松 山 大 学 論 集
第 23 巻 第 3 号 抜 刷
2011 年 8 月 発 行

モンテカルロ・シミュレーションによる予測の精緻化に関する数理モデル

檀　　裕　也

モンテカルロ・シミュレーションによる予測の精緻化に関する数理モデル

檀　　裕　也

1　はじめに

乱数を用いたモンテカルロ法によるシミュレーション［5］は，でたらめな乱数列を用いるにもかかわらず，例えば円周率の計算（定積分の数値計算）など，計算コストをかけることで正確な近似値を求めることができる。このような取り組みは，コンピュータによる繰り返しの高速計算と自然な乱数列の生成アルゴリズムによって可能になった。また，偶発現象の解析やマルチエージェントシミュレーションなど多くの人工社会において，モンテカルロ法が取り入れられ，その結果として解析的な取り扱いの難しい問題に対して有効な近似解を求めることができるようになった。現在，モンテカルロ・シミュレーションは，さまざまな応用範囲を持っているといえる。

本稿では，乱数を用いたモンテカルロ法によるシミュレーションによって，社会調査などで広く利用されている無作為抽出法に基づく標本調査の偏向（バイアス）を修正し，推定精度の向上および復元に関する手法を提案したい。そのための数学的評価について述べ，提案手法の有効性について数理モデルの提示とともに，計算機実験によるシミュレーションで定量的な評価を行う。

本稿の構成は以下の通りである：まず，第２章で標本調査の問題点について，回答データの欠損によるバイアスの発生という観点から指摘する。そして，第３章で標本調査の統計理論について無作為抽出法（ランダムサンプリング）を中心に概観し，その数学的な結果を述べる。また，第４章で回答の欠損

によるバイアスの発生について数学的に評価する。続いて，第5章では，標本調査のバイアスを修正し，推定精度を向上させるために提案する数理モデルを述べる。その後，第6章で，計算機実験によるシミュレーションを行い，提案手法の有効性を確認する。最後に，第7章で本稿をまとめる。

2 問題の背景

2.1 標本調査

ある地域に在住する特定の属性（年代や性別など）を対象とした市場調査を行う場合，例えば対象となるすべての人にアンケート方式などの調査をすれば，知りたい情報を手に入れることができる。このような全数調査[1)]の手法は，国勢調査などの限られた調査で採用されている。

しかし，全数調査の手法で回答を収集および集計するには，一般に膨大な時間と費用がかかるため，その費用対効果を考えると現実的ではない。

そこで，調査の対象となる属性の集合を母集団と捉え，その中から無作為に抽出された一部の標本（サンプル）を対象に限定的な調査［2］を行うことで，一定の統計的誤差[2)]は伴うものの，母集団の統計的代表値を推定することができる。このような標本調査[3)]は，マーケティングの分野におけるブランド志向などの市場調査やテレビ視聴者を対象とする視聴率調査，内閣支持率や各種選挙における投票意向など有権者を対象とする世論調査などで採用されている。

2.2 無作為抽出

標本調査の統計理論によると，標本の選び方について無作為に抽出すること[4)]が本質的に重要である。その仮定の下で標本調査の統計的誤差が決まり，その精度によって意思決定をすることになる。

1）全数調査 complete survey
2）標本誤差 sampling error
3）標本調査 sample survey
4）無作為抽出 random sampling

現在では，乱数表を用いた標本抽出のほかに，コンピュータの疑似乱数を用いた抽出方法であっても実用的である。例えば，1998 年に登場したメルセンヌ・ツイスタ [6] のアルゴリズムによって，乱数性は飛躍的に向上している。その上，電気ノイズによる乱数発生器だけでなく，物理的に乱数列を生成する量子デバイスも登場しており，本来の意味での乱数を取得することが技術的に可能となった。したがって，標本を無作為に抽出することは容易であるかのように思われる。

