計量政治学における因果的推論

今井　耕介

近年政治学一般における重要な傾向の一つとして、計量分析を用いた実証研究が多くなったことが挙げられる。これは、選挙や世論調査に限らず様々なデータの入手が容易になったこと、統計学という二〇世紀の初頭に始まった若い学問が電子計算技術の進歩とともに飛躍的な発展を遂げてきたことによるのであろう。また、仮説検定などの統計的原則は多様な学問分野における共通の科学的手法として受け入れられてきており、政治学においても統計的考え方を定性的研究（qualitative research）に導入しようという試みもなされてきた（King, Keohane, and Verba, 1994）。

本稿では、政治学における計量分析の重要な目的の一つである因果的推論（causal inference）について考えてみる。具体的には、まずはじめに因果的効果を統計的に定義した上で、政治学で使われている様々な研究デザイン（research design）が因果的推論を行う上でどのような役割を果たすのか、そしてそれぞれの研究デザインでどのような統計的手法が因果的効果推定に適切なのか、といった問題を著者の最近の研究をもとに論じていく。

1　因果的推論とはなにか

統計学においては、因果的推論は潜在的結果（potential outcomes）に基づいて分析されることが主流である（e.g., Holland, 1986）。例えば、二項処理変数（binary treatment variable）$T_i \in \{0, 1\}$ があるとしよう。処理群（treatment group）は $T_i = 1$ で、制御群（control group）は $T_i = 0$ である。さらに、それぞれの観察単位 $i = 1, \ldots, n$ について、二つの潜在結果変数（$Y_i(1)$, $Y_i(0)$）が定義できる。$Y_i(1)$ は処理条件下の結果を表し、$Y_i(0)$ は制御条件下の結果を意味する。すると、単位 i に関して、因果的効果（causal effect）あるいは処理効果（treatment effect）は例えば以下のように定義できる。

$$\mathrm{TE}_i \equiv Y_i(1) - Y_i(0). \tag{1}$$

しかしながら、実際に観察することのできるのはそれぞれの i に対して二つの潜在結果変数のうちの一つだけであるから、因果的推論には事実（factual）から反事実（counterfactual）を

推定することが求められる。つまり、因果的推論の難しさは、観察可能な結果変数（すなわち $Y_i \equiv T_i Y_i(1) + (1 - T_i) Y_i(0)$）から、二つの潜在的結果変数の関数である因果的効果を推定しなければならないところにある。この点は、因果的推論と他の統計的推論の本質的な違いである。

通常は、第(1)式に定義された個々の因果的効果を推定するよりも、ある有限母集団（finite population）における平均値を推定目標とすることが多い。例えば、N が母集団の大きさを表すとすると、母集団平均因果的効果（population average causal effect）は以下のように定義できる。

$$\text{PATE} \equiv \frac{1}{N} \sum_{i=1}^{N} [Y_i(1) - Y_i(0)] \qquad (2)$$

2　因果的推論のための研究デザイン

次に、政治学で使われている様々な研究デザインが第(2)式の PATE の推定とどう関連しているかを少し厳密に考えてみたい。Imai, King, and Stuart (2006) はこの問題を、推定誤差（estimation error）に関する新しい分解（decomposition）を導きだすことによって明らかにしようとしている。ここでは、その議論を紹介する。まず、ある有限母集団から得られた大きさが n の標本があると仮定しよう。一般性を失うことなく、さらに n は偶数で、標本のうち半分が処理群、残りの半分が統制群であるとすると、最も簡単な PATE の推定量は二グループの平均値の差である。I_i が単位 i が標本に含まれるか否かを示す変数とすると、この推定量は以下のように定義される。

$$D \equiv \left(\frac{1}{n/2} \sum_{i \in \{i|I_i=1,\, T_i=1\}} Y_i \right) - \left(\frac{1}{n/2} \sum_{i \in \{i|I_i=1,\, T_i=0\}} Y_i \right)$$

