

傾向スコアを巡る対話

鈴木 督久（日経リサーチ）

星野 崇宏（統計数理研究所）

鈴 傾向スコアが調査業界でも注目されている。発端は2000年の米大統領選挙の予測調査だ。ブッシュとゴアの接戦で予測報道が二転三転して大混乱したが、Harris InteractiveのWeb調査だけが「得票率はタイであると」予測した。「Web調査は偏りがあるので世論調査には利用できない」と考えられていたが、傾向スコア（Propensity Score）を利用することで予測に成功したという。この宣伝効果は大きかった。これが調査業界で傾向スコアが注目された背景だけれど、そもそも傾向スコアが開発されたのは疫学の分野らしいね。

星 そうですね。RosenbaumとRubinという研究者が1983年に、統計学で最も権威のある雑誌の1つであるBiometrika誌に載せた論文が最初のようにです。彼らの関心はというと、例えば、喫煙がある病気の発生に関係しているのではないかと、ということ疫学調査から調べる場合の「因果効果」の推定です。本当は、「喫煙する群としない群」（処遇＝喫煙）にランダムに割り振る（無作為割り当て）ことができれば、それから例えば10年後、それぞれのグループでの発病率を調べることで、「喫煙がそれ単独でどれくらい発病に寄与しているか」（＝因果効果）が分かるわけです。このように、研究者が処遇（ここでは喫煙）を割り当てることができる研究を実験研究といいます。

しかし人間を相手にする研究の場合、倫理的な問題や、費用がとても高くつくことから、実験研究ができない場合がほとんどです。例えばこの場合、喫煙を嫌がっている人に無理やりタバコをくわえさせることはできません。そこで疫学調査のように、既に喫煙している人と、非喫煙者での発病率を見る（＝観察研究）ことになります。

しかしそこで、大きな問題が生じます。

もともと喫煙している人は、非喫煙者よりもお酒を飲む頻度が多かったり、よりストレスフルな生活を送っている可能性は非常に大です。もし（多分そうでしょうが）お酒やストレスとその病気に関係があったら、それらの効果をごっちゃになってしまい（「交絡」という）、こういう観察研究をしても、喫煙が単独でどれくらい発病に寄与しているかという「因果効果」を知ることはできません。

そこで観察研究から何とか、実験研究でしかわからない「因果効果」を推定するための方法がいくつか提案されていますが、その中で今もっとも有力とされている方法が、この「傾向スコア解析法」という方法です。

その後、疫学だけでなく、経済学や教育の効果の研究など様々な「実験することが難しい」

分野に利用されるようになって来ています。

鈴 では、調査データへの適用は、ひとまず横において、オリジナルな場面で考えられた傾向スコアの基本的な考え方を教えてよ。

星 傾向スコアの基本的なアイデアは、「複数ある共変量を1つの値にまとめる」というものです。ここで共変量とは、先ほどの飲酒やストレスのように、発病率（結果変数）にも、喫煙するかどうか（処遇）にも関係する変数のことです。

今の例では2つの共変量しか挙げませんでした。実際は年齢・性別・他の病気など、さまざまな共変量が考えられます。もし共変量が年齢だけなら、年齢層ごとに層別して解析をすればいいかもしれません。でも疫学調査や社会科学の調査などのように、複数の共変量があると考えられる場合は、それらをいちいち層別することは不可能です。そこで、それらの共変量を用いてそれぞれの人が「喫煙者になる確率」を計算します。それが「喫煙に対する」傾向スコアです。

この傾向スコアをうまく利用すると、ある条件の下で理論的には実験研究でしか普通は知ることができない因果効果（＝もし全員が喫煙した場合の発病率－もし全員が喫煙していない場合の発病率）を推定することが可能です（傾向スコアの詳しい利用方法やその問題点などは星野・繁樹(2004)参照）。

