2-3-1 データの欠測

Missing of Data

狩野 裕 今田美幸

1. はじめに

情報通信技術の急速な発展に伴い，我々を取り巻く環境においても，アンケート形式で行われるような母集団の中からある一部だけを抽出して調査する標本調査や科学的な調査に加え，インターネットやセンサなどの機器を活用して多種多様なデータが比較的容易に入手できるようになってきた．

標本調査ではターゲットの母集団を定め，母集団のどの構成要素もサンプルとして選ばれる可能性が等しい（機会の平等）ような標本抽出が前提になっている．そのための有効な方法が無作為抽出法である．しかし，実際は，回収率が低く無回答によるバイアス（選択バイアス（selection bias）の一種）や，回答はあったものの幾つかの項目に無回答（欠測値データ，欠損値データ，missing data）であることから生じるバイアスの問題がある．これらはその後の統計解析を複雑にし，複雑な分析方法をとったとしても無視できないバイアスが生じることも多い．

科学的な調査を行うには，ある特徴を持つ構成要素が選ばれにくいまたは回答しないといった状況を作ってはいけない．この意味で，機会均等性を無視して標本サイズを大きくするというアプローチは誤りである．調査しやすい構成要素が選ばれやすいということが容易に起こる．歴史的には米国大統領選挙で N=200 万の調査が予測を外し N=3,000 の調査が当選者を当てたという例もある[1]．

インターネット調査では，調査の Web ページにアクセスできる任意の人を対象とする調査もあるが，多くは，何らかの対価を支払うことを条件に事前に調査に協力できる対象者をプールした人工母集団を作っておき，その集団を調査対象として調査することが行われる．人工母集団は，目的に応じた本来の母集団（e.g., 有権者全体，大学生全体）と，性別，年齢層，職業など属性変数を整合させるなどの工夫が見られる．この方法は，標本調査論で言う割当法(用語)と呼ばれるものと近い．この方法の欠点は，例えば，男性・50 歳代・商工業地区在住を満たす有権者 50 名を抽出することになったとき，該当者の集団から標本を無作為抽出することが極めて困難であることである．割当法による標本調査は成功例もあるものの失敗もあり，近年では余り用いられない．

最近は，様々なセンサや機器が開発され莫大なデータがオンラインで瞬時に得られる．これらによるデータは，データ入力などの作業が不必要で，瞬時に分析可能なデータとなることも魅力的である．また，医療の分野でも電子カルテ化が進んでいる(注1)．企業では，膨大な売上げデータや社員の健康管理データもある．このようなセンサなどで自動収集できるデータと個人や企業のデータやその解析結果を個人／企業へフィードバックする場合はよいかもしれないが，ある集団に対する知見を得たい場合は，データが得られたプロセスを注視し，サ

本会誌では，用語は①文部省（文部科学省）学術用語集電気工学編，②本会編の改訂電子情報通信用語辞典，③本会編のエンサイクロペディアハンドブック，に基づき統一している．本稿中の「最尤」「尤度」は，上記①～③に従うと「最ゆう」「ゆう度」であるが，ここでは著者の希望により「最尤」「尤度」で掲載した．

狩野 裕 大阪大学大学院基礎工学研究科数理科学領域
E-mail kano@sigmath.es.osaka-u.ac.jp
今田美幸 正員：シニア会員 日本電信電話株式会社 NTT 未来ねっと研究所
E-mail imada.miyuki@lab.ntt.co.jp
Yutaka KANO, Nonmember (Graduate School of Engineering Science, Osaka University, Toyonaka-shi, 560-8531 Japan) and Miyuki IMADA, Senior Member (NTT Network Innovation Laboratories, NIPPON TELEGRAPH AND TELEPHONE CORPORATION, Musashino-shi, 180-8585 Japan).
電子情報通信学会誌 Vol.100 No.11 pp.1274-1279 2017 年 11 月
©電子情報通信学会 2017

(注 1) 日本の特徴として，国民皆保険制度や介護保険制度の下，活用できるデータが豊富にあることが指摘できる．このような貴重なデータを，プライバシー問題を考慮しつつどのようにしてオープン化を図るかが喫緊の課題である．