ところが，標本の候補が無作為に抽出されたとしても，アンケートなどの回答をそのまま回収することは案外難しい。実は，アンケートの方法によっては，アンケートに回答する層とアンケートに回答しない層が分離し，集計結果にバイアスがかかるのである。具体的な事例として，郵送によるアンケート方式を検討してみると，回答を返送する層と返送しない層で分離する。標本の候補が無作為に選ばれているとしても，集計結果には回答を返送した層というバイアスがかかることになる。

同様の問題は，街頭における調査や電話による調査，インターネットによる調査でも発生する。街頭における調査では，その時間および場所に存在すること，そして声をかけられて調査に協力する層というバイアスがかかる。また，電話による調査では，その時間に存在すること，電話を受けることに加えて，例えば地域限定で調査する場合には固定電話を持っている層という限定されたバイアスがかかる。さらに，インターネットによる調査では，そもそもインターネットを使わない層は，はじめから調査の対象外となる。

つまり，これらの一般的な調査では収集された回答は，無作為抽出にはなっていないのである。一般に，回答率が 100％でないアンケートから母集団の統計的性質を推定することは，統計的には誤った解釈を生み出すことになる。

2.3　先行研究

単純な無作為抽出に基づく単純な標本調査法には結果の精度において一定の

限界があることは，半世紀以上も前から指摘［7］されており，統計学［9］の分野では，さまざまな手法が問題解決のために提案されてきた。例えば，計算機指向型手法としてジャックナイフ法[5]［8］やブートストラップ法[6]［3］などの改良アルゴリズムが提案されている。ジャックナイフ法は，標本集団から再抽出[7]において重複を許さず，いくつかのデータを抜いた状態から標本を生成するという特徴があり，任意の統計量に対して誤差が計算できる。また，ブートストラップ法は，標本集団から再抽出を繰り返して母集団の統計的性質を推定する手法で，精度の向上を図っている。

3　標本調査の統計理論

いま N 件の母集団から無作為に抽出された n 件の標本について考える。n 件の標本データ $x_1, x_2, \ldots, x_n$ に対し，標本平均 $\overline{x}$ は

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j \tag{1}$$

と定義される。この標本調査について j 件目の標本データがある項目に該当するときは $x_j = 1$ と表し，該当しないときは $x_j = 0$ と表すことにすると，標本平均は比率を意味することになる。統計的推定の理論に基づき，標本平均 $\overline{x}$ から母集団における比率 p（母比率）を一定の精度で推定することができる。

ここで，一般の確率変数 X に対し，次のように期待値 $E(X)$ と分散 $V(X)$ を導入する：

$$E(X) = \sum_{x \in \{0,\ 1\}} xP(x) \tag{2}$$

および

5）ジャックナイフ法 jackknife method
6）ブートストラップ法 bootstrap method
7）再抽出 resampling

$$V(X)=\sum_{x\in\{0,\ 1\}}(x-p)^2P(x) \tag{3}$$

ただし，確率密度分布 P は

$$P(x)=\begin{cases}1-p & (x=0)\\ p & (x=1)\end{cases} \tag{4}$$

で与えられる。このとき，$E(\overline{x})$ は母比率 p に等しい。実際，確率変数 $x_1, x_2, \ldots, x_n$ は線形かつ互いに独立なので，

$$E(\overline{x})=\frac{1}{n}E\left(\sum_{j=1}^{n}x_j\right)=\frac{1}{n}\sum_{j=1}^{n}E(x_j) \tag{5}$$

ここで，

$$E(x_j)=\sum_{x\in\{0,\ 1\}}xP(x)=0\cdot(1-p)+1\cdot p=p \tag{6}$$

だから

$$E(\overline{x})=\frac{1}{n}\cdot p\sum_{j=1}^{n}1=\frac{1}{n}\cdot np=p \tag{7}$$