でも実際は、どういふ変数が共変量となるかはわかりません。疫学や社会科学では、これまでの先行研究から、結果変数（発病）にも、処遇（喫煙）にも関係が「あるであろう」と推測されるものを事前にピックアップするということをしています。

自分が今仮定した共変量が有効かどうかは、データからある程度わかりますが、普通は全部調べることは、無理でしょうね。特に調査データへの適用では、どんな変数を共変量として調整に使えばいいかはまったくわかりませんので、手探りで当たりをつけていくしかありません。

鈴 共変量の選択と測定は、実際の調査研究においては難しい問題となりそうだな。疫学において因果効果の推定のために開発された傾向スコアを、世論調査・市場調査に適用してやろうと発想した人は誰だろうね。世間では、Web 調査を傾向スコアで「補正」すれば、無作為標本に近似できると宣伝されている。Harris Interactive は実際には、どのような手順を使ったのかな。

星 Harris Interactive は肝心の部分を企業秘密にしているので、基本的にはどんなことをやっているのかは想像するしかないのですが、論文(Taylor,2000,2001)やその目的から、以下のようなステップを実施している模様です。

- [0]. 大規模なオンラインパネルを構築する.
- [1]. オンラインパネルへの Web 調査と既存の（無作為抽出に近い）調査法の調整を可能にする共変量を探る研究を行う. 具体的には, 調整対象となる項目と, 事前に選定された共変量の候補となる項目についての既存調査と Web 調査どちらも実施する.
- [2]. 調整がうまく行える共変量のセットを見つける.
- [3]. 本調査としての Web 調査を行う. またここで, 本調査の中には, 調査目的に直接関係する項目だけではなく, 無作為抽出による調査と共通の項目も含める.
- [4]. 既存調査での共変量のデータと Web 調査での共変量のデータを用い, 傾向スコアを算出する.
- [5]. 傾向スコアを用いた重み付け法(**Propensity Weighting**)で Web 調査データを調整し, Web 調査の結果から既存の調査法での結果を推定する.

一度調整をうまくできる共変量を見つけたら, [1]で実施した既存の調査法のデータは複数の Web 調査の調整に「使いまわし」します ([3]から[5]だけを行う). もちろん社会情勢の変化などで既存の調査法でのデータが古くなると調整がうまく効かなくなるので, 一定期間を置いて[1]からやり直す. これが **Harris Interactive** のやっている方法だと考えられます.

鈴 米国でも企業内研究が公開されにくい状況は同じなのだな. 特許などで知的財産を法的に保護する必要性は認めるけれど, 統計学で公知の方法論の応用に関しては, 学会などのセミナーで発表する限り, 再試可能な情報を公開して欲しい気持ちもある. まあ, 結局自分でやってみればいいということなのだけれどね. ところで, **Propensity Weighting** と呼んでいる方法は, 単に傾向スコアと称している方法とは, どこに相違があるのかな.

星 傾向スコアを用いた調整法は大別して5つほどあります (星野・繁樹,2004) が, その中で, **Propensity Weighting** は先ほどの「因果効果」だけでなく, 「周辺平均」も推定できる方法の1つです. この方法を使うと, 例えば Web 調査という偏りのある集団での回答から, 「無作為抽出で選ばれた人々が Web 調査に答えた場合の平均」を推定することができるのが, 他の (マッチングなどの) 方法との違いです.

鈴 実際にやってみることが大切だ. ということで, 日経リサーチの「ブランド戦略サーベイ」への適用例をひとつ紹介してよ. 具体的な共変量は, 一応秘密ということにしておこうか. まだ研究中だから.

