6L-08

Webアンケートデータを用いた分布推定における信頼性向上手法に関する一考察

佐藤真悟† 戸次陸† 岡野拓巳† 國安裕太† 杉本瑛生† 須子統太†
早稲田大学社会科学部†

１．はじめに

企業が行う市場調査など，モニターを対象にした Web アンケートが急速に普及している．しかしながら大隅[1]が指摘するように，Web 調査で得られる回答は従来の郵送調査等に比べ対象母集団を代表していない場合が多く，信頼性が低い．

本研究では，EC サイトにおいて実施された実際の Web アンケートデータに対し，標本の偏りを考慮した回答分布の推定を行う．そのもとで，単純な集計を行った場合との比較を行い，分析結果の違いについて考察を行う．

２．Web アンケートデータの収集モデル

Web アンケートのような有意抽出データの扱い方については様々な手法が提案されている[3]が，本研究ではアンケート回答に対する共変量がアンケート回答者の属性のみであるという単純なモデルを仮定し分析を行う．

あるアンケートに関して，回答者の属性が k 個あり，それぞれを，$x_1, x_2, \cdots x_k$とし，$x_i \in X_i$とする．また，アンケート項目に対する回答を$\mathrm{y} \in \mathrm{Y}$とする．$P(x_1, x_2, \cdots x_k)$を母集団における真の属性分布とすると，母集団に対するアンケート回答の分布$P(y)$は以下の式で表される．

$$P(y) = \sum_{X_1} \cdots \sum_{X_k} P(y|x_1, \cdots, x_k) P(x_1, \cdots, x_k) \quad (1)$$

いま，アンケートの回答者の属性分布が母集団の属性分布に従わず別の分布$\tilde{P}(x_1, x_2, \cdots x_k)$，に従っていると仮定する．このとき，アンケート回答者の回答分布$\tilde{P}(y)$は以下の式で表される．

$$\tilde{P}(y) = \sum_{X_1} \cdots \sum_{X_k} P(y|x_1, \cdots, x_k) \tilde{P}(x_1, \cdots, x_k) \quad (2)$$

A Study on Reliability Improvement Method for Estimation of Distribution Using Web Questionnaire Data
Shingo Sato, Riku Hetsugi, Takumi Okano, Yuta Kuniyasu, Eisei Sugimoto, Tota Suko Waseda University Social Science Study

このモデルを仮定した場合，Web アンケートにおける単純な集計結果は，真の回答分布$\mathrm{P(y)}$ではなく，$\tilde{P}(y)$から発生した標本からの推定値と捉えることができる．そのため，属性分布の違いが回答分布の推定結果に影響を及ぼしていると考えられ，この偏りを考慮した回答分布の推定が必要になることがわかる．

３．アンケート回答分布の推定方法

いま，母集団における真の属性分布である$P(x_1, x_2, \cdots x_k)$を既知とする．例えば，母集団が日本国民全体である場合であれば，国勢調査から真の属性分布を得ることが可能である．また，母集団がある企業の運営する EC サイトの顧客全体である場合は，登録している顧客情報から母集団の属性分布を得ることが可能である．このように，真の属性分布が既知である状況は多数考えられる．

このとき，(1)式より，$P(y)$の推定問題は$P(y|x_1, \cdots, x_k)$の推定問題に帰着することがわかる．よって，考えられる推定方法としてアンケートに対する回答とそれに対応する属性データをすべて用いて単純に$P(y|x_1, \cdots, x_k)$を推定すれば良い．

しかしこの場合，属性の数が増えるに従い，$P(y|x_1, \cdots, x_k)$の推定に使用できるサンプルサイズが相対的に小さくなるという問題が生じる．例えば，あるアンケートに関してN個の有効回答を得たとする．仮に全ての属性x_iの取りうる値が２値だとすると，$x_1, x_2, \cdots x_k$は2^kの組み合わせが存在する．そのため$P(y|x_1, \cdots, x_k)$の推定に利用できるサンプルサイズは平均で$N/2^k$となり，kに対し指数的にサンプルサイズが小さくなってしまう．

そこで、全ての属性を利用するのではなく，回答と独立ではない属性のみを用いて$P(y)$を推定する．まず$x_1, x_2, \cdots x_k$それぞれに対し，yと独立性の検定を行う。検定の結果独立でない変数を$x_1', x_2' \cdots x_p'$とする．そのもとで$\mathrm{P}(y|x_1', x_2' \cdots x_p')$

を推定し，これを用いて，

$$\hat{P}(y) = \sum_{x_1'} \cdots \sum_{x_p'} \hat{P}(y|x_1', x_2' \cdots x_p') P(x_1', x_2' \cdots x_p') \quad (3)$$

