定量的選挙研究における結果の解釈をめぐる問題 *

矢内 勇生 †　　SONG Jaehyun ‡

2019 年 6 月 30 日

* 日本選挙学会 2019 年度総会・研究会
分科会 D（政党・議会部会）「エリート・有権者研究の実態と方法」
† 高知工科大学経済・マネジメント学群 講師/神戸大学大学院 法学研究科 研究員
E-mail:yanai.yuki@kochi-tech.ac.jp
Homepage: http://www.yukiyanai.com
‡ 早稲田大学高等研究所 講師
E-mail:jaehyun.song@aoni.waseda.jp
Homepage: http://www.jaysong.net

1　問題背景

定量的選挙研究は、一体何を目指しているのだろうか。

定量的手法を用いて書かれた選挙に関する研究論文を読むと、「〜は有意である」「〜は5% 水準で統計的に有意である」「〜は有意な負の値を示している」... といった表現が数多く見つかる。これらの表現の後に、分析の結論として「仮説は支持された」と続くことが多い。私たちはこれらの表現（結論）からどのような知見を得られるだろうか。その答えは 1 つである。「説明変数がが応答変数（結果変数）に与える影響は 0 ではない」ということである。

選挙研究（あるいはより広く政治学・社会科学）において、私たちが知りたいのはこのような結論だろうか。本当に「0 でない」ことが重要だろうか。もちろん、ゼロでないことが絶対的に重要なこともあるだろう。しかし、本当に知りたいのは、特定の変数にどれほどの効果があるのか、また、推定された結果がどれだけ確実なのかではないだろうか。

「統計的に有意である」ことは p 値が教えてくれるが、p 値は効果の大きさや結果の不確実性については教えてくれない。それにもかかわらず、p 値のみに依存して結果を解釈する論文がたくさん存在する。このような状況は国内外を問わず多くの分野に共通して観察されるものだが、近年は p 値偏重の定量的分析に対する批判が強まっている。たとえば、統計学においては Wasserstein and Lazar (2016) や Greenland et al. (2016)、生物学において Halsey (2019)、心理学においては Calin-Jageman and Cumming (2019)、大久保 (2016) などが、p 値が持つ情報量の貧弱さを示し、その代案を示してきた。

本稿は、『選挙研究』（あるいは選挙学会）における研究も p 値偏重になりすぎており、研究から得られたはずの知見の一部しか共有できていないことを指摘する。また、データ分析の結果をどのように解釈・提示すればよいかについての指針を示す。学術的な真実に迫るためには学会全体として知見を共有・蓄積することが重要である。そのためには選挙研究における結果の解釈と提示の方法を変えることが必要であり、それを達成することはそれほど困難ではないことを明らかにする。

2 『選挙研究』における定量的分析

2009 年から 2018 年までの 10 年間に刊行された『選挙研究』(25 号 1 巻から 34 号 2 巻まで) には、169 本の論文が掲載されている[1)]。そのうち、統計を使った論文は 115 を数える[2)]。 この数字をそのまま素直に受け取れば、『選挙研究』に掲載されている研究の過半数は定量的研究である。定量的分析は、日本の選挙研究において欠かせないツールであると言えそうだ。

しかし、115 本中 42 本の論文については、推測統計の手法は用いておらず、記述統計を資料として議論を進めている。記述統計のみを用いた研究を除き、統計的推定を行っている論文だけを定量的研究として数えれば、その数は 169 本中 73 本であり、全体の 4 割程度にとどまる[3)]。統計的推定を用いた 73 の定量的研究のうち、6 つの論文が因子分析・主成分分析を主に用いた研究、残りの 67 の研究が回帰分析を使った研究である[4)]。本稿は、回帰分析を用いた 67 の研究を対象として議論を進める。

これらの研究に共通する特徴は、仮説を検証するための手段として回帰分析を用いているということである。それぞれの研究に検証すべき仮説があり、それがデータとして測定可能な変数を用いた作業仮説に置き換えられる。さらに、その作業仮説を対立仮説とし、対立仮説を採択するために棄却したい帰無仮説が設定される。その上で回帰分析を実行し、特定の回帰係数に関する帰無仮説を棄却することで、自らの仮説 (対立仮説) が正しいことを示す。これが典型的な分析の流れである。

この分析の流れ自体に大きな問題はない[5)]。問題は、この後である。多くの論文では、分析がここで終わっている。つまり、帰無仮説が棄却されたこと自体が、統計分析セクションの結論になっていて、分析結果の詳しい検討があまり行われていない[6)]。本稿は、

---

1) 研究ノートを含む。ただし、書評、書評論文、講演会記録、資料は除く。

2) この他に定量的分析の方法についての論文が 3 本あるので、それらも合わせれば 118 本ある。

3) 筆者は、『選挙研究』(あるいは選挙学会) が日本の政治学の中でも「計量化」が最も進んでいる雑誌 (学会) だと思っていたので、意外と少ないという印象をもった。アメリカ合衆国の状況については、Tom Pepinsky がまとめた "Single Country Research in Comparative Politics" (https://tompepinsky.com/2018/06/15/single-country-research-in-comparative-politics/) が参考になる。

4) 後者には、因子分析・主成分分析と回帰分析を共に用いたものを含む。

5) ただし、回帰分析で使う統計モデルがはっきり説明されていないものや、データ生成過程 (data generating process) を検討せずに「使いやすい」(ソフトの使い方を知っている) モデルで推定したことが疑われるもの、複数のモデルが詳しい説明なしに並記されているものなどがあり、本来はこれらも問題にすべきかもしれない。p 値が特定の「統計モデル」に基づいて計算される以上、統計モデルを明示することが必要だろう。この点については、別の機会に論じたい。

6) もちろん例外もある。例えば、福元・中川 (2013); 上條 (2017); 松林 (2016); 中井 (2014) などは、効

これが定量的選挙研究において克服すべき重要な課題であると主張する。

帰無仮説を棄却することは、仮説検証型の定量的研究において達成したい事項の一つかもしれない。しかし、それは目的ではないだろう。目的は、自らの仮説を（検定ではなく）「検証」することであり、作業仮説レベルの対立仮説が採用されただけでは、必ずしも理論仮説を検証したとは言えないはずである。

統計分析の結果からは、帰無仮説を棄却するかどうかだけでなく、より多くの情報を得ることができる。定量的分析を統計的仮説検定だけに使うのは、もったいない[7)]。リサーチクエスチョンに適切な回答を与えるためには、それらの情報を慎重に検討することが求められる。

以下では、一般的に用いられる統計分析において特に誤解が生じやすいと思われる点について解説し、統計的仮説検定だけでは満足できない理由を示す。特に、効果量を検討することの重要性を強調し、そのために利用すべき方法をいくつか紹介する。

