　秘匿値に関する区間情報の計算

秘匿箇所	$t=0$		$t=1$		$t=2$		$t=3$	
	$u_{ij}^{(0)}$	$l_{ij}^{(0)}$	$u_{ij}^{(1)}$	$l_{ij}^{(1)}$	$u_{ij}^{(2)}$	$l_{ij}^{(2)}$	$u_{ij}^{(3)}$	$l_{ij}^{(3)}$
a_{52}	113	0	113	13	113	13	113	13
a_{55}	327	0	327	214	314	214	314	214
a_{71}	78	0	78	0	78	0	78	0
a_{74}	78	0	78	0	78	0	78	0
a_{91}	3776	0	3776	1917	3130	2152	3130	2152
a_{95}	1859	0	1859	610	1645	646	1624	646
$a_{11,4}$	640	0	640	0	640	0	640	0
$a_{11,5}$	640	0	640	0	640	0	640	0
$a_{13,1}$	6115	0	6115	2985	3885	2985	3885	2985
$a_{13,4}$	3696	0	3696	2796	3696	2796	3696	2796
$a_{14,2}$	100	0	100	0	100	0	100	0
$a_{14,5}$	100	0	100	0	100	0	100	0
$a_{18,4}$	182	0	182	0	182	0	182	0
$a_{18,5}$	182	0	182	0	182	0	182	0

(2)　シミュレーションによる補完法に関する評価

先の例と同様にシミュレーションを100回実施し，その結果を表13に示す．これによると，予測値を修正する方法でも，最大値，最小値による区間がほぼ区間情報内の値をとり，補完値の分布として大きな問題はないと考えることができる．

5. 考　　察

前章では2つの適用例を示したが，はじめの適用例から，各補完法とも実際値に近い補完値が得られており，疑似データとして使用することに大きな問題がないことを確認した．ここで

表12　各補完法による補完値の状況（神奈川県町村部）

秘匿箇所	(1) 回帰の予測値を修正する方法		(2) 制約条件付きの回帰による方法	(3)秘匿値に関する区間情報を考慮したEMアルゴリズム		秘匿値に関する区間情報	
	補完値	予測値	補完値	補完値	予測値	上限値	下限値
a_{52}	96	100	96	83	68	113	13
a_{55}	231	245	231	244	258	314	214
a_{71}	0	2774	66	39	68	78	0
a_{74}	78	399	12	39	61	78	0
a_{91}	2779	3493	2785	2748	2545	3130	2152
a_{95}	997	679	991	1028	975	1624	646
$a_{11,4}$	152	91	141	175	78	640	0
$a_{11,5}$	488	228	499	465	214	640	0
$a_{13,1}$	3336	3879	3264	3327	3359	3885	2985
$a_{13,4}$	3345	3163	3417	3354	3215	3696	2796
$a_{14,2}$	17	30	17	30	22	100	0
$a_{14,5}$	83	147	83	70	67	100	0
$a_{18,4}$	121	156	126	129	100	182	0
$a_{18,5}$	61	49	56	53	44	182	0

表13　シミュレーションによる補完値の分布（神奈川県町村部）

秘匿箇所	(1) 回帰の予測値を修正する方法		(3)秘匿値に関する区間情報を考慮したEMアルゴリズム		秘匿値に関する区間情報	
	最大値	最小値	最大値	最小値	上限値	下限値
a_{52}	113	67	102	38	113	13
a_{55}	260	214	289	225	314	214
a_{71}	78	0	59	22	78	0
a_{74}	78	0	56	19	78	0
a_{91}	3052	2656	3021	2590	3130	2152
a_{95}	1120	724	1186	755	1624	646
$a_{11,4}$	396	0	462	31	640	0
$a_{11,5}$	640	244	609	178	640	0
$a_{13,1}$	3459	2985	3475	3064	3885	2985
$a_{13,4}$	3696	3222	3617	3206	3696	2796
$a_{14,2}$	46	0	75	11	100	0
$a_{14,5}$	100	54	89	25	100	0
$a_{18,4}$	168	101	164	44	182	0
$a_{18,5}$	81	19	138	18	182	0