Imai, King, and Stuart (2006) は推定誤差 Δ、つまり D と PATE との差が、以下のように分解することを証明した。

$$\Delta \equiv \text{PATE} - D = \Delta_S + \Delta_X + \Delta_U. \qquad (3)$$

この等式において Δ_S は標本選択（sample selection）による誤差を表し、これは PATE と標本平均因果的効果（sample average causal effect; SATE $\equiv \frac{1}{n} \sum_{i=1}^{n} [Y_i(1) - Y_i(0)]$）との差に等しい（$\Delta_S$ = PATE − SATE）。それに対して、Δ_X と Δ_U は、それぞれ観察可能な処理前共変量（pre-treatment covariate）X_i と観察不可能な交絡量（confounder）U_i に関する、標本における処理群と統制群との非バランス（imbalance）による誤差を示している。

厳密に言えば、非バランスとは、処理群と統制群の経験分布の差である。例えば、$\hat{F}$ が経験累積分布関数を表すとすると、X に関する非バランスは $\hat{F}(X|T=1, I=1)$ と $\hat{F}(X|T=0, I=1)$ の差である。以上のことを前提にすると、Δ_X と Δ_U は、$Y_i(j) = g_j(X_i) + h_j(U_i)$ *for* $j = 0, 1$ という加法的モデルのもとでは、以下のように書くことができる。

$$\Delta_X = \int \frac{g_1(X) + g_0(X)}{2} d[\hat{F}(X|T=0, I=1) - \hat{F}(X|T=1, I=1)],$$

$$\Delta_U = \int \frac{h_1(U) + h_0(U)}{2} d[\hat{F}(U|T=0, I=1) - \hat{F}(U|T=1, I=1)].$$

第(3)式の分解は、政治学で使われている様々な研究デザイン

を統計的に理解するうえで便利である。まずはじめに、実験に基づく研究デザインを考えてみよう。政治心理学や政治行動学などで行われる実験の一つに**実験室実験**（laboratory experiment）がある（Kinder and Palfrey, 1993）。典型的な実験室実験では、処理は無作為化（randomization）されているが、ボランティアの学生を被験者とするなど標本選択には代表性がなく、また標本サイズも比較的小さい実験が多い。つまり、このような実験ではΔ_I, Δ_S, Δ_Eの三項ともに大きい可能性があり、処理の無作為化は仮想的な繰り返しの実験におけるΔ_IとΔ_Eの平均値がゼロに等しいことしか保証しない（$E(\Delta_I) = E(\Delta_E) = 0$）。そして$\Delta_S$に関しては、その期待値すらゼロであることはあり得ない。すなわち、実験室実験が外的妥当性（external validity）にかける由縁は、実験が実験室という特異な場所で行われることに加えて、標本の代表性が欠けていることにある。

第二に、最近経済学や政治学において注目されている、**フィールド実験**（field experiment）と呼ばれる実験室の外で行われる実験をとりあげたい（Rosenzweig and Wolpin, 2000; Green and Gerber, 2002）。フィールド実験の長所は、例えば実験室における仮想的な選挙における実験とは異なり、実際の選挙において実験を行うことで外的妥当性を高めようとするところにある。しかしフィールド実験では標本サイズは大きいことはよくあるが、倫理的あるいは財政的な理由から、特定の町や母集団における実験（e.g., Gosnell, 1927; Gerber and Green, 2000; Horiuchi, Imai, and Taniguchi 2007）や、競争的でない選挙区における実験（Wantchekon, 2003）など、標本選択には代表性が欠けている場合が少なくない。処理の無作為化と大標本によって$E(\Delta_E) = E(\Delta_S) = 0$、そして$\Delta_E \to 0, \Delta_S \to 0$が保証される一方で、$\Delta_S$は大きいままの可能性がある。さらに実験が実際の社会で行われるため、予測可能あるいは予測不可能な要因による、新たな推定誤差が生まれることが頻繁である（Imai, 2005; Horiuchi, Imai, and Taniguchi 2007）。また、この研究デザインの重要な限界は、倫理的理由等から研究可能な事象が限られてしまうことにある。政治学におけるほとんどのフィールド実験が投票率に関するものであることは偶然ではない。