3　p 値の解釈についての問題

本節では、p 値のみに依存する統計的有意性検定の問題について述べる。まず、p 値に対する誤解とその正しい意味について解説し、その後で p 値のみに依存する統計的有意性検定がもたらす弊害について述べる。

3.1　p 値の意味

2016 年に出されたアメリカ統計学会 (American Statistical Association) の声明でも改めて強調された通り、p 値[8)]の意味を一言で述べるなら、「特定の統計モデルの下で、

果量などについて詳しく検討している。

7) 統計分析をしない（統計分析が嫌いな?）人から「帰無仮説を棄却するかしないかばかり検討して何が楽しいのか」と問われたことがあるが、それに対する筆者の答えは、「全然楽しくないよ」である。仮説検定は統計分析の主な結果である必要はないし、最終目的でもない。計量分析をしない人たちが計量分析に興味をもたない（嫌いになる）理由が、このような誤解にあるとすれば、それはとても不幸であると言えるだろう。

8) 初期の統計分析ソフトウェアでは「有意確率」と訳されることもあった（英語にも、“marginal significance probablity” という表現がある）が、誤解されやすい表現であり、避けるべきものだ。有意確率という言葉からイメージされるのは、特定の変数の効果が「（統計的に）有意になる確率」という意味だろう。しかし、p 値にそのような意味はない。また、p 値には、それとは反対の「有意ではない確率」という意味もない。したがって、有意確率という言葉は、多くの人に p 値の意味を誤解させる原因の一つになってきたと推測される。『選挙研究』に掲載された論文にも、「p 値」の代わりに「有意確率」という言葉を用いているものがいくつかある。p 値の意味を知らない（定量的研究を行わない）読者に配慮したのかもしれないが、上で述べたような誤解をする可能性を高めるだけだろう。それなら「p 値というなんだかよくわか

データの要約統計量が観察された値と同じか、それより極端な値を取る確率[9]」である (Wasserstein and Lazar 2016, p.131)。ここには重要なポイントが 2 つある。一つ目は、「特定の統計モデル」という箇所である。たとえば、「X が Y に与える影響は 0 である」という帰無仮説を検定するとしよう。この場合、説明変数として X のみを入れた統計モデルと、Z というもう一つの共変量を投入したモデルにおいて X の係数の p 値は異なる可能性が高い。もし、後者 $(Y = f(X, Z))$ では X の係数の p 値が小さく（例えば、$p < 0.01$）、前者 $(Y = f(X))$ ではそうでもない（例えば、$p = 0.10$）という結果が得られたとき、推定対象となる母集団において $X \to Y$ の関係は成立すると言えるだろうか。これだけでは「わからない」のというのが正しい答えだろう。p 値とは特定の統計モデルから得られたものであり、二つの異なる統計モデルから異なる p 値が得られても、p 値に基づいてモデルの優劣を決めることはできない。

第二のポイントは、「データの要約統計量が観察された値と同じか、それより極端な値を取る」という箇所である。統計的有意性検定は、帰無仮説を棄却するかどうかを判断するための検定である。したがって、「効果がある」ことを示すために設定される「効果がない（ゼロ）」という帰無仮説を用いると、要約統計量（t 値や z 値などの検定統計量）が 0 からどれだけ離れているかに注目することになる。この状況で、p 値の意味を「帰無仮説が正しい確率」と考えがちだが、これ誤解である。確率分布に従う統計量を使って検定を行っている以上、実際に帰無仮説が正しくても（つまり、帰無仮説が正しい確率が 1 でも）、p 値が小さい値（例えば、$p = 0.01$）をとることはあり得る。p 値は、帰無仮説を含む統計モデルとデータの整合性を示しているが、それはあくまで確率であり、不確性が内在している[10]ことを忘れてはならない。

p 値の意味を正しく捉えれば、特定の変数の p 値が小さいからと言って、その変数が結果に対して重要であるとは限らないということも理解できるだろう。小さな p 値は、仮定した統計モデルと観測データの整合性が低いことを示すが、それ以上のことは示さない。実質的には無視しても差し支えないほどに小さな効果であっても、効果が 0 だという仮説との整合性が十分低ければ、p 値は小さくなる (浅野・矢内 2018, pp. 165–168)。したがって、p 値から変数の重要性を判断することはできないので、効果量についての検討が必要になる[11]。また、p 値を計算するために必要な仮定は帰無仮説だけではないので、仮

らないものを使っている」と思われる方がまだマシだろう。有意確率ではなく、p 値と書くべきである。

9) “[A] p-value is the probability under a specified statistical model that a statistical summary of the data would be equal to or more extreme than its observed value.”

10) つまり、データ生成を非常に大きな回数にわたって繰り返した場合の頻度を表しているだけで、データ生成が 1 度しか行われていない状況で「正しい」か「正しくない」かは判断できない。

11) この点については後で詳しく述べる。

定した統計モデルとデータとの整合性が低い（つまり、p 値が小さい）理由は、帰無仮説が間違っていること以外にもあり得る。

また、p 値が十分小さいとは言えず、帰無仮説が棄却できない場合には、係数が 0 である（効果がない、帰無仮説が正しい）」という主張はできない[12)]。「効果があるとは言えない（効果が確認できない）」という表現が正しい[13)]。$p = 1$ になっていない限り、帰無仮説以外に観測データとより整合的な仮説（例えば、効果が 0.001 など）が存在するはずなので、効果がないという証拠としては使えない。

p 値の正しい意味が分かったとしても、その用途を p 値が有意水準 (α) より小さいか否かで統計的有意性を確認することだけに限定すれば、ここまで述べてきたような誤用は避けられるかもしれない。しかし、α の値が恣意的な基準であることを忘れるべきではない。$\alpha = 0.05$ という基準を提唱したのは Fisher (1925, p.44) であるが、その理由は「便利だから[14)]」である。この便利さに科学的根拠はない。何一つ科学的根拠のない基準で「有意性」を測ることが適切かは、問題に応じて個別に検討すべき問題であり、「慣習だから」という理由で正当化されるものではない。そのような検討を伴わない研究が量産された結果、アメリカ統計学会の声明が出されたり (Wasserstein and Lazar 2016)、*Basic and Applied Social Psychology* における p 値掲載禁止 (Trafimow and Marks 2015) の措置が取られたりしたものと考えられる。

3.2 Fisher と Neyman–Pearson の p 値

周知のように、p 値は Fisher (1922) によって考案されたものである。ただし、Fisher は統計的有意性検定を目的に p 値を考案したのではなく、エビデンスの「強弱」の指標として用いることを想定していた。p 値と仮説検定が結びつくようになった契機は、Neyman and Pearson (1933) の研究である。Fihser の p 値は、値の大小を解釈する一方、Neyman–Pearson の p 値は、$p < \alpha$ か否かが重要であり、具体的な数値には意味を与えないという点で対立している。しかし、多くの論文ではこの 2 つの考え方が混在している。たとえば、小野 (2017) は $p < 0.001$ の結果から「有意性が高い (p. 49)」と解釈する一方、「(...) 独立変数の係数はいずれも統計的に有意だった ($p < .001$) (p. 48)」と述

12) この意味で実験データ、マッチング後データのバランスチェックに統計的有意性検定を行うことも適切でない。そもそも、統計的有意性検定の対象は母集団のだが、バランスチェックはサンプル内のバランス達成が目的である。詳細は Imai et al. (2008) を参照されたい。

13) 『選挙研究』に掲載された論文では、この点について誤解しているものはほとんどなかった

14) "The value for which P = .05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not."