となる。また，

$$V(x_j)=\sum_{x\in\{0,\ 1\}}(x-p)^2P(x)=(0-p)^2(1-p)+(1-p)^2p=p(1-p) \tag{8}$$

より，同様にして

$$V(\overline{x})=\frac{1}{n^2}\sum_{j=1}^{n}V(x_j)=\frac{p(1-p)}{n} \tag{9}$$

を得る。

ゆえに，一般的な統計的検定で用いられる信頼度 95% で標本比率から母比率について推定を試みると

$$E(\bar{x}) \pm 1.96\sqrt{V(\bar{x})} = p \pm 1.96\sqrt{\frac{p(1-p)}{n}} \tag{10}$$

と統計的に評価することができる。

表1　標本数と精度の関係

標本数	各比率に対する精度				
	10%	20%	30%	40%	50%
50	8.3%	11.1%	12.7%	13.6%	13.9%
100	5.9%	7.8%	9.0%	9.6%	9.8%
200	4.2%	5.5%	6.4%	6.8%	6.9%
500	2.6%	3.5%	4.0%	4.3%	4.4%
1,000	1.9%	2.5%	2.8%	3.0%	3.1%
2,000	1.3%	1.8%	2.0%	2.1%	2.2%
5,000	0.8%	1.1%	1.3%	1.4%	1.4%
10,000	0.6%	0.8%	0.9%	1.0%	1.0%

以上の議論で得られた精度の評価式を典型的な標本数 n に適用すると，表1を得る。精度の評価式は，$p = 0.5$ で最大値を取るため，比率50%の列で示された精度を基準にして標本数を決めると，無作為抽出における標本調査の精度は，それを上回ることはない。

4　欠損によるバイアスの効果

N 件の母集団から無作為に抽出された n 件の標本 $x_1, x_2, \ldots, x_n$ について，その平均 $\bar{x}$ は

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j \tag{11}$$

と書ける。しかし，各標本の有効回答率，すなわち，標本の候補として選ばれた場合における回答確率 $\nu_1, \nu_2, \ldots, \nu_n$ を考慮すると，実際には

$$\overline{x}' = \frac{1}{n'} \sum_{j=1}^{n} \nu_j x_j \tag{12}$$

の値を観測することになる。ただし，

$$n' = \sum_{j=1}^{n} \nu_j \tag{13}$$

であり，各回答確率は $0 \leq \nu_j \leq 1$ である。例えば，$n = 2$ で $\nu_1 < \nu_2$ のとき，

$$\min\{x_1,\ x_2\} \leq \overline{x}' \leq \max\{x_1,\ x_2\} \tag{14}$$

の関係は保証されるが，$x_1 < x_2$ ならば $\overline{x} < \overline{x}'$ および $x_1 > x_2$ ならば $\overline{x} > \overline{x}'$ となってしまう。これが欠損によるバイアスの効果である。

5　モンテカルロ予測の数理モデル

本節では，乱数を用いたモンテカルロ法によるシミュレーションによって標本調査の精度を向上させる数理モデル（モンテカルロ予測）を提案する。

5.1　調査対象の属性と回答の表現

標本調査法は，母集団 A から無作為に抽出した標本集団 S を構成する。前節で述べた統計理論を適用するには，標本集団 S が無作為に抽出されていなければならない。しかし，抽出関数

$$\sigma : A \rightarrow S \tag{15}$$

の性質が良くても，回答の有無によってバイアスがかかる影響を避けることはできない。そこで，本節では，母集団 A と相似な補正集団 B を提案手法によって構成する数理モデルを構築する。

いま，母集団の要素 $x \in A$ の性質を K 次元の実数値ベクトルで表現する。すなわち，

$$x=(x_1,\ x_2,\ \ldots,\ x_K)\in\mathbb{R}^K \tag{16}$$

とする。通常の調査では，要素ごとに属性などの既知データと回答などの未知データが含まれている。ここで，既知データの次元を k とすると，未知データの次元は $K-k$ とできる。したがって，母集団の要素 $x\in A$ は，既知データ

$$(x_1,\ x_2,\ \ldots,\ x_k)\in\mathbb{R}^k \tag{17}$$

および未知データ

$$(x_{k+1},\ x_{k+2},\ \ldots,\ x_K)\in\mathbb{R}^{K-k} \tag{18}$$

に分割することができる。

なお，既知データとは，郵送調査における郵便番号や住所・氏名，電話調査における電話番号（市外局番・市内局番など），インターネット調査におけるドメイン名や回答送信時刻などの調査対象者の属性である。例えば，性別の属性であれば，男性を0，女性を1のように実数に変換したものを考えると，本モデルを適用することができる。