ンプルが母集団をどの程度代表するかを常に意識・検討しておかなければならない．

近年の標本代表性を軽視し大標本が強調される調査の方法は，ともすれば，先に述べた大統領選挙予測の時代（1930年代）へ退行したという指摘も聞く．しかし，筆者は，インターネット調査や現在のビッグデータの利活用は大いに進めるべきであると考えている．データ収集と分析の即時性は大きな魅力であるし，先に述べたセンサの発展によって今まで採取できなかったデータが取れるようになったことも重要な発展である．また，とりあえずはざっくりとした結果であるが速報性が有用といったこともある．本稿でお伝えしたいことは，情報技術の利活用によるデータ収集と解析においても，少なくとも伝統的な統計学の果実である選択バイアスの議論や欠測値データ解析の理論と方法を念頭に置いておくべきではないかということである．

2. 選択バイアス

選択バイアスとは，調査に参加する個人やグループといった対象（調査対象）を選ぶ際に生じる誤差である．調査対象を得た段階で，それが母集団を代表しない場合，結果が有効でなくなるため，この誤差に起因する推定量のバイアスを言う．

ある母集団において母平均μを知りたいとし，その母集団から無作為にサイズnの標本$Y_1, \cdots, Y_n$をとる．つまり，

$$Y_1, \cdots, Y_n \overset{i.i.d.}{\sim} E(Y_i)=\mu \qquad (1)$$

このとき，標本平均はμの一致推定量(用語)となる．

$$\bar{Y}_n=\frac{1}{n}\sum_{i=1}^{n}Y_i \xrightarrow{P} \mu, \quad n\to\infty$$

では，$P(R|Y)$がYに依存しているとき，標本が偏って選ばれたとすれば何が起こるだろうか．確率変数Rを，Yが選ばれる（観測される）とき$R=1$，選ばれないとき$R=0$と定義する．選択バイアスが生じる状況とは，式(1)は母集団からの無作為標本ではなく，$R=1$が与えられた下でのYの条件付き分布を持つ母集団からの標本となる．

このとき，データの標本平均$\bar{Y}_n$は$E(Y|R=1)$に確率収束し，一般に$E(Y|R=1)\neq E(Y)=\mu$であり母平均の一致推定量にならない(注2)．

例えば，大学から離れたところに実家がある大学受験生がアパート探しをする状況を考える．近年，不動産業界は競争が し烈であるため，A不動産は合格発表前にアパートの先行予約を受け付け，不合格になった場合は無料でキャンセルに応じるというキャンペーンをやっている．A不動産が持つデータによる大学合学率は受験生全体の合格率を上回る．つまり，合格率の推定値としてはバイアスがあるのである．その理由は，先行予約する受験生は一般に自信・実力があるからである(注3)．つまり，A不動産の持つデータは自信のある受験生を選択的に選んだことになっており選択バイアスが生じているからである．大学に合格（不合格）のとき$Y_i=1(0)$，A不動産で先行予約する（しない）とき$R_i=1(0)$とすると，一般に$P(Y=1)<P(Y=1|R=1)$と考えられ，YとRは独立ではないのである．

さて，RとYとの関係を表す量として以下を定義する．

$$e_i := P(R_i=1|Y_i)=E(R_i|Y_i), \quad i=1, \cdots, n$$

このとき

■ 用　語　解　説

割当法　非確率抽出法の一種．標本抽出の際幾つかの変数に関する構成比が母集団のそれと等しくなるように標本を集める方法．構成比さえ等しければどの構成要素を抽出してもよいので，そこに恣意性が入るという問題がある．

一致推定量　サンプル数をどんどん増やしていくと，ほぼ確実に，真の値を正しく当てることができる推定量．

最尤法（尤度，最尤推定量）　確率分布を規定する母数(e.g., 正規分布における平均や分散）を推定する一つの方法．得られたデータが生起する確率（密度）が最大になるように母数の値を定める推定方法で，データの同時分布を母数の関数と見たとき，それを尤度，この方法で得られた推定量を最尤推定量（MLE）という．