星 日経リサーチは日本で初めて, 傾向スコア解析のために同じ項目の訪問調査と Web 調査を実施しました. 先ほどの **Harris Interactive** の方法の[1]のところですよ. そして, 調整

に有効な様々な共変量を選び、それをもとに調整を行いました。具体的には、大手パソコンメーカーや製菓メーカーなどのブランド調査を訪問留置調査と Web 調査で行い、Web 調査の結果から訪問調査の結果を予測することに一定の成功を収めています。例えば、総合電機メーカーの「かっこよさ」を Web 調査で聞くと 55%の人が Yes と答えていますが、訪問では 29%です。傾向スコアを用いて調整すると、34.5%という推定が得られました。他の項目でも、全体として訪問調査の結果に近づいています（表1・図1・2参照）。宣伝になってしまいますが、具体的なことは星野(2003)、星野・鈴木(2004)を参照してください。

表1:「訪問調査」の結果との誤差(全項目の二乗誤差和)

	Webの誤差	調整後の誤差
パソコン	1.3453	0.7883
総合電機	3.2952	1.8729
食品	0.7691	0.6721
ネット関連	5.7308	0.8598
ファーストフードチェーン	2.3644	1.9600

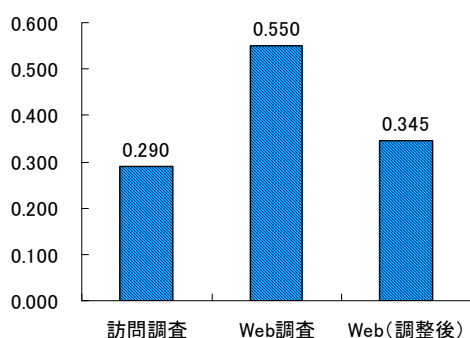


図1: 総合電機かっこよさ

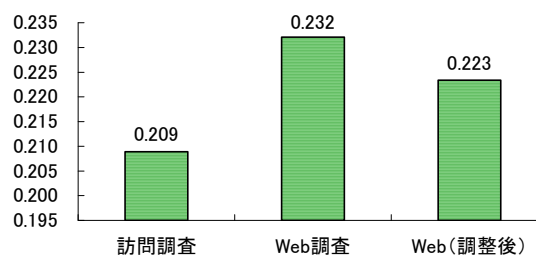


図2: パソコンメーカー情報源(評判)

鈴 Web 調査を手軽にやって、苦勞の多い無作為標本調査に近似できるなんて夢のような「いい話」だけれど、いい話にはきつと落とし穴があるよね。日経リサーチの「ブランド戦略サーベイ」においても、うまくいった項目や企業ブランドもある一方で、とんでもない結果になった「ある種の」企業ブランドもあった。仔細に見ると、そこに傾向スコアの性質の反映もあるような気がする。こういうところは「ノウハウ」になりやすいかな。

Harris Interactive の選挙予測に対する宣伝にも素朴な疑問がある。Web 調査が傾向スコアを使って近似しようとした目標は、無作為標本による調査だ。実際には RDD 調査だけれど、この近似が「成功した」という。予測に失敗した他社の調査は RDD で実施されている。だとしたら、近似なんていう傾向スコアを経由した予測より、直接に RDD 調査をやっている方が、予測に成功してもいいと思う。

それに関連するけれど、選挙結果を「真」だとすれば、RDD 調査も誤差を含み、Web 調査も誤差を含んでいる。しかもこの誤差は「標本誤差+非標本誤差」だから、RDD 調査で

さえ「真+標本誤差」というわけにはいかない。実際、「外れた」RDD 調査ではあるが、おそらく米国の調査得票率は「ナマ数字」であって、日本のマスコミのように過去の調査データも含めて解析・修正したあとの「予測得票率」ではない。Harris Interactive はそのような RDD 調査に向かって Web 調査を傾向スコアで調整したにもかかわらず、「真」に到達しえたという論理は、どこかに矛盾ないし欺瞞がないだろうか。傾向スコアが RDD 調査の非標本誤差さえも消去するとは、Harris Interactive も言わないでしょ。私たちは夢を見せられたのか。夢ならば科学的に覚醒すべきだ。