フィールド実験の短所をさらに克服しようとする研究デザインが、**自然実験**（natural experiment）と**政策実験**（policy experiment）である。自然実験とは、実際の社会において処理が無作為化されたあるいは無作為化された状況に近いと考えられる事例を見つけ、分析するものである。政治学においては、カリフォルニア州の選挙で投票用紙に印刷される候補者氏名順の無作為化が義務づけられていることを利用した、投票用紙順効果（ballot order effect）の分析が行われた（Ho and Imai, 2006a,b）。一方で、政策実験とは政府などの公的機関によってよく行われる政策評価（policy evaluation）が目的の実験のことである。経済学における職業訓練に関する実験が有名であるが、政治学や公衆衛生の分野においても筆者が現在携わっているメキシコ政府による医療保険に関する政策実験などがある。こういった実験では現実味も代表性も確保され、外的妥当性は満たされるこ

とが多い。そして標本サイズも大きいのが普通であるから $b_E \to 0$ である。さらに処理の無作為化と大標本が $E(\Delta_E) = E(\Delta_C) = 0$、そして $b_E \to 0$, $b_C \to 0$ も保証してくれる。ただし、自然実験や政策実験では、フィールド実験と同じく、実際の社会で行うことに伴う新たな推定誤差が生まれることが多い。また、この研究デザインの最大の短所は、自然実験や政策実験自体が稀なことである。

最後に、実験ではなく観察に基づく研究デザインがある。それは**観察研究**（observational study）とよばれているもので、大多数の政治学における実証研究がこの方法を用いている。観察研究がよく使われる理由は、政治学の重要な研究課題のほとんどが、フィールド実験や自然実験によって検証することが不可能であることにある。観察研究は実験に基づいた研究デザインと比較して大標本であることが普通であるから、代表性の問題は少ないことが多いが（$b_E \to 0$）、処理が無作為化されていないため、処理群と統制群の非バランスが X と U の両者にわたって顕著である可能性がある。つまり、b_E と b_C は大標本であるにもかかわらず無視することができず、その期待値さえもゼロであることが保証されていない。残念ながら、U は観察不可能であるから、観察研究においては処理群と統制群の X における非バランスだけが問題であるという仮定を立てて、因果的推論が行われるのが常である。この仮定は、統計学では ignorability そして社会科学では no omitted variable と呼ばれている（厳密には、この仮定は $p(Y_i(1), Y_i(0) \mid T_i, X_i) = p(Y_i(1), Y_i(0) \mid X_i)$ と定義される）。この仮定のもとでは、標本サイズが大きければ $b_E \to 0$ であるから、様々な統計手法を用いて b_E を最小化することが分析の主目的となる。しかしながら、観察研究の最大の問題点は、観察不可能な U に関する非バランスは無視できるという仮定は楽観的な希望にすぎないことが多く、その仮定の真偽さえもデータからは直接検証不可能であるという点にある。観察研究は外的妥当性はあるが、内的妥当性（internal validity）に欠けるといわれるのは、そのためである。

こうして考えてみると、政治学で使われている様々な研究デザインもそれぞれ短所と長所があることがはっきりする。自然実験や政策実験は理想的な研究デザインであるが、そのような実験の存在自体が稀である。その他の研究デザインは推定誤差の三つの要因のうち、少なくとも一要因は研究者のコントロールが及ばないところにあるから、どれが卓越した手法であるかは一般的には断言することはできない。そしてさらにここで強調したいのは、理想的な自然実験や政策実験でさえも、実際の社会で実験が行われる限り予想可能あるいは不可能な問題に直面することは不可避だということである。従って、そのような事態に対応するための新たな統計手法の開発が重要な課題になるのである。