因子分析（共通因子，因子負荷量）　因子分析は，多次元データ（観測変数）間の相関をできるだけ少ない潜在変数で説明するための多変量解析法．その潜在変数を共通因子といい，各共通因子が説明する観測変数の変動の大きさを因子負荷量という．

直接尤度　データに欠測があるときは，存在するデータの確率分布を用いた尤度（これを直接尤度という）に基づく最尤法が用いられることがある．一般には，欠測メカニズムを付加した完全尤度を考えなければならないが，ある特定の条件の下では，直接尤度最大化でも適切な推定量を構成することができる．

EMアルゴリズム　不完全データから最尤推定値を算出するアルゴリズム．Eステップでは欠測部分を条件付期待値で置き換えデータを完全化し，Mステップでは完全化された尤度を最大化することで（一時的な）推定値を得る．このプロセスを収束するまで繰り返すことで最尤推定値を得る．因子分析モデルで登場する共通因子は直接観測されないが，これを欠測とみなして，EMアルゴリズムを適用することもできる．

(注2)　もちろんYとRが独立であれば一致推定量になるが，一般にはYとRは関係する．

$$\frac{1}{n}\sum_{i=1}^{n}\frac{P(R_i=1)}{P(R_i=1|Y_i)}Y_i$$

は，母平均 μ の一致推定量になる．選ばれた標本だけの単純な平均値はまずいが，e_i の逆数に比例した重みを付けた推定量を用いなければならないということである．つまり，選ばれにくい個体にはより重い重みを付与するということであり，この考え方は重要である．この方法を用いるためには，事前に $P(R_i=1|Y_i)$ と $P(R_i=1)$ を見積もっておかなければならない．

3. 欠　　測

標本が当初想定した形で得られないとき，欠測（missing）があると言う．まず，直接観測された変数（観測変数）が一次元の場合の欠測値問題を考える．

$$\begin{bmatrix}Y_1\\R_1\end{bmatrix},\cdots,\begin{bmatrix}Y_n\\R_n\end{bmatrix}$$

ここで，Y_i に欠測が生じ得るとし，Y_i が観測されるとき $R_i=1$，欠測するとき $R_i=0$ とする．ここでは選択バイアスはないものとし，Y_i は母集団からの無作為標本とする．

この母集団からの無作為標本の設定は 2. の選択バイアスと酷似する．存在するデータのみに基づいて推測を行うことは，$R_i=1$ のデータを使うということであり，選択バイアスの場合と同一である．違いは，ここでは，標本サイズ n が分かっており R_i のデータがあることから，$P(R=1)$ をデータから推定できることである．

次に観測変数（ベクトル）が二次元である場合を考える．以下の無作為標本を考える．

$$\begin{bmatrix}Y_1\\X_1\\R_1\end{bmatrix},\cdots,\begin{bmatrix}Y_n\\X_n\\R_n\end{bmatrix}$$

ここで，欠測は Y_i にのみ生じ得るとし，Y_i が観測されるとき $R_i=1$，欠測するとき $R_i=0$ とする．X_i は常に観測されるとする．X_i に関するこの条件は本質的でなく一般的な状況は後ほど考える．欠測があるとき，Y の母平均 $E(Y)=\mu_y$ がどのように正しく（または誤って）推定されるのか考えよう．$I_R=\{1\le i\le n|R_i=1\}$，$N_R=\#I_R$ と置く(注4)．

存在する（欠測のない完全データ）Y_i のみを用いて推測を行う方法をリストワイズ削除（listwise deletion）または完全ケース解析（complete-case analysis）と言う．すなわち

$$\frac{1}{N_R}\sum_{i\in I_R}Y_i=\frac{n}{N_R}\frac{1}{n}\sum_{i=1}^{n}R_iY_i$$

によって μ_y を推定することになる．これは先に述べたように，μ_y の一致推定量ではなく，適用してはいけない方法であることが分かっている．

そこで，Inverse Probability Weighting Estimator（IPWE）という推定量が提案されている．

$$\frac{1}{n}\sum_{i\in I_R}\frac{Y_i}{e_i}=\frac{1}{n}\sum_{i=1}^{n}\frac{R_iY_i}{e_i}$$

