先生、それPハクです: うむ、お金で解決しよう

伊藤成朗

2024年05月16日 10:03

Brodeur et al. ([2023](#ref-BrodeurCarrellFiglioLusher2023))

[Abel Brodeur, Scott Carrell, David Figlio, Lester Lusher](https://www.aeaweb.org/articles?id=10.1257/aer.20210795): “Unpacking p-hacking and publication bias,” American Economic Review, 113, November 2023, pp. 2974-3002. <doi:10.1257/aer.20210795>

# なぜか分からないけど、効果があってほしい

データを使った実証研究は統計学的な仮説検定をします。たとえば、最低賃金引き上げは失業を増やすという仮説の検定は、

という式で係数が正か検定します。具体的には、統計プログラムなどを使って回帰分析をしてを推計し、推計値(と表記)を得ます。推計値が正だったとしても、誤差で正になっている可能性もあるので、誤差を考慮しても正か、つまり、真の値が統計学的にゼロと違うと判断できるか検定します。[[1]](#footnote-22)

この統計学的推論では、が正しいと想定し、得た推計値が(からすると)どれだけ極端かを問います。真の値がの場合に推計値以上の値を観察する確率が分かれば、得た推計値がどれだけ極端かの判断材料になります。この「帰無仮説が正しいときに、得た推計値以上の値を観察する確率」を値といいます。[[2]](#footnote-23)値が小さければ、

値が小さい  
⇒  
が正しいと想定すると極端なことが起こっている  
⇒  
が正しいと想定するのは誤りなのでは  
⇒  
真の値は統計学的にゼロではない

と判断します。

このように、実証研究では、値が小さいと「効果ありという発見」といえます。[[3]](#footnote-24) 逆に、値が大きいと「効果なしという発見」といえます。[[4]](#footnote-25) 読者の皆さんは、「効果ありという発見」と「効果なしという発見」のどちらに興味をそそられるでしょうか。

筆者にとっては「効果ありという発見」の方が目を引きます。え、そうなの、という反応になることが多いのに対し、「効果なしという発見」は、あ、そう、で終わりがちです。予想通りの結果であったとしても、効果ありの場合は、やっぱりそうか、なのに対し、効果なしの場合は、そんなの当たり前でしょ、になりがちです。なぜそうなるのか分かりませんが、多くの人が筆者と同じ反応をすると思っています。

実は研究者の多くも同じです。今日紹介する論文でミクロ経済学研究者に実施した匿名調査では、値が小さくないと雑誌に掲載されないのでは、と思っている人の割合は8割を超えているのです。[[5]](#footnote-26)

# 効果がないので効果を出そう: ハッキング

雑誌などが値が小さい研究を選んで掲載することを出版バイアスといいます。出版バイアスを予期している、効果ありの方が良い、など様々な動機から、研究者が値を小さくする作業をハッキング(または分岐道forking paths)といいます。

ハッキングは意図的に作り出されることが多いです。推計方法をいじくり回して、値が小さくなるように仕向ける。幾つか推計をして、値が小さい結果だけを報告する。効果が出るまで実験を繰り返す。

一方、意図せずハッキングになってしまうこともあります。推計結果が出てから結果に合うように仮説を選ぶのは、値が小さいことが先に決まっているので検定とはいえず、ハッキング(Hypothesize After Results are Known, HARKingといいます)です。ほかにも、推計をして値が大きいと、モティベーションを失って論文を書かないことも意図しないハッキングです。なぜならば、大きな値の研究をお蔵入りさせ、小さな値の研究だけを世に出しているので、結局は値を小さくする作業になるからです。

# ハッキングは誤解を広める

ハッキングが横行すると、効果ありという(値が小さい)研究ばかり世に出て、効果なしという研究は日の目を見ません。すると、効果ありなんだ、という誤解が世に広まります。だから、ハッキングは困った行為なのです。

ハッキングをもたらすこれらの行為は、疑わしい研究行為(questionable research practices, QRPs)の一部です。文科省などは、研究倫理に照らしてやってはいけない、と指導しています。QRPをする研究者は研究者同士の信頼を失い、研究予算を得にくくなります。しかし、作業過程を隠せばバレないですし、不公正な行為という意識が乏しいことも手伝い、経済学論文でハッキングは横行している…か検討したのが今回紹介するブロデューたちの研究です。