は，そのアプローチの仕方により，いくつかの特徴がみられる各補完法に関する適用上の考察を行う．

各補完法を簡便性の面で評価すると，予測値を修正する方法が最も簡単に実施することができる．この方法は他の2つの方法に比べれば簡単な方法ではあるものの，適用例における補完値の状況としては他の2つの方法とほぼ同様な結果を得ている．しかし，モデルにおけるパラメータを評価する際には，予測値を計算する段階において秘匿値に関する情報を全く考慮していないため，他の方法に比べて劣っていることになる．提案した各補完法はモデル選択のための方法とは言えないが，モデル選択を行う際には，秘匿箇所に想定する値は実際値に近いことが望ましい．区間情報を考慮したEMアルゴリズムは，パラメータ推定の際に用いる代入値が実際値に近いことがわかっているため，パラメータを評価する上では最も優れている方法である．一方，数理計画法に関するアプリケーションが利用可能な場合には，制約付きの回帰による方法の適用が容易となる．また，制約付きの回帰による方法を除いた他の2つの方法は，最後に秘匿値に関する部分合計を満たすための最適化計算を行っている．しかし，この計算は，補完値と秘匿値に関する部分合計との整合性をとる方法にすぎないため，得られた補完値がすべての場合において実際値に近づく保証がないことに注意しなければならない．

適用例においては，補完値の分布を評価するためのシミュレーションを実施しており，このシミュレーションでは，*Multiple imputation* の可能性も確認している．この結果，予測値を修正する方法では，区間情報を考慮しないことから，比較的広い幅をもつ分布となる．また，区間情報を考慮したEMアルゴリズムでは，秘匿箇所における実際値が区間情報の上限値あるいは下限値に近いときには，得られた補完値の分布が実際値を含まない場合がある．したがって，これらの補完法に関するシミュレーションを *Multiple imputation* として用いるには問題が残ることになる．また，この問題は，より有効なモデルの選択により，ある程度解決可能であると考えられるものの，本論文において提案した3つの補完法はモデル選択に適切であるとは言えない．このため，今後はモデル選択に関する方法について研究を進める必要がある．

謝　辞

査読者の方には数々の有益なご指摘を頂きました．深く感謝いたします．

参考文献

[1] Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when expected marginal totals are known, *Ann. Math. Statist.*, **11**, 427-444.
[2] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Statist. Soc., B* **39**, 1-38.
[3] Hasselblad, V., Stead, A. G. and Gralke, W. (1980). Analysis of coarsely grouped data from the lognormal distribution, *J. Am. Statist. Assoc.*, **75**, 771-778.
[4] Healy, M. J. R. and Westmacott, M. (1956). Missing values in experiments analyzed on automatic computers, *Appl. Statist.*, **5**, 203-206.
[5] 稲葉由之，岩崎　学．(1996)．クロス集計表における秘匿の影響に関する数値的評価，「応用統計学」，**25**，61-72.
[6] Ireland, C. T. and Kullback, S. (1968). Contingency tables with given marginals, *Biometrika*, **55**, 179-188.
[7] 川崎　茂．(1993)．統計提供における秘密保護（第1回），「統計」，**44**，1，77-80.
[8] 川崎　茂．(1993)．統計提供における秘密保護（第2回），「統計」，**44**，2，67-70.
[9] 川崎　茂．(1993)．統計提供における秘密保護（第3回），「統計」，**44**，3，70-73.
[10] Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, John Wiley & Sons.
[11] Madow, W. G., Nisselson, H. and Olkin, I. (1983). *Incomplete Data in Sample Surveys, Vol. 1 : Report and case studies*, Academic Press.

[12] Madow, W. G. and Olkin, I. (1983). *Incomplete Data in Sample Surveys, Vol. 3 : Proceedings of the Symposium*, Academic Press.
[13] Madow, W. G., Olkin, I. and Rubin, D. B. (1983). *Incomplete Data in Sample Surveys, Vol. 2 : Theory and Bibliographies*, Academic Press.
[14] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons.
[15] Schmee, J. and Hahn, G. J. (1979). A simple method for regression analysis with censored data, *Technometrics*, **21**, 417-432.
[16] Smith, J. H. (1947). Estimation of linear functions of cell proportions, *Ann. Math. Statist.*, **18**, 231-254.

Appendix : 秘匿値に関する区間情報を考慮したEステップの修正

本論文で提案した補完法「秘匿値に関する区間情報を考慮したEMアルゴリズム」におけるEステップでは，秘匿値に関する区間情報を考慮して期待値計算の修正を行っている．ここでは，この期待値の修正に係わる計算式導出を示す．

いま，変数 y が平均 μ，分散 σ^2 の正規分布に従っているものとする $(y\sim N(\mu,\sigma^2))$．このとき，y がある区間 $(p\leq y\leq q)$ 内に存在することが判明していると考える．そこで，$p\leq y\leq q$ という条件の下での y の期待値を求めることが，期待値計算の修正となる．