ここで，

$$e_i=P(R_i=1|Y_i,X_i)\qquad(2)$$

である．式（2）は傾向スコア（propensity score[2]）若しくは，欠測メカニズム（missing-data mechanism）と呼ばれている．

母集団の分布が正規分布などで仮定できるパラメトリックモデルが利用可能な場合は，最尤法(用語)が適用できる．

$$(Y_i,X_i)\sim f(y_i,x_i|\theta)$$
$$P(R_i=1|Y_i=y_i,X_i=x_i)=e(y_i,x_i|\tau,\theta)$$

としよう．ここで f と e は既知の関数で，θ と τ が未知母数である．観測ベクトルの順番を入れ換えて

$$\begin{bmatrix}Y_1\\X_1\\1\end{bmatrix},\cdots,\begin{bmatrix}Y_m\\X_m\\1\end{bmatrix},\begin{bmatrix}*\\X_{m+1}\\0\end{bmatrix},\cdots,\begin{bmatrix}*\\X_n\\0\end{bmatrix}$$

とする．ここで，$*$ は欠測を表す．このとき，尤度

$$L(\theta,\tau)=\prod_{i=1}^{m}f(Y_i,X_i|\theta)e(Y_i,X_i|\tau,\theta)\times\prod_{i=m+1}^{n}f(X_i|\theta)(1-e(X_i|\tau,\theta))$$

を最大化する母数の値 $(\hat{\theta},\hat{\tau})$ を最尤推定量とする．

最尤法を適用するには欠測メカニズム $e(\cdot)$ の項が面倒である．そこで，

$$P(R_i=0|Y_i=y_i,X_i=x_i)=P(R_i=0|X_i=x_i)\qquad(3)$$

が仮定できると，簡略化された尤度

(注3)　たとえ無料であっても，記念受験組は決して先行予約しない．
(注4)　N_R は集合 I_R に含まれる要素の数を表す．

$$L(\theta)=\prod_{i=1}^{m} f(Y_i, X_i|\theta)\times \prod_{i=m+1}^{n} f(X_i|\theta)$$

に基づく最尤推定量が一致性を有することが示される．仮定（3）を欠測メカニズムが MAR（Missing At Random）であるという．

文献（3）に倣うと，欠測メカニズムには三つのパターンがある．観測が想定されている全変数を $\boldsymbol{Y}$ と書き，$\boldsymbol{Y}$ の欠測メカニズムを $\boldsymbol{R}$ とする．$\boldsymbol{Y}$ を $\boldsymbol{R}$ に従って観測される部分 $\boldsymbol{Y}_{obs}$ と欠測部分 $\boldsymbol{Y}_{mis}$ に分ける．つまり，$\boldsymbol{R}$ の 1 の成分に対応する $\boldsymbol{Y}$ の成分を集めた確率変数の集合が $\boldsymbol{Y}_{obs}$ である．なお，$\boldsymbol{Y}=(\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$ である．このとき

- 欠測メカニズムが MCAR（Missing Completely At Random）：

$$\boldsymbol{R} \perp\!\!\!\perp \boldsymbol{Y} \qquad (4)$$

- 欠測メカニズムが MAR：

$$P(\boldsymbol{R}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})=P(\boldsymbol{R}|\boldsymbol{Y}_{obs}) \qquad (5)$$

- 欠測メカニズムが NMAR（Not Missing At Random ; or MNAR）：

$$P(\boldsymbol{R}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})\neq P(\boldsymbol{R}|\boldsymbol{Y}_{obs}) \qquad (6)$$