を求める．

4．EC サイトの Web アンケート分析

前述の手法を用いて，某 EC サイトの顧客データと顧客に対して行われた Web アンケートデータを利用し，アンケート回答の分布推定を行う．使用したデータは，2016 年 3 月 17 日～2016 年 3 月 23 日の間に行われたアンケートデータで，全顧客（母集団）のなかでアンケートに回答のあった 3,144 人分のデータを用いた．また，EC サイトにおける全顧客データは，2015 年 4 月 1 日～2016 年 3 月 31 日の間に購入履歴のあった 103,144 人分の属性データを利用した．表１に母集団と回答集団の人数と属性分布を表した．集団間の属性分布に，特に年代の面で隔たりが確認できる．

表 1：母集団・回答集団の人数と属性分布

	回答集団	母集団
合計人数	3144 人	103144 人
（性別割合）男性	31.33%	34.11%
（年代割合）10 代	3.56%	4.37%
20 代前半	9.22%	16.16%
20 代後半	15.84%	19.20%
30 代前半	20.07%	20.18%
30 代後半	19.78%	18.17%
40 代以上	31.52%	21.92%
（地域割合）北海道	4.61%	3.51%
東北	5.60%	5.84%
関東	38.39%	40.50%
中部	16.25%	15.22%
近畿	18.73%	18.58%
中国	5.44%	5.23%
四国	2.19%	2.63%
九州	8.78%	8.48%

利用したアンケートは主に個人の意識に関するものである．項目は全部で 107 件あり，そのうち自由記述の 5 項目を除いて 102 件を分析に用いた．102 項目のうち 55 項目が 2 択，47 項目が 4 択の質問であり，今回の分析では単純化のために 4 択の項目をその基準を考慮して 2 択に振り替えた．

全アンケート項目に対し，前述の方法で推定した回答「Yes」の割合と単純集計した場合の「Yes」の割合との差を求め，ヒストグラムを作成した（図１）．また，全アンケート項目に対し，独立ではない属性の組み合わせの数を表 2 に示す．

図１より，単純集計と推定結果で最大で約 6%の誤差がでる事がわかる．また，表 2 より，アンケート項目によって関連のある属性が異なることがわかる．特にアンケート回答と全属性が独立でない項目は 11 件に過ぎず，それ以外の項目では関係のある属性を絞ることで，相対的にサンプルサイズが増え，$P(y)$の推定の精度が向上していると考えられる．

表 2：アンケート回答と独立でない属性の数

独立でない属性	全推定項目に対する数
性別・地域・年代	11
性別・地域	37
地域・年代	5
性別・年代	3
性別	15
年代	23
地域	1

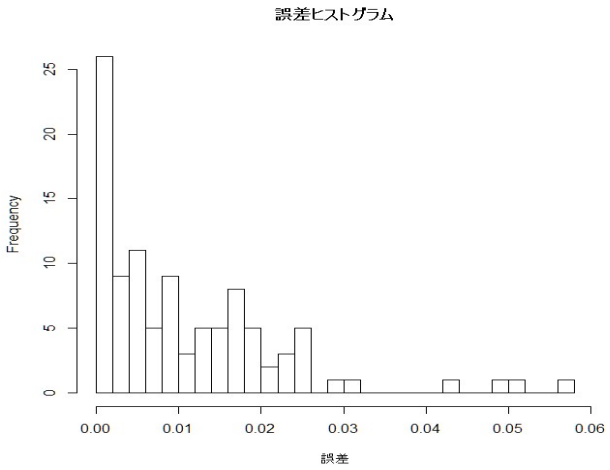

図 1：単純集計と推定結果の誤差

5．まとめ

本研究では，EC サイトで行われた実際の Web 顧客アンケートを用いて，母集団属性分布と回答者属性分布の偏りを考慮したアンケート回答分布の推定を行った．属性分布の偏りを考慮することで，単純集計では得られない回答分布が推定できることを示した．

本研究では共変量として顧客属性のみを利用したが，購買行動履歴などその他の変数を利用することで，より精度の高い推定が可能となると考えられるが，これについては今後の課題としたい．

参考文献

[1]大隅昇，”インターネット調査の適用可能性と限界”p26-35, 行動計量学第 29 巻第 1 号，2002 年
[2]石田浩ほか，”信頼できるインターネット調査法の確立に向けて”p33-47，ＳＳＪＤＡ－４２，2009 年 3 月．
[3]星野崇宏，調査観察データの統計科学～因果推論・選択バイアス・データ融合，岩波書店,2009 年.