まず，$\int_p^q g(y)dy=1$ となる密度関数 $g(y)$ を考えると，

$$g(y)=\frac{1}{C}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\},\ p\leq y\leq q,$$
$$C=\int_p^q\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}dy,$$

となる．この $g(y)$ を用いて y の期待値を求めると，以下のようになる．なお，関数 f は標準正規密度関数，関数 F は標準正規分布関数を示す．

$$\begin{aligned}
&E(y|p\leq y\leq q)\\
&=\int_p^q yg(y)dy\\
&=\int_p^q \mu g(y)dy+\int_p^q (y-\mu)g(y)dy\\
&=\mu\frac{1}{C}\int_p^q\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}dy+\frac{\sigma}{C}\frac{1}{\sqrt{2\pi}}\int_p^q\frac{(y-\mu)}{\sigma^2}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}dy\\
&=\mu\frac{C}{C}+\frac{\sigma}{C}\frac{1}{\sqrt{2\pi}}\left[-\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}\right]_p^q\\
&=\mu-\frac{\sigma}{C}\left\{f\left(\frac{q-\mu}{\sigma}\right)-f\left(\frac{p-\mu}{\sigma}\right)\right\}.
\end{aligned}$$

このとき，$z=(y-\mu)/\sigma$, $c=(p-\mu)/\sigma$, $d=(q-\mu)/\sigma$ とおくと，C はつぎのようになる．

$$\begin{aligned}
C&=\int_c^d\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{z^2}{2}\right\}\sigma dz\\
&=\int_c^d\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{z^2}{2}\right\}dz\\
&=F(d)-F(c).
\end{aligned}$$

これから，$E(y|p\leq y\leq q)$ は，

$$E(y|p\leq y\leq q)=\mu-\sigma\frac{f(d)-f(c)}{F(d)-F(c)},$$

となる．また，同様な計算により $E(y^2|p\leq y\leq q)$ は，

$$E(y^2|p\leq y\leq q)=\left\{\mu-\sigma\frac{f(d)-f(c)}{F(d)-F(c)}\right\}^2+\sigma^2\left\{1-\left(\frac{f(d)-f(c)}{F(d)-F(c)}\right)^2-\frac{df(d)-cf(c)}{F(d)-F(c)}\right\},$$

となる．いま，$y(\sim N(\mu,\sigma^2))$ を，$y_{ij}^{(t)}(\sim N((\mu^{(t)}+\beta_i^{(t)}+\gamma_j^{(t)}),\sigma^{2(t)}))$ に置き換えて，$p=\log(l_{ij}^*/b_{ij})$，$q=\log(u_{ij}^*/b_{ij})$ とすると，$E(y|p\leq y\leq q)$, $E(y^2|p\leq y\leq q)$ は以下のようになり，Eステップにおける $y_{ij}^{(t)}$, $y_{ij}^{2(t)}$ の計算式となる．

$$E(y_{ij}^{(t)}|\log(l_{ij}^*/b_{ij})\leq y_{ij}\leq\log(u_{ij}^*/b_{ij}))=(\mu^{(t)}+\beta_i^{(t)}+\gamma_j^{(t)})+\sigma^{(t)}\delta_{ij}^{(t)},$$

$$E(y_{ij}^{2(t)}|\log(l_{ij}^*/b_{ij})\leq y_{ij}\leq\log(u_{ij}^*/b_{ij}))=\{(\mu^{(t)}+\beta_i^{(t)}+\gamma_j^{(t)})+\sigma^{(t)}\delta_{ij}^{(t)}\}^2+\sigma^{2(t)}(1-\xi_{ij}^{(t)}),$$

$$\delta_{ij}^{(t)}=-\frac{f(d_{ij}^{(t)})-f(c_{ij}^{(t)})}{F(d_{ij}^{(t)})-F(c_{ij}^{(t)})},$$

$$\xi_{ij}^{(t)}=\delta_{ij}^{(t)2}+\frac{d_{ij}^{(t)}f(d_{ij}^{(t)})-c_{ij}^{(t)}f(c_{ij}^{(t)})}{F(d_{ij}^{(t)})-F(c_{ij}^{(t)})},$$

$$c_{ij}^{(t)}=\frac{\log(l_{ij}^*/b_{ij})-(\mu^{(t)}+\beta_i^{(t)}+\gamma_j^{(t)})}{\sigma^{(t)}},$$

$$d_{ij}^{(t)}=\frac{\log(u_{ij}^*/b_{ij})-(\mu^{(t)}+\beta_i^{(t)}+\gamma_j^{(t)})}{\sigma^{(t)}}.$$