5.2 調査対象における距離空間

次に，調査対象の属性に距離を導入する。

いま，2つの調査対象 $x=(x_1,\ x_2,\ \ldots,\ x_K)\in A$ と $y=(y_1,\ y_2,\ \ldots,\ y_K)\in A$ のうち，属性について考察する。すなわち，$x=(x_1,\ x_2,\ \ldots,\ x_k)\in\phi(A)\subset\mathbb{R}^k$ と $y=(y_1,\ y_2,\ \ldots,\ y_k)\in\phi(A)\subset\mathbb{R}^k$ に対し，距離

$$d(x,\ y)=\sqrt{\sum_{i=1}^{k}\alpha_i\,(x_i-y_i)^2} \tag{19}$$

を定義する。ただし，$\alpha_1,\ \alpha_2,\ \ldots,\ \alpha_k$ は非負の定数で，各属性の重みを表現するスケール因子である。また，射影

$$\phi : x = (x_1,\ x_2,\ \ldots,\ x_K) \in \mathbb{R}^K \rightarrow (x_1,\ x_2,\ \ldots,\ x_k) \in \mathbb{R}^k \tag{20}$$

は，行列

$$\begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix} \tag{21}$$

によって表現される線形変換である。

距離 $d(x,\ y)$ は，距離関数の定義を満たす：

- $\forall x,\ y \in \mathbb{R}^k;\ d(x,\ y) \geq 0$
- $\forall x,\ y \in \mathbb{R}^k;\ d(x,\ y) = d(y,\ x)$
- $\forall x,\ y \in \mathbb{R}^k;\ x = y \Leftrightarrow d(x,\ y) = 0$
- $\forall x,\ y,\ z \in \mathbb{R}^k;\ d(x,\ z) \leq d(x,\ y) + d(y,\ z)$

したがって，$(\mathbb{R}^k,\ d)$ は距離空間となる。

実数値ベクトル空間 $\mathbb{R}^k$ の距離関数として，Euclid 距離

$$d_E(x,\ y) = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \tag{22}$$

や Euclid 距離の一般化である

$$d_G(x,\ y) = \sqrt[n]{\sum_{i=1}^{k} (x_i - y_i)^n} \tag{23}$$

や Manhattan 距離

$$d_M(x,\ y)=\sum_{i=1}^{k}|x_i-y_i| \tag{24}$$

や Chebyshev 距離

$$d_C(x,\ y)=\max_{i=1,\ \ldots,\ k}|x_i-y_i| \tag{25}$$

などがある。いずれの距離関数も提案手法で用いる $d(x,\ y)$ と同値である。$d(x,\ y)$ は，一般に馴染みのある Euclid 距離に，項目間の重み付けを加えたものである。

5.3 補正集合に付加する要素の選択

標本集団 S の要素を使って，母集団 A と相似な補正集合 B を構成する。

まず，母集団の要素 $x\in A$ に対し，次の同値集合を考える：

$$\Omega(x)=\{y\in S;\ d(x,\ y)=0\} \tag{26}$$

x の同値集合を構成するとき，x の既知データしか使わない点に注意しておく。すると，$\Omega(x)$ の要素数に応じて，次の3つに場合分けをすることができる。

(1) $\#\Omega(x)>1$ のとき

同値集合 $\Omega(x)$ の中には2件以上の要素が含まれているため，このうち1個の要素 $y\in\Omega(x)$ を無作為に抽出して補正集合 B の要素に付け加える。