べる。他にも一つの統計モデルを解釈する際に、係数ごとに異なる有意水準 (α) を使う研究もある[15)]。これらの論文の共通点は、特定の α の値を想定した Neyman–Pearson 流の検定を前提としつつ、p 値が小さい方が「良い」結果だという Fisher 流の考え方も併せて用いていることである。

Neyman–Pearson 流の統計的有意性検定は、具体的に言えば、統計的有意性判定 (judgment) である。そのためには、判定の閾値が必要である。そこで有意水準 (α) と検出力 ($1-\beta$) を設定する必要があり、社会科学ではそれぞれ、0.05、0.8 に設定することが一般的である。これらの 2 つの数値と想定される効果量と母集団の分散に基づいてサンプルサイズを計算し、それによって決められたサンプルサイズで p 値を計算することが「正しい」手順である。ここで重要なのは、サンプルサイズが予め（分析の前に）決まっていることである。つまり、p 値を判断基準として使うためには、サンプルしサイズを事前に定めることが必要である[16)]。決められたサイズよりも大きいサンプルを使う場合、p 値はもはや有意性判定の判断材料にはならない。たとえば、二群の平均値の差が 5 以上なら実質的に効果があると解釈でき、それぞれの群の標準偏差が 20 である場合、必要なサンプルサイズはそれぞれの群で、約 251 である[17)]。p 値から統計的有意性を判定するためには、$n = 500$ のデータを用いるべきである[18)]。もし、$n > 500$ の場合、推定結果から $p < 0.05$ の統計量が得られたとしても、統計的に有意であると判定できない。

選挙研究では、分析単位が有権者の場合にサンプルサイズは大きくなり、議会や国になると小さくなる。また、事前にサンプルサイズを決めたり、検出力分析が行われることは非常に稀である。これは日本の選挙研究に限った話ではなく、国内外を問わず、政治学一般に見られる状況である。そもそも、母集団における標準偏差などが未知である場合が多く、サンプルサイズを「決める」のが困難なケースが多いのも事実である。また、実質的に有意とみなす基準も恣意的なものになり得る。正しい手順で行われる定量的研究が蓄積していくと、メタ分析が可能となり、この問題も一定程度解決されると考えられる。

15) 例えば、「X の係数は 5% 水準で、Z は 1% 水準で統計的有意」などの表現である。

16) 上で述べた「特定の統計モデル」には、帰無仮説だけでなく、この予め決められたサンプルサイズも含まれる

17) 両側検定を行う場合

18) これは同じ手順でサンプリング、分析を無限回行う場合、効果量が 5 なら、分析結果の 80% において $p < 0.05$ の結果が得られることを意味する。

3.3　効果量と不確実性への無関心

ここまで述べたように、p 値は誤用されやすい。厳密に考えると、適切なサンプルサイズのもとで計算されていない p 値は統計的有意性判定の材料にもならない。他にも p 値が提供する情報は、統計分析全体から得られる情報に比して非常に少ないことも自覚すべきである。

p 値によって判断することができない重要な要素の一つに、効果量 (effect size) がある。既に述べたとおり、『選挙研究』に掲載された定量的分析の多くが、p 値と統計的有意性の判定に囚われており、効果量を議論していない。Calin-Jageman and Cumming (2019) は、効果量が「新しい統計学」の重要な構成要素だと主張する[19)]。効果量を示すというのは、ある説明変数が応答変数（結果変数）に与える影響の大きさを数値として示す[20)]ということである。

繰り返しになるが、原因 X が結果 Y に与える影響を推定した結果から得られた p 値（の小ささ）は、その影響力の大きさを示す訳ではない。影響力の大きさは、効果量によって示される。p 値とは異なり、効果量は帰無仮説が正しくない程度を量的に表す指標である (大久保・岡田 2012)。他の条件が同じなら、効果量が大きい場合、p 値は小さくなる。しかし、p 値が小さくなる条件は大きい効果量以外にも、分散が小さいこと、サンプルサイズが大きいことが挙げられる。これは、効果量が大きくても、p 値が大きくなる可能性があることを示唆する。逆に、非常に微々たる効果量であっても、サンプルサイズによっては小さな p 値が得られる可能性もある。

選挙研究に限らず、あらゆる研究分野において、効果量の議論は極めて重要である。血糖値 1mg/DL を下げる 100 万円の薬を購入する意味がないように、投票率 0.1% ポイントを上げるために数億円を投資することは費用–便益の観点からは望ましくないだろう。帰無仮説を棄却し、「効果がゼロでない」ことを示すだけでなく、推定された効果が実質的に意味があるかどうかまで検討してはじめて仮説を「検証」したと言えるだろう (浅野・矢内 2018, pp. 165–168)。効果量の議論は、学問の実践や社会貢献のためには必要不可欠な要素である。

効果量を議論する際に欠かせないのが、その効果の不確実性 (uncertainty) である。残念ながら、p 値は不確実性の指標でもない[21)]。統計的推定には不確実性が常に存在する。

---

19) 効果量以外の要素は、信頼区間、メタ分析、オープンサイエンスである。

20) 影響を与える/与えないだけではなく、どれほど与えるかを定量的に示す。

21) むろん、真の効果量が 0 でない限り、不確実性と p 値は比例するだろう。不確実性が小さいことは標準

推定から得られた因果効果の点推定値（平均値の差分、処置変数の係数など）は確実なものではなく、もう一度同じ条件でデータの収集と分析を行えば、異なる推定値が得られるだろう。その結果、効果量が 0 に近付くこともあれば、0 から離れるより大きな効果量を示すこともあるだろう。