# データ

データは、2013-2018年にJHR誌に投稿された全3607論文、各査読段階の判定結果、各論文に割り当てられた編者と査読者の情報です。論文からは、主たる結果の値を抜き出します。さらに、著者たちの見解を調べるために、投稿者全員561名に匿名調査を依頼し、143名(25.49%)から回答を得ています。

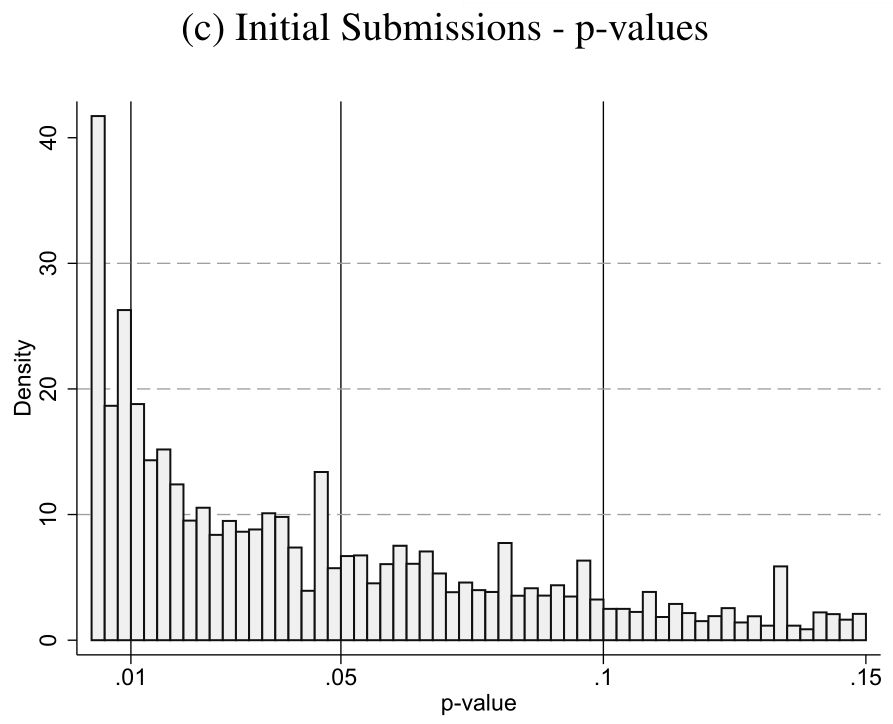
# ハッキングと出版バイアスを検定する方法

多数の論文から値を抜き出し、ヒストグラムを描いたとします。その頂点を結んだ曲線をカーブといいます。[[6]](#footnote-31) [[7]](#footnote-32) このカーブの傾きでハッキング有無が分かります。

真に効果ありの場合、推計値は0よりも大きい場合が多いので、値の小さな研究が多数あります。真の効果が十分に大きい場合、誰もハッキングしなければ、値が有意水準として参照される.05近傍よりも、.01近傍、.02近傍の研究が多くなります。つまり、カーブは右下がりになります。一方、ハッキングがあると、有意水準として参照される.05以下に小さくする作業を研究者がするので、カーブは.05未満部分で増えます。つまり、0から右に進んでいくときに、.05直前で増えるためにその周辺で右上がりになります。.05や.01などの参照される水準で、カーブが右上がりかを検定すれば、ハッキング有無の検定になるのです。

経済学で同様の計測研究は他にもありますが、本論文の強みは、とある学術雑誌(Journal of Human Resources)の初稿から最終稿の詳細情報を入手することで、著者側が仕組んだハッキングと雑誌側が仕向けた出版バイアスの影響を別々に計測した点です。つまり、初稿は著者たちの行動の影響だけが反映されているのに対し、最終稿には著者たちの行動+雑誌側の方針が反映されているはずです。よって、初稿と最終稿を比較すれば、雑誌による出版バイアスの影響をある程度計測できるのです。[[8]](#footnote-33)

# ハッキングは蔓延している



結果は一目瞭然です。カーブで右上がりの部分が.01、.05、.1で見て取れます。[[9]](#footnote-38) 先行研究でも同じ傾向で、経済学全体にハッキングが蔓延していることが分かります。著者たちの匿名調査では、トップ・ジャーナルに論文を掲載した一流の研究者の20%-40%が過去5年に各種QRPに手を染めた、とも回答しており、問題の深刻さが分かります。[[10]](#footnote-39)