3　因果的推論のための統計手法

本節では、第2節で挙げられた様々な研究デザインのもとで使われる統計手法の簡単な紹介をしたい。まずはじめに、**実験**

研究の統計手法であるが、デザイン段階と分析段階において用いられるものに分けることができる。実験のデザインにはいろいろなものがあるが、社会科学の実験では、それぞれの被験者を無作為に処理群と統制群に分ける単純無作為化（simple randomization）、あるいはあらかじめ決められた数の被験者を無作為に分割する完全無作為化（complete randomization）が用いられることが圧倒的である。実際、Time-Sharing Experiments for the Social Sciences のもとで行われた実験を調べてみると、ブロッキング（blocking）やマッチング（matching）等の統計学ではよく知られた手法を用いて処理の無作為化を行った例はほぼ皆無だった。「ブロックできるものはブロックし、できないものは無作為化せよ」と言われるように（Box, Hunger, and Hunter, 1978, p.108）、ブロッキングやマッチングは無作為化の前に同じ属性をもった被験者のグループあるいは対（pair）を形成し、その中で処理の完全無作為化を行うという手法である。そうすることで、前節で詳述した分解における $\mathbb{E}$ を無作為化を行う以前にゼロに近づけることができるのである。ブロッキングやマッチングをする際には結果変数に密接に関連していることが知られている変数を用いることが大切である。Horiuchi, Imai, and Taniguchi (2007) は投票率と政策情報に関する実験の中で、性別と選挙直前の投票意思の有無をもとにブロッキングを行った。最近では、限られた標本サイズのもとで数多くの属性がある場合にどのようにマッチングを行うかの研究もさかんになされている（Greevy, Lu, Silber, and Rosenbaum, 2004）。いずれにせよ、ブロッキングやマッチングは簡単に実行可能な、正確性向上には欠かすことのできない手法であり、社会科学実験においても積極的に用いられるべきである。

自然実験や政策実験は理想的な研究デザインではあるが、フィールド実験と同様、実際の社会で行われるがゆえの複雑な問題が起きることがよくあることは前に述べた。例えば、先に述べたカリフォルニア州の投票用紙順効果の研究では、州法に定められた特異な無作為化法をどのように統計的に分析するかが重要な方法論的課題であった（Ho and Imai, 2006b）。またより一般的な問題としては、特に不遵守（noncompliance）と欠測データ（missing data）が挙げられる。不遵守とは、実験において、処理割当（treatment assignment）に従わない被験者の行動を指す。この問題に関しては、無作為化された処理割当を実際の処理変数の操作変数（instrumental variable）としてみなして、分析する手法が Angrist, Imbens, and Rubin (1996) によって提唱されている。この場合、いわゆる「処理の意図の効果」（Intention-to-Treat effect）ではなく、実験のプロトコールに従順な母集団における平均因果効果（complier average causal effect）が推定目標となる。しかし、この母集団には割当に関係なく処理を受け入れない被験者（never-taker）等の不遵守者（noncomplier）は含まれないだけでなく、母集団自体の定義が個別の実験プロトコールに左右されることから、批判も多い。例えば、Balke and Pearl (1997) は不遵守の問題のある実験のもとで、母集団平均因果効果の値域（bounds）を導出した。このアプロ

ーチは不遵守に関する仮定を一切なくして、データから直接得られる情報のみを使った場合、どこまで一般の母集団における因果効果について学ぶことができるかを問うものである（Manski, 1995）。

欠測データもフィールド実験等にはつきまとう問題である。特に、サーベイを用いて結果変数を計測するような実験ではある程度の不回答は避けられないであろう。また、長期間にわたる実験などでは、被験者のいわゆるドロップアウトによる欠損（attrition）の問題もでてくる可能性がある。このような実験における欠測データに関して最近開発された統計的手法を文献紹介程度にまとめたい。一般的に使われてきた仮定は無作為欠測（missing at random; MAR）である（Little and Rubin, 2002）。例えば、結果変数 Y が何人かの被験者に対して欠測しているとすると、この MAR 仮定のもとでは、Y の欠測確率は観察された属性 X と処理変数 T にのみ依存するということになる。これに対し、筆者は Y の欠測確率は X と（観察不可能な）Y の値そのものに依存するが、処理変数とは条件付き独立であるという新しい nonignorability（NI）仮定を提唱した（Imai, 2006b）。この NI 仮定は、政治学における選挙実験や経済学における職業訓練実験など、欠測確率が被験者の実際の投票行動や所得といった結果変数に関連する場合に最適であると考えられる。これらの手法とは別に、Horowitz and Manski（2000）は先程述べた不遵守の問題と同様のアプローチに基づいて、欠測データがある場合における因果的効果の値域を導いた。更に、不遵守の問題と欠測データの問題が同時に存在するような実験のための統計手法の開発も行われてきている。Yau and Little（2001）は MAR 仮定を不遵守の問題がある実験データに拡張した。NI 仮説の同様の拡張も既になされており（Imai, 2006b）、その他にも Frangakis and Rubin（1999）の手法などが応用研究でも幅広く活用されつつある。政治学の分野でも、Horiuchi, Imai, and Taniguchi（2007）が投票行動に関する実際の実験を例にとり、不遵守と欠測データの問題を具体的な例をあげて論じている（堀内、今井、谷口、二〇〇五も参照）。