(2) $\#\Omega(x)=1$ のとき

同値集合 $\Omega(x)$ の中には，ちょうど1件の要素が含まれているため，その要素 $y\in\Omega(x)$ を補正集合 B の要素に付け加える。

(3) $\#\Omega(x)=0$ のとき

同値集合 $\Omega(x)$ は空集合であるため，補正集合 B の要素に付け加えるもの

を $\Omega(x)$ から探すことはできない。そこで，標本集団 S の全要素について，要素 x との距離に応じた抽出を試みることにする。

いま，標本集団 S の全要素 $\{y^{(1)}, y^{(2)}, \ldots, y^{(n)}\}$ のそれぞれについて，要素 x との距離

$$\{d(x, y^{(1)}), d(x, y^{(2)}), \ldots, d(x, y^{(n)})\}$$

を求める。標本集団 $\{y^{(1)}, y^{(2)}, \ldots, y^{(n)}\}$ から 1 件の要素を選択するとき，単に無作為抽出をするのではなく，距離の逆数

$$\left\{\frac{1}{d(x, y^{(1)})}, \frac{1}{d(x, y^{(2)})}, \cdots, \frac{1}{d(x, y^{(n)})}\right\}$$

に比例した確率の重みを付けてルーレット式に 1 件の要素を取り出すことにする。この操作は，要素 x の属性に近いものを S の中から探すことになり，距離が近ければ近いほど乱数によって選ばれる確率が高まることを意味している。なお，逆数の計算で $d(x, y)=0$ を満たす $y \in S$ は存在しないので，上記の確率は一意に定まる。

5.4 補正集合の構成

母集団 A に属するすべての要素について，前節の操作を施すことで，N 件の要素からなる補正集合 B を構成することができる。仮に，標本集合 S に欠損データが含まれていたとしても，補正集合 B は母集団 A の縮図として，その統計的性質からバイアスを補正したことになる。

6 計 算 機 実 験

本節において，モンテカルロ予測の数理モデルに基づく計算機実験（シミュレーション）について述べる。

6.1 開発および実行環境

提案手法の有効性を検証する目的で，計算機実験によるシミュレーションを作成し，テストデータに対して実行することにした。計算機実験用のシミュレーションプログラムは，比較的規模の大きなデータを対象とするため，表2に示す開発および実行環境で動作させた。なお，疑似乱数のアルゴリズムは，標準ライブラリのものを使用した。実行プログラムは，C言語のネイティブコードとして作成し，コンパイル時における最適化オプションは標準の設定を適用した。その結果，配列の領域を最大限確保したにもかかわらず，プログラムの実行時間は1試行あたり10秒程度で収まった。

表2 実行環境

CPU	Intel Core 2 Duo T9600（2.80GHz）
メモリ	4.00Gバイト
OS	Microsoft Windows 7（64ビット）
コンパイラ	Microsoft Visual Studio 2010 / C++ Express Edition

6.2 実験の手順

まず，1件のデータあたり，1次元の既知データ（2値）と1次元の未知データ（2値）を含むエージェントを乱数を用いて生成し，母集団として65,536件の要素を作成した。母集団は，32,768件の属性Aと32,768件の属性Bに2分割される。母集団の要素における未知データは0または1の値を取ることから，その平均値は比率を表している。初期値の生成にあたって，属性Aの母比率は1/3＝0.333，属性Bの母比率は1/2＝0.500を仮定した。よって，母集団全体の母比率は5/12＝0.417となる。

次に，プログラム上で標本1,024件の無作為抽出を行う。この標本数は，母集団の1/64である。その際，属性Aの要素は1/5＝20%の確率で回答を拒否すると仮定した。そのため，単純な標本平均は，属性Bの効果が高まって本来の値より上昇すると考えられる。

最後に，提案手法であるモンテカルロ予測を用いて標本平均の補正を行った。また，統計誤差として，標本平均と同様に，標本集団のうち有効回答数を基準にしたものを流用した。