不確実性の指標として主に使われるのは、信頼区間 (confidence interval) である。信頼区間の厳密な意味は本稿の後半に述べるが、区間内の数値が真の値であっても、今回の結果は驚く結果でないことを意味する[22)]。たとえば、投票率という結果に影響を与える要因として、投票時間延長の効果の点推定値が 3% ポイントでその信頼区間が [0.5, 5.5] であり、投票所増設の効果の点推定値が 2% で、信頼区間が [1.0, 3.0] だとしよう。どちらも同じ費用を伴い、片方の政策しか採用できないとき、どちらを採用すべきだろうか。この場合、効果の点推定値だけでなく、不確実性も慎重に考慮に入れるべきである。しかし、このような不確実性の大きさと、その具体的な数値は p 値からは分からない。

図 1 のケース（1）と（2）を比べてみると、不確実性が小さいのは（1）であるが、p 値が小さいのは（2）である[23)]。また、ケース（3）と（4）を比べると、p 値が必ずしも効果量を示さないことが分かるだろう。これらの 4 つのケース（係数）を p 値のみを用いて比較すると、最も p 値が小さいケース 3 が最も重要な結果であると「誤解」してしまうかもしれない。しかし、ケース 3 の効果量は他のケースに比べて小さく、同じ単位で測定されているとすれば、結果に与える実質的なインパクトは最も小さい（したがって、重要性も低い）はずである。政策評価の場面でより高い評価が可能な結果はどれだろうか。不確実性の大きさを勘案してもケース（4）であろう。

このように、社会科学における定量的研究の場合、効果量と不確実性に関する議論は不可欠であるものの、p 値はこれらの情報を読者に教えてくれない。『選挙研究』では、効果量だけでなく、推定の不確実性も軽視されていると言わざるを得ない。推定の不確実性を示すために信頼区間を用いている論文は、検討した 10 年間で 11 本しかない。また、推定値と p 値の掲載はあるが、不確実性の指標であり、信頼区間を作るために必要な標準誤差が掲載されていない論文も多い[24)]。

誤差が小さいことでもあり、自然に p 値も小さくなる。

22) 近年、confidence interval を compatibility interval という用語で置き換えることが提案されている。

23) p 値が同じなら係数の大きさと標準誤差の大きさは比例するからである。

24) 中には、p 値すら載せず、推定値に星が付けられているだけのものもある。

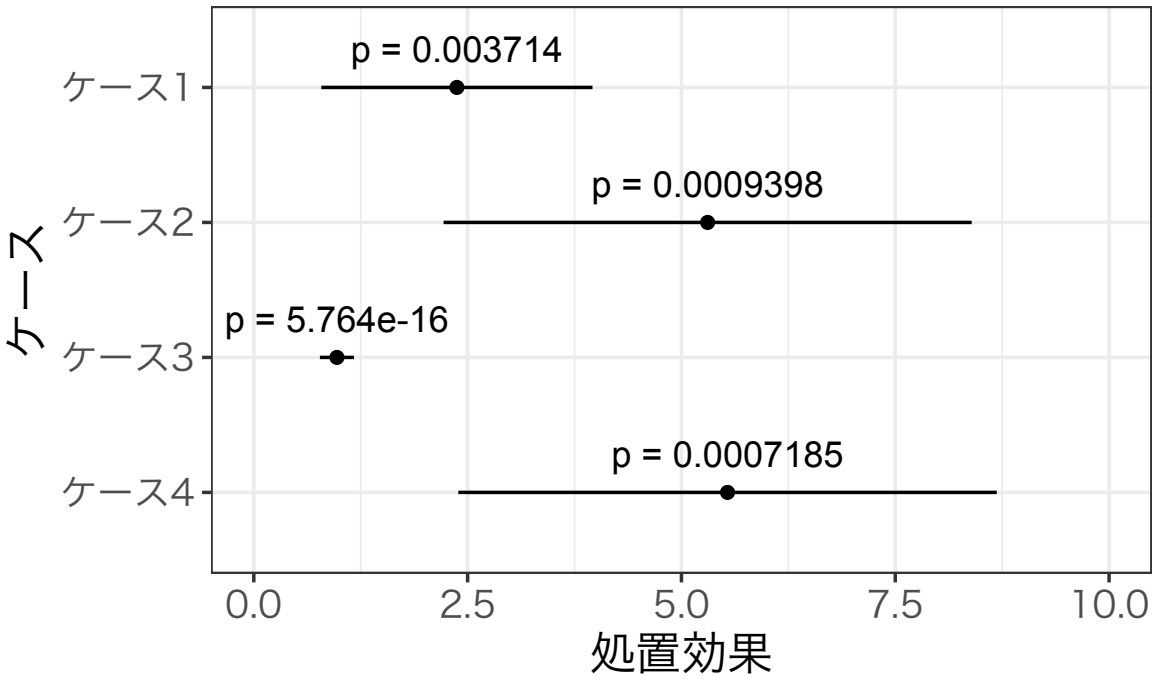

図 1　効果量、不確実性（信頼区間）、p 値の関係

4　問題の解決策

これまで見てきたように、p 値のみに依存する既存の統計的有意性検定は、推定から得られた多くの情報を捨象してしまう。ここでは、このような問題を克服するための 3 つの方法を提案する。その方法とは　(1) 効果量の議論、(2) 信頼区間の利用、(3) 可視化である。

4.1　効果量の議論

効果量を議論することは推定結果を分かりやすい単位で議論することである。説明変数がダミー変数なら、変数が 0 から 1 へ増加した際の応答変数の増加量を議論する。連続変数なら最小値から最大値への変化や標準偏差 1 つ分の変化に対する応答変数の増加量を示す。また、これらを議論する際は他の説明変数を適切な値に固定する必要がある。他にも、説明変数、応答変数を適切に単位変換、標準化することも有効であろう。以下ではいくつかの例を紹介する。

4.1.1　線形回帰 (OLS) の場合

交互作用や自乗項を含まない線形回帰分析の場合、効果量の算出は比較的簡単である。推定された説明変数の係数が、他の変数の値を一定に保ち、説明変数を 1 単位増加させた場合の応答変数の増加分だからである。例えば、松林 (2017) は 1 万人当たりの期日前投票所が 1 箇所増えると投票率は 0.518% ポイント増加することを示した。1 万人当たりの期日前投票所が 1 標準偏差分（= 5.026）増加する場合、投票率は約 2.603% ポイント増加する。これは応答変数の約 0.282 標準偏差である。このように、効果量の単位さえ予め指定しておけば、効果量は係数の掛け算のみで算出可能である。また、横軸に説明変数、縦軸に予測値と信頼区間（または、予測区間）を示せば、読者の理解が促されるだろう。

4.1.2　一般化線形モデル (GLM) の場合

一般化線形モデルの場合、効果量の算出はやや複雑である[25)]。なぜなら、ある説明変数と応答変数間における関係（= 傾き）が説明変数の値によって変化するだけでなく、他の説明変数の値からも影響を受けるからである。