星 確かにこの問題は非常に謎です。Harris Interactive が情報をほとんど開示していないので、私もこの件について答えることはできません。ただ 1 ついえるのは、他社の選挙前の RDD 調査より、Harris Interactive の RDD 調査の方が準備時期なども含め、直前に実施していない分、代表性があったかもしれないということです。また、オンラインパネルによる Web 調査は個人抽出ですが、RDD は一般的に世帯抽出であり、選挙直前の調査など期間が非常に短い場合では大きな偏りが生じる可能性もあります。

鈴 Harris Interactive も選挙前に Web 調査だけでなく、実は RDD 調査をしていた、ということになると比較可能だね。今のところ謎のままだ。ブランド調査での研究で思ったけれど、市場調査と違って選挙予測の場合は、目的変数は得票率という 1 個の変数だ。傾向スコアを市場調査に日常的に利用するとなると、さまざまな項目を目的変数にしたいくなる。いくつか絞っても、やはり複数なければ商売にならない。けれども共変量の適性は目的項目に応じて異なるのではないかな。

星 その通りだと思います。基本的に調査目的に応じて調整に使う共変量を変える必要があるでしょう。例えば Harris Interactive では、マーケティングに使う共変量と、政治問題への意見を聞く場合に使う共変量は別になっています。実際、私と日経リサーチとの共同研究でも、目的項目や企業の業種ごとに有効な共変量を変えるという試みを行っています。1 つの目的変数だけをなるべく正確に調整するように共変量を決めても、他の目的変数には有効ではないことが多いです。従って、調整の精度と汎用性のトレードオフを考える必要があるでしょう。例えばブランド調査なら企業ごとに共変量を選択するのではなく、業種ごとに有効な共変量のグループを構成する、というのが現実的だと思います。

鈴 吉村幸氏や大隈昇氏は、「調査モード」の異なる調査間で、傾向スコアによる調整をすることに注意喚起している（吉村，2003）。これについて、どう思う？

星 調査モードの影響は 2 つに分けて考えるべきです。1 つは共変量の調査モードの影響、もう 1 つは調整したい目的項目の調査モードの影響です。前者は傾向スコア調整にあまり

影響がありません。もし共変量の上でモードの違いがあっても、それも含めて傾向スコアが算出されるからです。しかし、後者は影響を受けます。例えば Web 調査と訪問調査で同じ「趣味の数」を聞いたとしても、一般に Web 調査の方が平均的に多くなるなどという現象(吉村・大隅 2003)は傾向スコアで調整しても消えません。この意味で、吉村先生や大隅先生が仰っていることはもっともだと思いますし、調査モードの違いによる回答への影響は今後研究されるべきテーマでしょう。また、一度調査モードの違いによる効果が推定できれば、それをを用いた新たな補正も可能だと思います。

しかし発想を転換して、既存の調査法を近似しようと考えなければ、傾向スコアは使える可能性が大きいと思います。つまり「無作為抽出された人々が訪問調査に答えた場合の平均」を推定するのではなく、「無作為抽出された人々が Web 調査に答えた場合の平均」を推定することを目的とすれば、後者は問題になりません。モードの違いを重要視するかどうかは、目的をどこに置くか、の考え方の違いではないでしょうか。

鈴 なるほど、確かに目的項目については調整できないね（この点からも **Harris Interactive** の大統領選予測は謎だが、調整できないこと自体が成功因か・・・）。調査実務家は、傾向スコアを使えば「手軽な Web 調査」を「伝統的な無作為抽出標本に対する訪問調査」に向かって調整できると受けとめたようだ。

昔から、測定刺激を変更すると同一回答者集団においてさえ、異なる回答結果をもたらすことは経験的にも実験的にも数多く実証されてきた。このために「比較」という観点では、測定装置が Web であれ、電話であれ、郵送であれ、訪問であれ、その測定装置のもので「同じやり方で」繰り返し測定することが重視され実践されている。