以下は，１試行あたりの実行結果である：

モンテカルロ予測のシミュレーション
母平均 ＝ 0.415985
**　属性 A ＝ 0.331970（N ＝ 32768）**
**　属性 B ＝ 0.500000（N ＝ 32768）**
標本平均 ＝ 0.426124
統計誤差 ± 0.031715
モンテカルロ補正
標本平均 ＝ 0.417862
統計誤差 ± 0.031631

この試行結果によると，無作為抽出に基づく標本調査の手法によって，標本平均として 0.426±0.032 の結果を得た。また，モンテカルロ予測に基づく補正をかけた結果，推定平均として 0.418±0.032 となったことを示している。なお，母平均の正解値は 0.416 である。

6.3　実験結果

従来手法と提案手法を比較するため，同一の母集団に対する平均の推定を両者の方法で 100 回繰り返した。

その結果，表 3 に示した実験結果（一部抜粋）を得た。

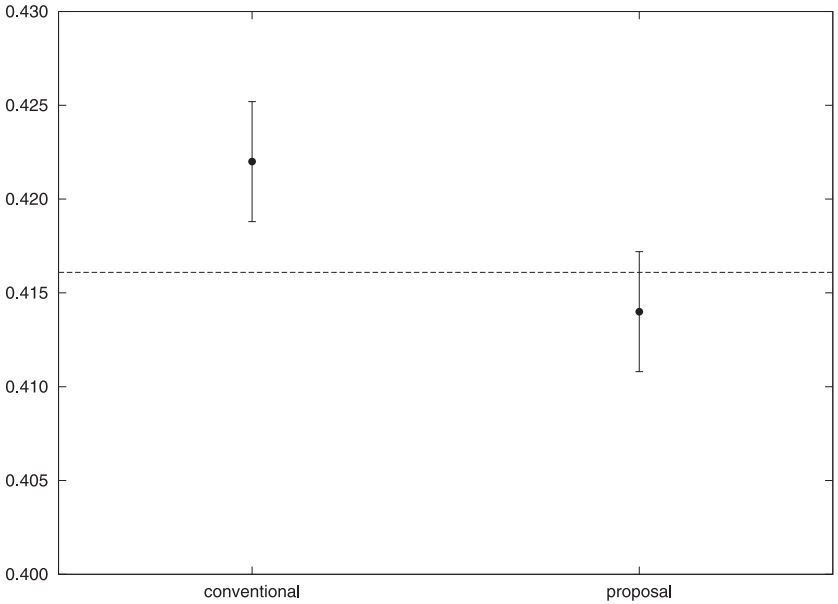
無作為抽出法に基づく標本調査では，一部の標本が回答を拒否した場合，推定される母平均は 0.422±0.003 となった。一方，本稿で提案したモンテカルロ予測に基づく補正を適用すると，推定される母平均は 0.414±0.003 である。なお，いずれの評価においても，標準偏差は $\sqrt{100} = 10$ で割ることができ

表3 計算機実験の結果

試行回数	従来手法		提案手法	
	平均値	±誤差	平均値	±誤差
1	0.420	0.032	0.417	0.032
2	0.410	0.032	0.402	0.032
3	0.430	0.032	0.413	0.032
4	0.400	0.032	0.393	0.032
5	0.447	0.032	0.439	0.032
6	0.400	0.032	0.390	0.031
7	0.413	0.032	0.409	0.032
8	0.427	0.032	0.416	0.032
9	0.428	0.032	0.420	0.032
10	0.423	0.032	0.420	0.032
11	0.424	0.032	0.411	0.032
12	0.428	0.032	0.418	0.032
13	0.418	0.032	0.410	0.032
14	0.413	0.032	0.409	0.032
15	0.426	0.032	0.412	0.032
16	0.426	0.032	0.418	0.032
17	0.445	0.032	0.430	0.032
18	0.433	0.032	0.418	0.032
19	0.439	0.032	0.431	0.032
20	0.407	0.032	0.392	0.032
…	…	…	…	…
99	0.433	0.032	0.427	0.032
100	0.407	0.032	0.394	0.032
最小値	0.380	0.031	0.379	0.031
平均値	0.422	0.032	0.414	0.032
最大値	0.460	0.032	0.454	0.032