たとえば梶原 (2014) は、選挙公報で「広義の」地方分権改革に言及する候補者の属性を、ロジスティック回帰分析を用いて分析した。推定された民主党ダミー変数の係数は 1.508 である (梶原 2014, p. 100)。しかし、この係数を直接解釈するのは困難なので[26)]、他の方法で効果量を示すべきであろう。この場合、説明変数がダミー変数であるため、民主党ダミーが 1 のときの予測値から民主党ダミーが 0 の時の予測値を引くだけで十分である[27)]。そのとき、その他の説明変数をどの値に固定したかは論文内で示す必要がある。筆者らが計算した民主党ダミーの効果量は第 42 回衆議院議員総選挙の場合、約 0.338 である[28)]。これは民主党所属の候補者は、そうでない候補者に比べ、選挙公報で「広義の」地方分権改革に言及する確率が約 33.8% ポイント高くなることを意味する[29)]。これは非常に大きい値と考えられるため、効果量を言及することによって本論文の重要性がよりア

---

25) ただし、統計ソフトウェアを使う場合、OLS でも GLM でも効果量算出の手間はほぼ変わらないだろう。

26) ロジスティック回帰分析の場合、係数の意味を大雑把に計算させる “divide by 4 rule” (Gelman and Hill 2007) があるものの、この方法では予測確率が 0.5 周辺の効果（つまり、傾きが最も急な位置での効果）しかわからない。また、この手法で効果量が近似できるのは、係数が 0 に十分近い場合のみである。

27) 説明変数が連続変数の場合は、最大値から最小値、平均値を中心に +1 標準偏差から-1 標準偏差などが考えられるが、いずれにしてもどのように効果量が算出されたかを明記すべきである。他にも横軸に説明変数、縦軸に予測値を示すことも有効であろう。

28) 自民党、公明党、新人ダミーは 0 に、他の変数は平均値に固定した。

29) 第 43 回衆議院議員総選挙の場合の効果量は 35.4% ポイント、第 44 回衆議院議員総選挙の場合の効果量は 27.3% ポイントである。いずれもかなり高い効果量であると考えられる。

ピールできると考えられる。

4.1.3　交互作用がある場合

交互作用は条件付き効果や、効果の不均一性を推定する際に用いられる。交互作用項は、主に関心を持つ説明変数とその他の変数とを掛け合わせて作られることが多い。そうすると、主な説明変数の値を変化させると、交互作用項も一緒に変化してしまうため、他の変数を一定にして効果量を考えることができない。よって、異なるアプローチで効果量を計算すべきである。Brambor et al. (2006) の文献サーベイによると、1998 年から 2002 年までのトップ 3 の政治学ジャーナル (*American Political Science Review, American Journal of Political Science, Journal of Politics*) に掲載された交差項を含む論文 156 本の中で、適切に交互作用を解釈した論文[30)]は 16 本のみであった。この論文が発表された後、2006 年から 2015 年まで公刊されたトップ 5 ジャーナルにおいて、ほとんどの論文が適切な解釈を示すようになった (Hainmueller et al. 2019)。その意味で、交互作用に関する Brambor et al. (2006) の論文が政治学に与えた影響は大きい。

Brambor et al. (2006) によると、交互作用モデルを適切に解釈するためには 4 つの条件を満たす必要がある。それらの条件は、(1) 交互作用項 (product term; $X \cdot Z$) だけでなく、交互作用を構成する項 (constitutive terms X, Z 自体) もモデルに投入すること、(2) 交互作用項の係数のみを独立に解釈しないこと、(3) 限界効果 (marginal effect) と (4) その不確実性を調整変数 (modifying variable; Z) の値ごとに示すことである。

たとえば、小川 (2018) は政党イメージ数に対する選挙制度の不均一ダミー変数の効果を推定するモデルを、負の二項回帰分析を用いて推定した。ここで制度不均一を表すダミー変数と政治関心との交互作用が検討されている。単純化すると、このモデルの線形予測子 (linear predictor) は式 (1) のように表現できる。

$$\text{政党イメージ数} = \beta_0 + \beta_1 \cdot \text{不均一} + \beta_2 \cdot \text{政治関心} + \beta_3 \cdot \text{不均一} \cdot \text{政治関心} \quad (1)$$

推定結果から得られた制度不均一変数の限界効果は、-1.258 とされる (小川 2018, p.140)。本論文が示している限界効果は $(\beta_1 + \beta_3 \cdot \text{政治関心})$ の数値ではなく、応答変数の値である。非線形モデルの場合、係数を直接解釈することは困難なため、このように予測値の変化量で示すことは読者の理解の助けになると考えられる[31)]。しかし、この限界効果がどのように計算されたかについては本文中に示されていない。文末注 (9) (p.144)

[30)] Brambor et al. (2006) は 4 つの条件を提示し、それぞれの条件を満たす論文数とすべての条件を満たす論文を掲載した。

[31)] ただし、予測値で限界効果を示す場合、交互作用の統計的有意性検定自体が困難となる。

によれば、共変量が平均値、交互作用項が 0 の場合の限界効果であると述べられている。しかし、政治関心は 1 から 4 までのスケールで、中心化されていないことを考えると、そのような設定は非現実的である。したがって、限界効果を出す際は、調整変数の変化に応じた限界効果を示すべきである。また、本論文には係数間の共分散行列が掲載されていないため、読者から厳密に限界効果とその信頼区間を計算することは不可能である。交互作用を検証する際には横軸に政治関心を、縦軸に $(\beta_1 + \beta_3 \cdot$ 政治関心$)$ の値とその信頼区間を示したプロットを必ず掲載すべきである。これは R の `margins` パッケージ、Stata の `margins` コマンドを使えば簡単に示すことができる。

4.2 信頼区間の利用

p 値のみに依存する統計的有意性検定に対して、信頼 (confidence)・信用 (credible)・予測 (prediction) 区間を示すことが勧められる場合もある (Wasserstein and Lazar 2016; Amrhein et al. 2019)。ここでは伝統的統計学における信頼区間について考える。

信頼区間を利用する際には、信頼区間の正しい理解が必要である。統計学を教えていると、「真の値が ◯◯% の確率でこの区間内に属する」のように誤解される場合が多い。しかし、一つのデータセットに対して特定の統計モデルを用いて計算した特定の信頼区間に真の値が含まれる確率は、0 または 1 である。信頼区間の信頼度は、データ収集（生成）とデータ分析を非常に大きな回数にわたって繰り返し行ったとき、真の値を含む信頼区間の比率が何パーセントになるかを示す。