「調査モード≒測定装置」と大雑把に理解しているのだけれど、標本の性質（非確率的であれ）が安定していれば、解釈可能性の観点から、調査モードの相違は問題ではない。ただ、Web 調査という時、測定装置の意味だけでなく、標本に関して、偏りさえもが不安定ではないかという疑いが残っていることも背景にあるのではないかな。特にオープン方式の標本。ここでも測定方法と標本抽出方法を混同せず区別して考えることが重要だ。

星野さんのように、「目的しだい」であって、選挙予測のように「予測」が純粋な目的なら、Web 調査だけであっても「可能な筈」だ。Web 調査という測定装置の性質（および利用する標本の性質も）を徹底的に知ることが予測可能性を高めていく。

ところで、標本サイズに関しては、何か知見はあるだろうか。日経リサーチの場合、実験とはいえ 1 個の調査は Web 調査も訪問調査も 200 人程度だったのだが。

星 これまでの経験からは基本的に傾向スコアを推定するために必要なサンプル数は、数百程度あれば十分と考えられます。しかし、それよりも重要なのは割合の問題です。共変量から傾向スコアを推定する際には、(先ほどのステップ[3]の) Web 調査と(ステップ[1]の) 既存調査の標本全体から算出するのですが、ここで Web 調査が多かったりすると、一

般にはあまり調整が効きません。

鈴 この他、これまでの研究も含めて、調査データにおいて傾向スコアを利用する時の問題点は？

星 やはりこの方法を使うときに最も大事なものは、共変量の選択に尽きます。これはたまたま訪問調査と Web 調査で同様の項目を聞いているから、ちょっとやってみようか、というように簡単にできる方法ではありません。きちっと計画を立てて、実験的な調査研究を何度も行い、適切な方法で共変量を選択し、どんな項目では調整はうまくいくのか、いかないのかを調べる必要があります。鈴木さんは先ほど夢のような「いい話」と仰いましたが、確かに傾向スコア調整法は一種の「マジック」に見えるかもしれません。しかしマジシャンがショーで大成功を収める裏には血のにじむ努力があるように、大きな先行投資も必要な方法であると思います。

また先ほども述べましたが、「偏った Web 標本が Web 調査に答えたときの結果」を「無作為抽出による標本が Web 調査に答えたときの結果」に調整できる可能性はありますが、既存調査そのものに調整することができるとは考えない方がいいでしょう。

文献

星野崇宏 (2003) "調査データに対する傾向スコアの適用". 品質, 第 33 巻 3 号 44-51.

星野崇宏・繁榊算男 (2004) "傾向スコア解析法による因果効果の推定と調査データの調整について". 行動計量学, 第 31 巻 1 号 43-61.

星野崇宏・鈴木督久 (2003) "傾向スコアを用いた Web 調査の無作為抽出への近似". 第 31 回行動計量学会大会発表抄録集.

大隅昇・吉村宰 (2003) 「インターネット調査を検証する一質の評価と標準化に向けて」 第 32 回 JMRA 特別研修セミナー.

Rosenbaum, P.R., & Rubin, D.B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41--55

Taylor, H. (2000) Does internet research work? *International Journal of Market Research*, 42(1) 51--53.

Taylor, H., Bremer, J., Overmeyer, C., Siegel, J.W. & Terhanian, G. (2001) The Record of Internet -based Opinion Polls in Predicting the Results of 72 Races in the November 2000 U.S. Elections. *International Journal of Market Research*, 43, 127--136.

吉村宰 (2003) "Web 調査の現状と課題—調査誤差の分類と対処の観点から—". 第 31 回日本行動計量学会大会チュートリアルセミナー.

吉村宰・大隅昇（2003）”インターネット調査の質の評価を考える”. ISM シンポジウム,
インターネット調査の現状を検証する—調査法としての評価方法と標準化をどう考える
か—.