る。従来手法と提案手法を比較すると，従来手法では回答拒否の結果として母平均を有意に上回る系統誤差が出ているのに対し，提案手法では真の平均値0.416を正しく推定できていることが分かった。

図１　実験結果



より分かりやすく表現するため，両者の結果を図１にまとめた。左側は従来手法による母平均の推定値で，統計誤差を含めて示している。また，右側は提案手法による母平均の推定値で，同じく統計誤差を含めて示している。横の破線は真の母平均を表しており，提案手法の有効性が確認できる。

以上のことから，乱数を用いたモンテカルロ法によるシミュレーションによって標本調査の精度を向上させることが可能であることが明らかになった。

7　ま　と　め

本稿では，実際の標本調査における統計理論の限界について指摘し，乱数を用いたモンテカルロ法によるシミュレーションによって，標本調査の精度を向上させる手法を提案した。その数理モデルを構築するとともに，計算機実験（シミュレーション）によって提案手法の有効性を確認した。

市場調査や世論調査などの社会調査では，時間や費用の制約から全数調査ではなく，サンプリングによる標本調査が採用されている。その中で，無作為に標本を抽出できたとしても，そのすべての標本から有効な回答が得られるとは限らない。有効回答率が100％を下回る調査では，標本調査の基礎を与えている統計理論が適用できず，理想的な状況に比べて統計的誤差が大きくなる。すなわち，完全な無作為抽出に基づく標本調査よりも精度が落ちるという問題がある。

本稿で提案したモンテカルロ予測では，母集団と標本集団の要素から既知のデータを見て，母集団と相似な補正集団を構成した。そのため，新たなコストが増えることなく，単純な標本平均に比べて母平均の推定精度を上げることができる。

しかし，その精度を無作為抽出法に基づく標本調査よりも良くすることはできない。あくまでも，バイアスの発生によって歪みが大きくなった統計誤差の精度を元に戻す方向に近づけるに過ぎない。

本稿の提案は，あらかじめ期待する精度を定めて標本調査を始めたにもかかわらず，回答拒否その他のバイアスによって調査の妥当性に疑義が発生した際に，新たな標本を追加することなく精度を補正するものである。もちろん，完全に精度が回復するかどうかは，バイアスのかかり方に依存する。

今後は，精度や計算コストについて他の手法との比較を行うとともに，実際の社会調査において提案手法を適用し，推定精度の改善を図ることが課題である。

参 考 文 献

[1] H. Akashi and H. Kumamoto, “Random sampling approach to state estimation in switching environments,” Automatica, Vol. 13, pp. 429-434. (1977)

[2] W. G. Cochran, *Sampling Techniques,* 3rd. ed. (John Wiley, 1977)

[3] B. Efron, “Bootstrap methods : another look at the jackknife,” The Annals of Statistics, Vol. 7, no. 1, pp. 1-26. (1979)

[4] A. Doucet, S. Godsill, and C. Andrieu, “On sequential Monte Carlo sampling methods for

Bayesian filtering," Statistics and Computing, Vol. 10, no. 3, pp. 197-208. (2000) doi : 10. 1023 /A : 1008935410038

[5] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller "Equation of state calculations by fast computing machines," J. Chem. Phys., Vol. 21, Iss. 6, pp. 1087-1092. (1953) doi : 10. 1063 / 1. 1699114

[6] M. Matsumoto and T. Nishimura, "Mersenne Twister : A 623-dimensionally equidistributed uniform pseudorandom number generator," ACM Trans. on Modeling and Computer Simulation, Vol. 8, No. 1, pp. 3-30. (1998)

[7] M. Quenouille, "Problems in plane sampling," The Annals of Mathematical Statistics, Vol. 20, no. 3, pp. 355-375. (1949)