伝統的な統計学において、母数 (parameter) は一つの数値である。したがって、特定の信頼区間が与えられれば、母数が区間内に含まれる確率は、0% が 100% のいずれかである。同じデータ生成過程を仮定し、同じサンプリング、分析を繰り返し行って 95% 信頼区間を計算するとする。この過程を無限回繰り返した場合、約 5% の分析から得られた 95% 信頼区間内に真の値が含まれず、残りの約 95% の 95% 信頼区間には母数が含まれるはずである。たとえば、同じサンプリングと分析を 1000 回繰り返したとすれば、約 50 回分の分析から得られた 95% 信頼区間には真の値が含まれていないことを意味する。この意味で、95% 信頼区間とは、「真の値が 95% の確率でこの区間内に属する」というのではない。「95% の 95% 信頼区間に真の値が含まれる」と言った表現が適切であろう[32)]。

これは簡単なシミュレーションで確認することができる。図 2 はデータ生成過程を $\hat{Y} = 0 + 1 \cdot X$ と設定し、誤差項が標準正規分布に従うことを仮定したモデルから計算さ

[32)] 言うまでもないが、p 値と同様、統計モデルが不適切な場合には、信頼区間も信頼できない。

れた X の係数の点推定値と 95% 信頼区間をプロットしたものである。サンプリングと推定は 1000 回繰り返した。サンプルサイズは 1000 である。シミュレーションの結果、95% 信頼区間に真の値 $(= 1)$ が含まれているケースは 954 回であり、概ね 95% である。しかし、1000 回分の結果を一つの図にすべて掲載することは難しいため、無作為に 100 回分の結果を掲載した。この図からは 94 回分の 95% 信頼区間内に真の値が含まれていることが分かる。

信頼区間は、図 2 のようなキャタピラプロットによって図示されることが多い。キャタピラプロットは点推定値だけでなく、信頼区間の情報も含まれているため、非常に便利な可視化手法である。Kastellec and Leoni (2007) の論文以来、日本の政治学論文においてもキャタピラプロットが少しずつ普及していると考えられる[33)]。

多くの場合、一人の研究者が同じサンプリング、調査、分析を数十 ~ 百回行うことは難しく、稀である[34)]。それでは、一回限りの分析から得られた 95% 信頼区間から何が分かるだろうか。例えば、大森・平野 (2017, 表 3) によると、「ニュースステーション」の視聴が外的政治的有効性感覚に与える影響の点推定値 -0.15 であり、標準誤差が 0.06 である[35)]。この場合、95% 信頼区間は $[-0.268, -0.032]$ である。もし、真の処置効果が -0.268 だとしても、つまり、今回の点推定値よりも大きい効果量であるとしても、今回得られたデータが異常でないことを意味する[36)]。逆にいうと、真の処置効果が -0.032 のように非常に小さいとしても[37)]、今回得られたデータが得られる可能性は十分にあると解釈できる。

しかし、多くの場合、信頼区間に 0 が含まれているか否かによって「有意である/有意ではない」の判断のみに使われているようである。たとえば、飯田 (2015) は 2 つのキャタピラプロットを示し、「(...) 点の左右に伸びている横線は独立変数の係数の推定値の 95% 信頼区間であり、これが垂直の破線で示される 0 をまたいでいなければ、推定値は 5% 水準で統計的に有意に 0 とは異なる」と説明している（p. 79)。しかし、これだとこれまでの $p < \alpha$ の議論と同じであり、信頼区間を図示する意味が薄れてしまう。なぜなら

33) ただし、既にに述べたとおり、『選挙研究』では信頼区間自体がほとんど扱われていないので、キャタピラプロットもあまり登場しない。
34) その意味で、複数の異なる研究をメタ分析によって統合する努力が必要になるだろう。
35) 効果量を具体的に計算するためには各変数の記述統計量が必要であるものの、情報が不足しているため、共変量が取りうる全ての値に対して効果量を計算した結果、約 7% であった。これは、ニュースステーションを視聴しない回答者が、視聴する回答者よりも外的政治的有効性感覚が約 3.8% 高いことを意味する。
36) 効果量の期待値は約 13% であり、かなり大きいと考えられる。
37) この場合、効果量の期待値は 1% であり、非常に小さい。