統計学では、実験データだけでなく、因果的推測のための**観察研究の統計手法**の開発も活発になされている。政治学のみならず社会科学一般においてはパラメトリック回帰分析（parametric regression analysis）が未だに主流であるが、そのような分析が強い仮定に基づいていることは従来より知られているところである。統計学ではこの問題に対処するために様々なノンパラメトリック手法が考案されて来た。ここでは中でも最近注目を浴びているマッチング（matching）を簡潔に取り上げてみたい（詳しくは Rubin（2006）等を参照）。マッチングとは先に述べた ignorability の仮定の下で、処理群と統制群の観察可能な処理前共変量 X の非バランスをなくすために、処理群のそれぞれの被験者とそれに最も近似した統制群の被験者を対にする手法である。より一般的には対だけでなく、X の値が近い複数の被験者を処理群と統制群から選びグループを形成する手法 subclassification もマッチングの一種と考えてよい。それぞれの処

理群について、すべての$\boldsymbol{X}$の値が同じである統制群の被験者を見つける手法は厳密マッチング（exact matching）とよばれ、前節の分解における$\Delta_{\boldsymbol{X}}$を完全にゼロにすることができる。

もっとも通常の観察研究データでは、標本サイズがそこまで大きいことは稀であるから、すべての$\boldsymbol{X}$について厳密にマッチングすることは不可能である。そこで、Rosenbaum and Rubin (1983) の傾向スコア（propensity score）にもとづくマッチングがよく使われている。傾向スコアとは処理を受け取る条件付き確率$\Pr(T_i = 1|\boldsymbol{X}_i)$のことであり、ignorability の仮定の下では、スカラーである傾向スコアをコントロールできれば通常多変量である$\boldsymbol{X}$をコントロールしたことになることが証明されている。この結果に基づいて、傾向スコアをデータから推定した上で、それを用いてマッチングを行うのが傾向スコアマッチング（propensity score matching）と呼ばれる手法である。さらに、Imai and van Dyk (2004) は傾向スコアを二項処理変数だけでなく一般の処理変数に一般化し、それを傾向関数（propensity function）と呼び subclassification を用いた平均因果効果の推定方法を提唱した。近年は他にも Rosenbaum (1989) による最適マッチング（optimal matching）などの新しい手法やマッチングの統計ソフト（e.g., Ho, Imai, King, and Stuart, 2005）も開発されており、生物統計学を中心に応用研究も多数行われている。政治学においては、Imai (2005) が無作為化の失敗したフィールド実験に傾向マッチングを応用した。

このように近年注目を浴びているマッチングであるが、筆者はこれを社会科学で広く使われているパラメトリック回帰分析にとって代わる手法というよりも、ノンパラメトリックな分析前データ処理（preprocessing）手法としてとらえている（Ho, Imai, King, and Stuart 2007）。つまり、マッチングによって分析前データ処理を行い処理群と統制群を近似したグループにすることによって、パラメトリック回帰分析につきものである様々な仮定に対する敏感度（sensitivity）を減らすのである。マッチングは他の方法と異なり、結果変数を使わないので分析者による新たなバイアスが入り込む余地もなく、分析前データ処理として適している。この観点からすると、マッチングをした後に、マッチングでゼロにすることができなかった処理群と統制群間の$\boldsymbol{X}$の非バランス$\Delta_{\boldsymbol{X}}$を、通常の回帰分析をもちいてさらにコントロールすればよいということになる。