[8] M. Quenouille, "Notes on bias in estimation," Biometrika, Vol. 43, no. 3 / 4, pp. 353-360. (1956)

[9] C. -E. Särndal, B. Swensson, and J. Wretman, *Model Assisted Survey Sampling.* (Springer, 2003)

[10] J. K. Tugnait, "Detection and estimation for abruptly changing systems," Automatica, Vol. 18, pp. 607-615. (1982)

疑似コード

計算機実験で用いたプログラムの疑似コード（C 言語）を以下に示す。

```
/* モンテカルロ予測のシミュレーション */

#include <stdio.h>
#include <stdlib.h>
#include <math.h>

#define N_size (1024*64) // 母集団の要素数
#define S_size 1024 // 標本の要素数

int main()
{
  int i, j; // ループ変数
  int x = 0; // 未知データの集計用
  int x1 = 0; // 未知データ（属性Ａ）
  int x2 = 0; // 未知データ（属性Ｂ）
  int n1 = 0; // 属性Ａの要素数
  int n2 = 0; // 属性Ｂの要素数
  int x_known[N_size]; // 母集団の既知データ
  int x_unknown[N_size]; // 母集団の未知データ
  int s = 0; // 標本の未知データ
  int ss = 0; // 有効回答数
  int s_known[S_size]; // 標本の既知データ
  int s_unknown[S_size]; // 標本の未知データ
  int r; // 乱数一時格納用
  double p; // 推定される比率
  double d; // 要素間の距離
  int n; // 補正集合の要素数
  int xx = 0; // 未知データの収集用
  int q[S_size]; // 補正集合の要素

  puts( "モンテカルロ予測のシミュレーション" );
  srand( 0 );

  // 母集団の生成
  for( i = 0; i < N_size; i++ ){
    x_known[i] = rand() % 2;
```

```
  if( x_known[i] ){
    x_unknown[i] = ( rand() % 3 ) % 2;
    if( x_unknown[i] )
      x1++;
    n1++;
  }
  else{
    x_unknown[i] = rand() % 2;
    if( x_unknown[i] )
      x2++;
    n2++;
  }
  if( x_unknown[i] )
    x++;
}

// 標本調査
for( i = 0; i < S_size; i++ ){
  r = rand() % N_size;
  s_known[ss] = x_known[r];
  if( s_known[ss] && rand() % 5 != 0 ){
    s_unknown[ss] = x_unknown[r];
  }
  else if( s_known[ss] ){
    // 回答拒否
    continue;
  }
  else{
    s_unknown[ss] = x_unknown[r];
  }

  if( s_unknown[ss] ){
    s++;
  }
  ss++;
}

// シミュレーション結果（標本調査）
p = (double)s / (double)ss;
printf( "母平均   = %f\n", (double)x / (double)N_size );
printf( "  属性A = %f (N = %d)\n",
         (double)x1 / (double)n1, n1 );
```

```
  printf( "  属性B = %f (N = %d)\n",
           (double)x2 / (double)n2, n2 );
  printf( "標本平均 = %f\n", p );
  printf( "統計誤差 ± %f\n",
           1.96 * sqrt( p * ( 1.0 - p ) / (double)ss ) );

  // モンテカルロ予測
  for( i = 0; i < N_size; i++ ){
    n = 0;
    for( j = 0; j < ss; j++ ){
      d = sqrt( ( x_known[i] - s_known[j] )
              * ( x_known[i] - s_known[j] ) );
      if( d < 0.5 ){
        q[n] = s_unknown[j];
        n++;
      }
    }
    if( n > 1 )
      xx += q[rand() % n];
    else{
      puts( "Empty set!" );
      exit( 1 );
    }
  }

  p = (double)xx / (double)N_size;
  puts( "モンテカルロ補正" );
  printf( "標本平均 = %f\n", p );
  printf( "統計誤差 ± %f\n",
           1.96 * sqrt( p * ( 1.0 - p ) / (double)ss ) );

  return 0;
}
```