式（3）が MAR を示すことを確かめておこう．MAR の定義式（5）は欠測パターンごとに確認する．観測が $\boldsymbol{Y}_{obs}=(Y, X)$ のとき，$\boldsymbol{Y}_{mis}=\phi$ であり，（5）は

$$P(R=1|Y, X)=P(R=1|Y, X)$$

となる．この式の成立は自明．観測が $\boldsymbol{Y}_{obs}=X$ のとき，$\boldsymbol{Y}_{mis}=Y$ であり，定義式（5）は

$$P(R=0|Y, X)=P(R=0|X)$$

となる．この式が（3）である．

次章では，大量欠測を伴う大規模データの分析方法を紹介する．

4. 大規模 Web 調査データと欠測値問題

大規模データの解析では計算量と計算精度が重要であり，計算機科学の観点からも様々な発展がある．大規模データの特徴としてデータサイズ（volume），即時性（velocity），非等質性（variety）がうたわれるが，そこに，大規模ゆえの欠測を指摘したい．確かに，優秀なセンサやログ記録をネットワークによって蓄積していく大規模データにおいては欠測は発生しないかもしれない．しかし，例えば，販売記録データでは，ポイントカード等の導入によって購買者の個人情報を利用可能とした ID-POS データが有用であるが，意図的に個人情報を提供したくない（すなわちカードを持ちたくない）購買者も少なからず存在し，彼らについては個人情報が欠測となり，その欠測はランダムとは言えないだろう．先に述べたように，近年 Web アンケートによる質問紙調査が盛んである．本稿では，対象母集団が不明確で選択バイアスの可能性があるという Web 調査の根源的な問題は横に置き，設問の幾つかのみを選択回答するという意味で生じる欠測値問題を取り上げ，その統計分析方法を論じる．

4.1 大量欠測が生じる調査デザイン

人を評価する軸は様々である．ある評価者は勤勉さが重要と考えるかもしれないが，別の評価者は勤勉さには興味がなく社交性が大事と考えるかもしれない．勤勉さに興味のない評価者には勤勉さを問う価値はなく，興味のない質問項目は，欠測となるかランダムに選ばれた回答が得られるにすぎない．そこで，こういった状況では，数多くの評価項目を用意し，評価者に興味のある項目を選択回答させる方法が考えられる．ところが，このような調査デザインは調査後の統計分析が困難であるため実際に採用されることはほとんどなかったと思われる．このように，回答者が項目を選択しなかったという行為によってデータが欠測する場合，状況によっては大量の欠測が生じる．図 1 のデータは，実験協力者に人物刺激を与え第一印象を評定させたもので，94 項目が用意されている．94 項目の中で 6 項目は必答で，他の 88 項目はその中から 4 項目を選択回答させている[(4)]．なお，人物刺激は 4 種（4 名）であり回答者は n=8,544

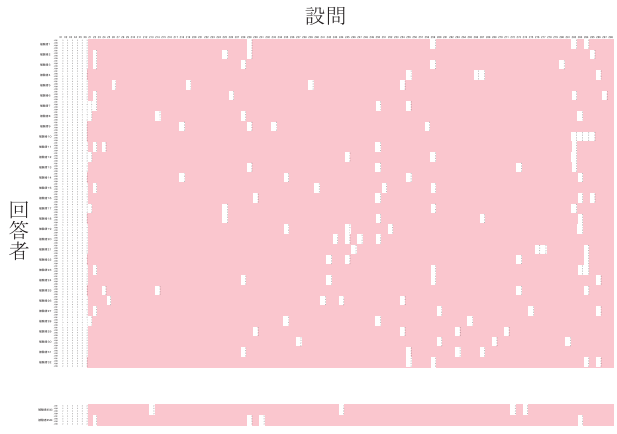

図 1　Web アンケート調査データ　　ピンクの部分が欠測．

である（調査期日：2011年12月28日～2012年1月10日）．このデザインの場合，データの約90%が欠測する．このような場合の統計的推測の方法及び推定値を効率的に求める計算アルゴリズムは容易ではない．ここでは，Hiroseほか[5]と狩野ほか[6]によって提案された方法を紹介する．

4.2 仮定と推測方法

本稿では，データの大半が欠測している場合の因子分析(用語)モデルの推定問題を扱う．ここで紹介する方法は他の統計モデルでも適用できるであろう．p次元の観測変数ベクトル$(Y_1, \cdots, Y_p)$に対してm個の潜在共通因子ベクトル$(F_1, \cdots, F_m)$を持つ因子分析モデルは

$$Y_i = \mu_i + \sum_{r=1}^{m} \lambda_{ir} F_r + \epsilon_i, \quad i = 1, \cdots, p$$