本節では、因果的推論のための統計手法を研究デザインごとに簡単に紹介した。最後に、これらの手法すべてを用いるにあたって重要な要因であるが社会科学の実証研究においては無視されがちな処理後バイアス（post-treatment bias）について述べたい。処理後バイアスとは処理後共変量（post-treatment covariate）$\boldsymbol{Z}$を誤ってコントロールすることによって生じるバイアスのことである（Rosenbaum, 1984）。この問題を直感的に理解するには、処理後共変量をコントロール変数として含んだ線形重回帰方程式$E(Y_i|T_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) = \alpha + \beta T_i + \boldsymbol{\gamma}\boldsymbol{X}_i + \boldsymbol{\delta}\boldsymbol{Z}_i$を考えるとよい。係数$\boldsymbol{\beta}$を因果的効果として解釈するためには、他の変数$\boldsymbol{X}$と$\boldsymbol{Z}$を一定に保ったうえで、処理変数$\boldsymbol{T}$を0から1に変えたと

きにどのように結果変数が変化するかを推定することになる。しかしながらこの手順の問題は、ZはXと異なり処理後共変量であるため、処理変数の値の変化によって影響を受ける可能性にある。従って、処理変数の因果的効果はβだけではなく、TがZにどのような影響を与えるか、そしてZが結果変数に因果的影響を与えるか否かに依存する。すなわち、処理後共変量をコントロールすることによってバイアスが生じる可能性があるのである。この処理後バイアスを避けるためには、回帰方程式を$E(Y_i|T_i,X_i)=\alpha+\beta T_i+\gamma X_i$と特定する必要がある。

この処理後バイアスの問題は、政治学における観察研究において、一つの重回帰方程式の複数の係数をそれぞれ因果的効果として解釈する際に多く見受けられる。また、実験研究においても似たような場合がいわゆる「死による切断 truncation by death」と呼ばれる問題である（Zhang and Rubin, 2003; Imai, 2006a）。この問題は例えば医学実験において、ある種の細胞数等への治療の効果を生存患者に限って調べるときに、処理後変数である生存率も治療によって影響を受ける可能性があることから生じる問題である。社会科学の実験においても、職業訓練の給与に与える因果的効果であるとか（この場合訓練後に就職できるかは処理後変数である）、政策情報が投票先の変更を引き起こすかどうか（この場合は投票が処理後変数となる）等を検証する際に発生する問題である。

こうしてみると、異なった研究デザインはそれぞれ別の統計手法を必要とするものの、本質的な点では色々な共通要素が多いことがよくわかる。因果的推論のための統計手法の開発は一九八〇年代以降急激に進展した。今後も現在まで考えられてこなかった問題に取り組んだり、既存の手法の向上を目的としたりする研究が次々と生まれてくることが予想される。政治学研究者もこのような最新の手法を駆使して、政治学における実証研究の精度を上げていく必要があるだろう。

4　応用統計学としての計量政治学

そもそも統計学は数学の一分野というよりも学際的な性質が大変強い学問である。特に因果的推論のための統計手法は、一九二〇年代の農業実験の分析から始まり、医学実験、さらには近年の社会実験や観察研究といった多様な実証科学研究の要求に応える形で、発展してきた。因果的推論は、計量分析を使って実証研究をする政治学者にとって学ぶべきテーマであると同時に、歴史の浅い政治学方法論（political methodology）が、計量経済学（econometrics）や生物統計学（biostatistics）のように、応用統計学の一つとしての地位を確固たるものにするための第一歩を踏み出す機会を与えてくれるはずである。本稿が、実証研究をする日本の政治学研究者にとって、因果的推論をするにあたって統計学から学べることは何かを考えるとともに、政治方法論が統計学一般にどのような貢献をできるのかという問題についても考察するきっかけになれば、と望む次第である。

〔付記〕　本研究は米国 National Science Foundation（SES-0550873）

プリンストン大学 Committee on Research in the Humanities and Social Sciences からの助成のもとで行われた。また、高橋設樹に有益なコメントをいただいた。さらに堀内勇作と福元健太郎の両氏、そして特に山本耕平氏には感謝を申し上げたい。

参考文献

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association* 91, 434, 444-455.