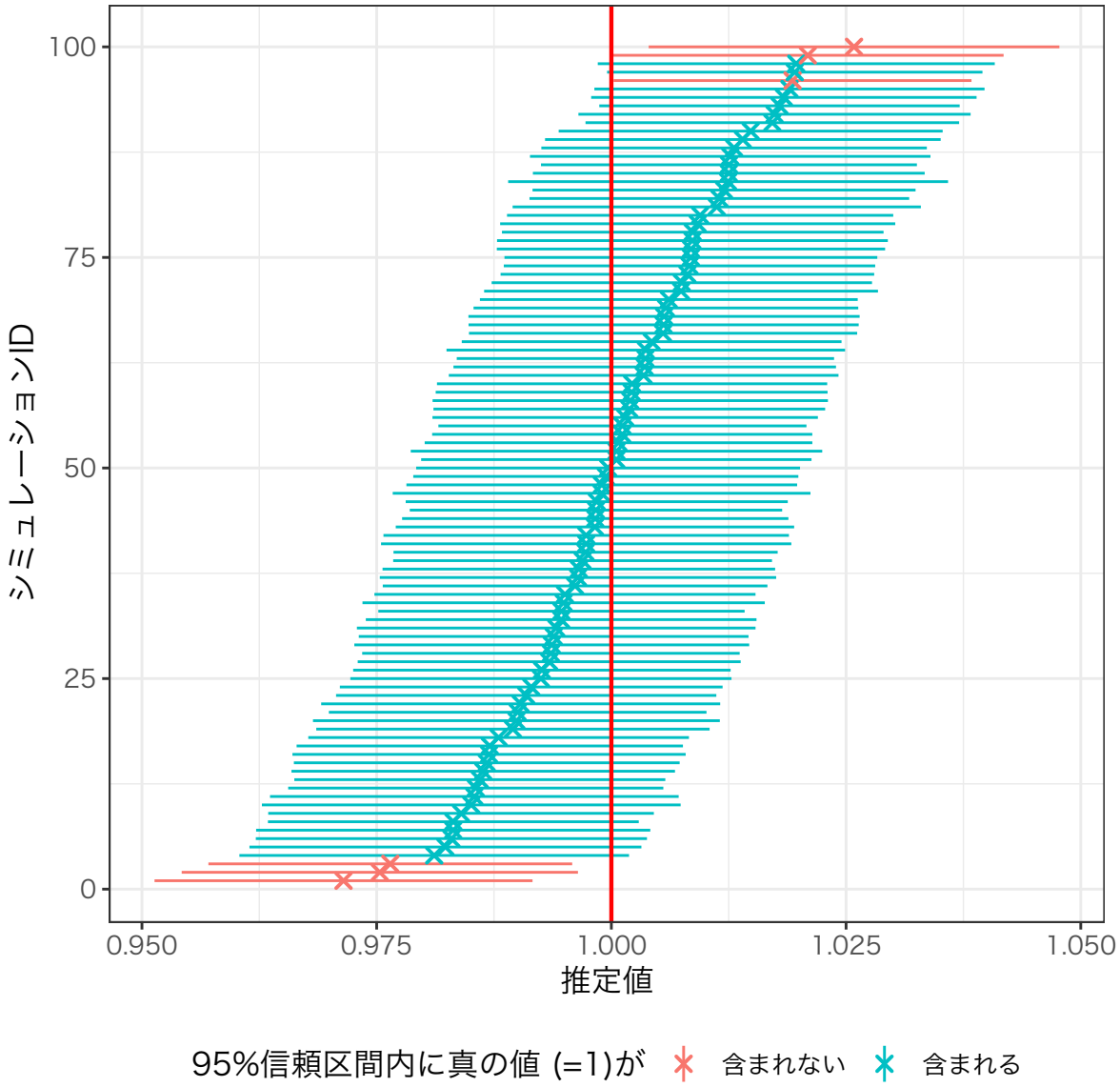

図 2　95% 信頼区間のシミュレーション。データ生成過程は $\hat{Y} = 0 + 1 \cdot X$ と設定し、誤差項は標準正規分布に従うことを仮定した。各施行のサンプルサイズは 1000 であり、1000 回繰り返しを行った。95% 信頼区間内に真の値 $(= 1)$ が含まれたケースは 954 回、含まれていないケースは 46 回である。図は 1000 回の結果から無作為に 100 個を抽出した結果である。

$p < 0.05$ だと 95% 信頼区間に 0 が含まれないのは当然であり、その逆も成立するからである。このような解釈方法は間違いではないが、多くの情報を捨象することになる。

信頼区間を適切に用いるためには、キャタピラプロットによって点推定値と信頼区間を図示した上で、点推定値が示す効果量と、信頼区間の両端における効果量について解釈すべきである。既に述べたように、点推定値だけでは推定の不確実性がわからない。信頼区間の両端についても解釈を与えることで、推定の不確実性が明確になり、実質的重要性についての判断もしやすくなる。推定の精度は信頼区間の長さによって表されるが、信頼区間の両端の意味を考えることによって、区間の長さだけではわからないことまで理解できる。例えば、信頼区間が長い（不確実性が高い）場合でも、両端の効果量がともに実質的に重要な大きさなら、推定結果は大きなインパクトを持つ。反対に、区間が短い（不確実性が低い）場合でも、両端のいずれにおいても効果量が十分大きいとは言えないなら、統計的に有意な結果でも実質的に意味がない結果かもしれない。また、統計的に有意でない結果でも、信頼区間の中に実質的に意味がある値が含まれるなら、その結果を無視すべきではない。今後の研究に生かすべき知見として、詳しく説明すべきである。

4.3 可視化

効果量と不確実性を同時に示す可視化も重要である。定量的研究において、可視化は常に重要なツールである。前節で取り上げたキャタピラプロットや限界効果のプロットも可視化の例である。同じデータを示す場合でも、可視化の方法は様々である。

たとえば、Calin-Jageman and Cumming (2019) は実験データの 2 つの可視化方法を比較した。図 3 の（A）は、実験群ごとの応答変数の平均値とそれぞれの信頼区間を示したものであり、多くの論文で見られる可視化方法である。この図からは両群間において統計的有意な処置効果があることと、効果量の点推定値が約 1.5 であるということが読み取れる。しかし、効果量の不確実性は不明である。一方、図 3 の（B）は、生データ[38]のみならず、それぞれの群の中央値とその信頼区間が示されている（青と赤の点）。他にも右には効果量と信頼区間が示されている（三角の点）。この信頼区間は [0.00001, 2.99] である。もし、真の効果量が 0.00001 という、実質 0 に近いものであっても、今回の結果は驚きに値しない結果であることが分かる。このような情報は図 3 の（A）からでは分からない。

データの可視化方法は無数にある。一般的に図の情報量と、可読性はトレードオフ関係

[38] ケースが多い場合は全ての観測値を示すことが困難であるため、バイオリンプロットで代替できるだろう。

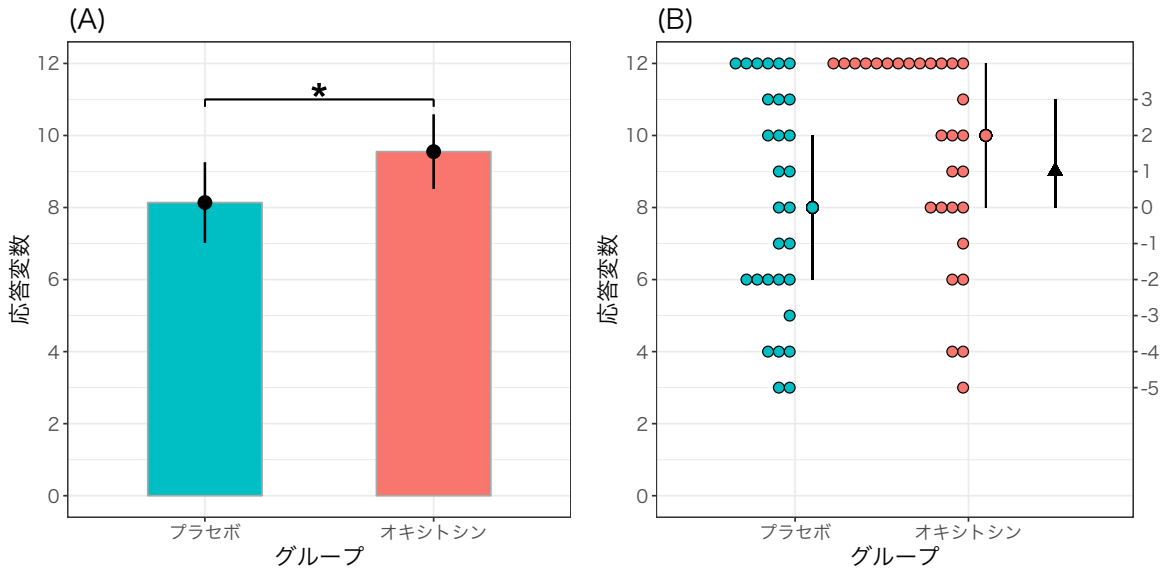

図 3　Calin-Jageman and Cumming (2019) の Figure 1 (p.272) を再現。

にあるが、工夫次第で多くの情報量を分かりやすく伝えることができる。分析者には結果を読者に分かりやすく伝える義務がある。その方法の一つが可視化である。近年、様々な可視化方法が開発され、優れた書籍・マニュアルが入手できる状況にある。これは、結果を分かりやすく、適切に伝えることへの関心の高さの証であろう。そして、可視化の方法を学ぶ機会が十分提供されている以上、可視化の方法がわからないので表だけを掲載するという作法は、今後は通用しなくなっていくだろう。