と書くことができる（e.g., 文献(7), (8)）．詳細は省くが，因子分析ではまず，因子数mの選定と因子負荷量λ_{ir}の推定を行う．欠測指標変数のベクトルを$(R_1, \cdots, R_p)$とする．$R_i = 1$は，回答者が質問項目Y_iを選択し回答することを示す．観測変数がp個の場合，欠測パターンの総数は2^pである．p個から必ずq個を選択する場合，その数は${}_pC_q$である．先に示した調査デザインでは，$p = 88, q = 4$であり，欠測パターンの数は${}_{88}C_4 = 2{,}331{,}890$となる．

回答者がY_iを選択するとき，Y_iへは真っ当な評価を行うと仮定し，すなわち，因子分析モデルが正しいとする．一方，Y_iを選択しないときは，もし強制的に回答させられたとすれば無作為に選択肢を選ぶ等，その分布は想定した因子分析モデルに従わないと考える．そういった反応を表す確率変数をW_iとする．このとき，観測変数Y_iの分布は

$$Y_i = R_i(\mu_i + \sum_{j=1}^{m} \lambda_{ij} F_j + \epsilon_i) + (1 - R_i) W_i,$$
$$i = 1, \cdots, p$$

と表現できる．ここで，次の仮定を置く．

$$(R_1, \cdots, R_p) \perp\!\!\!\perp (F_j's, \epsilon_i's, W_i's)$$
$$P(R_1, \cdots, R_p) \text{ is unrelated to } (\mu_i, \lambda_{ij}, \phi_{ii})'s$$

このとき，欠測メカニズムを含めたfull-likelihoodは

$$L = f(\boldsymbol{Y}_{obs}, \boldsymbol{R}) = f(\boldsymbol{Y}_{obs}|\boldsymbol{R}) P(\boldsymbol{R})$$
$$\propto f(\boldsymbol{Y}_{obs}|\boldsymbol{R}) = \prod_{n=1}^{N} N(\boldsymbol{Y}_{[n]}|\boldsymbol{\mu}_{[n]}, \Sigma_{[n]})$$

となる．ここで，$\boldsymbol{Y}_{[n]}$はn番目の回答者が選択し回答した質問項目から成る観測ベクトルであり，$\boldsymbol{\mu}_{[n]}$と$\Sigma_{[n]}$は$\boldsymbol{Y}_{[n]}$の平均ベクトルと分散共分散行列である．この尤度は，MARの下で適用される直接尤度(用語)（direct-likelihood or observed likelihood）と同等である(注5)．

4.3 最適化アルゴリズム

直接尤度の最適化には，擬似ニュートン法や，共通因子と欠測値を潜在変数とみなしたEMアルゴリズム(用語), (9), (10)などが用いられてきた．しかしながら，これらの手法は1970～80年代のサンプルサイズがそれほど大きくなかった頃に確立された手法であり，現在Webで取得されるような「サンプルサイズが膨大，観測変数の数が多い，欠測率＝90%が大きい」データに対しては，計算時間が非常に掛かり現実的でない．そこで，文献(5)では，共通因子のみを潜在変数とみなしたEMアルゴリズムを導出した．

提案したアルゴリズムのパフォーマンスを数値実験によって検討した．提案したアルゴリズムは，従来のEMアルゴリズムよりも数百倍程度計算速度が速いことが分かった．データの大半が欠測している場合，一般に大きなサンプルが必要となる．大量欠測時における推定精度とサンプルサイズの関係を調べたところ，90%の欠測の状況では数万程度の大きさの標本が望ましいことが分かった．（参考情報：90変数の場合，5,000人程度あれば5%以下の誤差で分析可能.）提案手法は，初対面第一印象データに適用し，対人認知構造を探ったり[11]，職場の人間関係予測[12]に応用した．