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association* 92, 1171-1176.

Box, G. E., Hunger, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters*. Wiley-Interscience, New York.

Frangakis, C. E. and Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatmentnoncompliance and subsequent missing outcomes. *Biometrika* 86, 2, 365-379.

Gerber, A. S. and Green, D. P. (2000). The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *American Political Science Review* 94, 653-663.

Gosnell, H. F. (1927). *Getting-Out-the-Vote: An experiment in the stimulation of voting*. University of Chicago Press, Chicago.

Green, D. P. and Gerber, A. S. (2002). *Political Science: State of the Discipline* (eds. Katznelson, I. and Milner, H. V.), vol. III, chap. Reclaiming the Experimental Tradition in Political Science, 805-832. W. W. Norton, New York.

Greevy, R., Lu, B., Silber, J. H., and Rosenbaum, P. (2004). Optimal multivariate matching before randomization. *Biostatistics* 5, 2, 263-275.

Ho, D. E. and Imai, K. (2006a). Estimating causal effects of ballot order from a randomized natural experiment: California alphabet lottery, 1978-2002. Tech. rep., Department of Politics, Princeton University.

Ho, D. E. and Imai, K. (2006b). Randomization inference with natural experiments: An analysis of ballot effects in the 2003 california recall election. *Journal of the American Statistical Association* 101, 475, 888-900.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2005). MatchIt: Nonparametric preprocessing for parametric causal inference. available at The Comprehensive R Archive Network (CRAN). http://cran.r-project.org.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* Forthcoming.

Holland, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association* 81, 945-960.

Horiuchi, Y., Imai, K., and Taniguchi, N. (2007). Designing and analyzing randomized experiments. *American Journal of Political Science* Forthcoming.

Horowitz, J. L. and Manski, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the Americal Statistical Association* 95, 449, 77-84.

Imai, K. (2005). Do get-out-the-vote calls reduce turnout?: The importance of statistical methods for field experiments. *American Political Science Review* 99, 2, 283-300.

Imai, K. (2006a). Sharp bounds on the causal effects in randomized experiments with "truncation-by-death". Tech. rep., Department of Politics, Princeton University.

Imai, K. (2006b). Statistical analysis of randomized experiments with nonignorable missing binary outcomes. Tech. rep., Department of Politics, Princeton University.

Imai, K., King, G., and Stuart, E. A. (2006). Misunderstandings among experimentalists and observationalists: Balance test fallacies in causal inference. Tech. rep., Department of Politics, Princeton University.

Imai, K. and van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association* 99, 467, 854-866.

Kinder, D. R. and Palfrey, T. R., eds. (1993). *Experimental Foundations of Political Science*. University of Michigan Press.

King, G., Keohane, R. O., and Verba, S. (1994). *Designing Social Inquiry*. Princeton University Press, Princeton, NJ.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data*. John Wiley & Sons, New York, 2nd edn.

Manski, C. F. (1995). *Identification Problems in the Social Sciences*. Harvard University Press.

Rosenbaum, P. R. (1984). Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association* 79, 565-574.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association* 84, 1024-1032.

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

Rosenzweig, M. R. and Wolpin, K. I. (2000). Natural 'natural experiments' in economics. *Journal of Economic Literature* 38, 827-74.

Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge University Press, Cambridge.

Wantchekon, L. (2003). Clientelism and voting behavior: Evidence from a field experiment in Benin. *World Politics* 55, 399 〃 422.

Yau, L. H. Y. and Little, R. J. (2001). Inference for the complier-average causal effect from longitudinal data subject to noncompliance and missing data, with application to a job training assessment for the unemployed. *Journal of the American Statistical Association* 96, 456, 1232-1244.

Zhang, J. L. and Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics* 28, 4, 353-368.

堀内勇作、今井耕介、谷口尚子（二〇〇五）「政策情報と投票参加：フィールド実験による検証」『年報政治学二〇〇五―一』一六一―一八〇頁