5　結論：再現性と再生性へ向けて

ここまで述べてきたように、p 値が内包する情報は私たちが求める内容の一部でしかない。また、この問題は p 値のみに依存する定量的分析がもたらす問題の一部である。もっとも深刻な問題は、研究の再現を困難にし、結果として学術的知識の蓄積を阻害することである。

研究の再現とは同じ手順で研究を行い、同じ知見が得られるかどうか確認することである。ここまで述べてきた不確実性を考えると、今回得られた $p < 0.05$ という結果はたまたま得られた可能性もある。これを確認するためには同じ手順でサンプリングから分析まで行う必要がある。しかし一人の研究者には困難な話である。学会・学界の存在意義は、知識を蓄積し、集合的に研究の再現を推進することによって、学術的知見を深めることにあるだろう。また、学術的知見を統合するため、メタ分析の重要性が高まっていくだろう。

しかし、同じテーマを扱った研究が複数存在しても、それらの結果に p 値しか掲載されていないと、蓄積された情報の統合が困難になる[39)]。メタ分析は効果量と不確実性を統合するものであるが、その材料となる研究に効果量・不確実性の情報が存在しないと、情報の統合ができない。本稿を通じて述べてきたように、p 値は重要な情報を教えてくれない。また、メタ分析を意味あるものにするためには、同じ手順で行われた分析であれば、統計的に有意でない結果も含まれるべきである。「有意」という言葉に惑わされ、成果の重要性を p 値に依存して決めるようでは、学術的知見を蓄積するという理想の達成は困難である。「有意でない」結果の投稿を躊躇わせる原因ともなっており、いわゆる「出版バイアス (publication bias)」という問題が避けられない。これからの「知の蓄積」のためには、Amrhein et al. (2019) が提案するように、帰無仮説を棄却することだけを目的にした統計的仮説検定はやめるべきときが来たのではないだろうか。

[39)] p 値すら掲載されていない論文の場合、この問題はより深刻である。

参考文献

Amrhein, Valentin, Sander Greenland, and Blake McShane (2019) "Retire Statistical Significance," *Nature*, Vol. 567, pp. 305–307.

浅野正彦・矢内勇生 (2018)『R による計量政治学』，オーム社.

Brambor, Thomas, William Roberts Clark, and Matt Golder (2006) "Understanding Interaction Models: Improving Empirical Analyses," *Political Analysis*, Vol. 14, No. 1, pp. 63–82.

Calin-Jageman, Robert J. and Geoff Cumming (2019) "The New Statistics for Better Science: Ask How Much, How Uncertain, and What Else Is Known," *American Statistician*, Vol. 73, No. sup1, pp. 271–280.

Fisher, Ronald A. (1922) "On the interpretation of χ^2 from contingency tables, and the calculation of P," *Journal of Royal Statistical Society*, Vol. 85, pp. 87–94.

——— (1925) *Statistical Methods for Research Workers*: Oliver and Boyd.

福元健太郎・中川馨 (2013)「得票の継承に対する世裂の効果：政党投票・候補者投票との比較」，『選挙研究』，第 29 巻，第 2 号，118–128 頁.

Gelman, Andrew and Jennifer Hill (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models*: Cambridge University Press.

Greenland, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman (2016) "Statistical Tests, *P* Values, Confidence Intervals, and Power: A Guide to Misinterpretation," *European Journal of Epidemiology*, Vol. 31, pp. 337–350.

Hainmueller, Jens, Jonathan Mummolo, and Yiqing Xu (2019) "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice," *Political Analysis*, Vol. 27, No. 2, pp. 163-192.

Halsey, Lewis G. (2019) "The Reign of the *p*-value Is Over: What Alternative Analyses Could We Employ to Fill the Power Vacuum?" *Biology Letters*, Vol. 15, No. 20190174.

飯田健 (2015)「有権者のリスク態度と政権基盤の強化—2013 年参院選における分割投票—」，『選挙研究』，第 31 巻，第 1 号，71–83 頁.

Imai, Kosuke, Gary King, and Elizabeth Stuart (2008) "Misunderstandings Among Experimentalists and Observationalists about Causal Inference," *Journal of the*

Royal Statistical Society, Series A, Vol. 171, No. 2, pp. 481–502.

梶原晶 (2014) 「国会議員の政策選好としての地方分権改革」,『選挙研究』, 第 30 巻, 第 2 号, 91–104 頁.

上條諒貴 (2017) 「多数状況における内閣総辞職：政策決定の集権性と党内支持」,『選挙研究』, 第 33 巻, 第 1 号, 57–70 頁.

Kastellec, Jonathan P. and Eduardo L. Leoni (2007) "Using Graphs Instead of Tables in Political Science," *Perspectives on Politics*, Vol. 5, No. 4.

松林哲也 (2016) 「投票環境と投票率」,『選挙研究』, 第 32 巻, 第 1 号, 47–60 頁.

——— (2017) 「期日前投票制度と投票率」,『選挙研究』, 第 33 巻, 第 2 号, 58–72 頁.

中井遼 (2014) 「中東欧新興民主主義国の投票規定要因：有権者個票データによる分析」,『選挙研究』, 第 30 巻, 第 1 号, 113–127 頁.

Neyman, Jerzy and Egon S. Pearson (1933) "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society of London*, Vol. 231A, pp. 289–338.

小川寛貴 (2018) 「制度間不均一が有権者に与える影響—政党差別化の分析—」,『選挙研究』, 第 34 巻, 第 1 号, 73–87 頁.

大久保街亜 (2016) 「帰無仮説検定と再現可能性」,『心理学評論』, 第 59 巻, 第 1 号, 57–67 頁.

——— ・岡田謙介 (2012) 『伝えるための心理統計—効果量・信頼区間・検定力』, 勁草書房.

大森翔子・平野浩 (2017) 「娯楽化したニュースと政治的有効性感覚—戦略型フレーム報道への接触に注目して」,『選挙研究』, 第 33 巻, 第 2 号, 73–87 頁.

小野恵子 (2017) 「米社会における格差の変容と 2016 年大統領選挙—白人高卒有権者にみる「バックラッシュ」とトランプ支持—」,『選挙研究』, 第 33 巻, 第 2 号, 41–57 頁.

Trafimow, David and Michael Marks (2015) "Editorial," *Basic and Applied Social Psychology*, Vol. 37, No. 1, pp. 1–2.

Wasserstein, Ronald L. and Nicole A. Lazar (2016) "The ASA's Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, Vol. 70, No. 2, pp. 129-133.