5. おわりに

データ分析を行う上で，データの欠測は無視できない．これからデータ分析を行う本稿の読者には，ここで述べた伝統的な統計手法を念頭に置き，データに欠測があった場合でも，安易に欠測のあるレコードを削除するのではなく，まず欠測の状態確認してほしい．どの変数の値がどれくらいの量，欠測しているかの状態にもよるが，欠測メカニズムを仮定することで分析できることがある．完全ケース解析の場合も，母集団に対して非常に少量のレコードを削除する場合はそれほど気にする必要はないが，大量のレコードを削除する場合はバイアスを考えてほしい．レコード削除によって，結果的に標本が偏って選ばれてしまうかもしれない．

本稿をきっかけとして，電子情報通信分野でしばしば発生するデータの欠測の問題を，統計学の研究者とのコラボレーションによって解決することで，今後の技術の進歩につながることを期待したい．

（注5） なお，ここでの欠測はNMARである．

文　　献

（1） 鈴木督久，佐藤　寧，アンケート調査の計画・分析入門，日科技連，2012.

（2） P.R. Rosenbaum and D.B. Rubin, “The central role of the propensity score in observational studies for causal effects,” Biometrika, vol. 70, no. 1. pp. 41-55, April 1983.

（3） R. Little and D. Rubin, “Statistical Analysis with Missing Data,” vol. 4, Wiley, 1987.

（4） 廣瀬　慧，金　順暎，狩野　裕，今田美幸，松尾真人，“大量欠測データに対する因子分析モデルの最尤推定，”2013 年度統計関連学会連合大会，p. 164, Sept. 2013.

（5） K. Hirose, S. Kim, Y. Kano, M. Imada, M. Yoshida, and M. Matsuo, “Full information maximum likelihood estimation in factor analysis with a large number of missing values,” J. Stat. Comput. Simul., vol. 86, no. 1, pp. 91-104, 2016.

（6） 狩野　裕，廣瀬　慧，今田美幸，松尾真人，“大量欠損データの探索的因子分析：大規模 Web アンケートデータの統計解析，”日本行動計量学会第 42 回大会発表論文抄録集，vol. 42, pp. 138-139, Sept. 2014.

（7） 柳井晴夫，前川真一，繁枡算男，市川雅教，因子分析―その理論と方法，朝倉書店，1990.

（8） 市川雅教，因子分析，朝倉書店，2010.

（9） D.B. Rubin and D.T. Thayer, “EM algorithms for ML factor analysis,” Psychometrika, vol. 47, no. 1, pp. 69-76, March 1982.

（10） D.B. Rubin and C. Liu, “Maximum likelihood estimation of factor analysis using the ECME algorithm with complete and incomplete data,” Statistica Sinica, vol. 8, pp. 729-747, 1998.

（11） 金　順暎，廣瀬　慧，今田美幸，吉田　学，松尾真人，藤井竜也，“個人属性が対人認知構造に及ぼす影響について―Web アンケートによる大規模調査の解析結果から，”信学技報，HCS2012-13, HIP2012-13, pp. 97-102, May 2012.

（12） M. Imada, K. Hirose, M. Yoshida, S. Kim, N. Toyozumi, G. Lopez, and Y. Kano, “An interpersonal sentiment quantification method applied to work relationship prediction,” NTT Technical Review, vol. 15, no. 3, 2017.

（平成 29 年 6 月 6 日受付）

狩野　裕

昭 59 阪大大学院基礎工学研究科中退．同年運輸省（当時）海技大学校助手．昭 61 阪大・工博．阪府大，筑波大，阪大・人間科学部などを経て，現在阪大大学院基礎工学研究科教授．基礎工学研究科長・学部長，阪大数理・データ科学教育研究センター副センター長兼任．統計科学，計量心理学，不完全データの解析と統計的因果推論などの研究に従事．平 25 日本統計学会研究業績賞，平 26 日本行動計量学会功績賞各受賞．

今田　美幸（正員：シニア会員）

平 2 横浜国大大学院修士課程了．同年日本電信電話株式会社入社．平 17 早大大学院博士課程修了．博士（工学）．現在 NTT 未来ねっと研究所勤務．営業行動予測，欠損を前提とした対人感情予測，プライバシー保護，ユビキタスネットワーキングサービス，フォールトトレラントシステムの研究に従事．平 27 人間情報学最優秀ポスタ賞受賞．