一方、初稿と最終稿の比較では、ハッキングの程度は両者で変わらないことが分かりました。つまり、この雑誌で出版バイアスは確認できません。ただし、掲載された論文は、優れているだけでなく、出版バイアスを織り込んで十分にハックしていた、という可能性もあります。よって、最終稿までに変化がないことだけを以て出版バイアスがない、と断定するのは慎重ではありません。この論文で分かったのは、出版バイアスはあったとしても小さい、という示唆でしょう。

出版バイアスは小さそうですが、この雑誌はハッキングのある論文もない論文も同じ程度に掲載しています。よって、この雑誌はハッキングを追認しているともいえそうです。ハッキングを見破るには投稿者以外による推計再現作業が必要なので、雑誌側の体制が整わず追認するしかないのかもしれません。

# なぜハックするのか

JHRのように出版バイアスがない雑誌でも、研究者たちがあると思い込めば、ハックするでしょう。[[11]](#footnote-41)著者たちは、研究者がハックしないように、分析前計画(pre analysis plan)を公開して、データを扱う前にどのような推計をするか決めてしまう制度などを推奨してます。

確かに予防効果はあると思いますが、データを見る前に分析方法を決められない場合もあります。[[12]](#footnote-42) こうした場合には、雑誌側が査読過程を公開するなど、出版バイアスがあるという誤解を解くことも有効だと思います。JHR誌の場合、実在しない心配から研究者がハッキングに手を染めるという、ばかげた状況を回避できます。また、すべての論文で推計再現作業をすると宣言すれば、ハッキングがバレると恐れた研究者がハッキングをしなくなるか、推計再現作業をしない論文に投稿先を変えるかもしれません。

そもそも、最低賃金が失業を増やす「効果あり」でも「効果なし」でも、研究者にとってどちらでも良いはずです。それなのに、とくに利益もないのに「効果あり」にしようと血道を上げるのは愚かです。標本サイズが大きければ、小さな効果も検知できます。ですから、標本サイズを大きくすることに努め、小さな効果でも「効果あり」と判定できる、だけど、小さいから無視可能、のような研究をすべきなのでしょう。しかし、そうすると大規模なデータや実験が必要で、若手研究者は困ってしまいます。予算のない研究者は、斬新なアイディアを見つけるか、ほどほどの標本サイズでも検知できる効果の大きい現象を見つけるしかないのでしょうか。アイディアとお金、どちらかがないと悩ましいです。

# 参考文献

Brodeur, Abel, Scott Carrell, David Figlio, and Lester Lusher. 2023. “Unpacking p-Hacking and Publication Bias.” *American Economic Review* 113 (11): 2974–3002. <https://doi.org/10.1257/aer.20210795>.

Kranz, Sebastian, and Peter Pütz. 2022. “Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics: Comment.” *American Economic Review* 112 (9): 3124–36.

1. 推計値、観察数(標本サイズ)をもとに、という帰無仮説が棄却されるか検定(片側検定)します。 [↑](#footnote-ref-22)
2. 値とは帰無仮説が成立する確率、と説明されることもありますが、正確ではありません。値は帰無仮説が正しいときに手元にあるデータを観察する確率です。よって、手元のデータを得たときに帰無仮説が正しい(成立する)確率ではありません。 [↑](#footnote-ref-23)
3. 正確には、「効果がないこと(=帰無仮説)を強く疑問視する発見」です。疑問視するとは、帰無仮説下で起こりづらいことが起きた、という意味です。 [↑](#footnote-ref-24)
4. 正確には、「効果がないこと(=帰無仮説)を強く疑問視しない発見」です。疑問視しないとは、帰無仮説下で起こりやすいことが起きた、という意味です。 [↑](#footnote-ref-25)
5. 正確には、編者の意思決定に統計的有意が重要な要因だ、と回答しています。 [↑](#footnote-ref-26)
6. 専門用語を使うと、カーブとは各論文から集めた値の確率密度関数で、程度の範囲のものを指します。 [↑](#footnote-ref-31)
7. 帰無仮説が正しい場合、カーブは水平になります。真の値が推計値以上になる確率が値ですが、ここで値が.05だとしましょう。仮に、推計値が少し小さくなったとき、真の値が推計値以上になる確率は5%から6%になるとします。つまり、値が1%ポイント増えるということは、真の値が(より小さい)推計値以上になる確率は1%増えます。値が.1であれ.2であれ(=カーブのどの点であれ)、値が1%増えるときには、真の値が推計値以上になる確率は常に1%増えています。つまり、ヒストグラムを描くと、値がどの水準でも、同じだけの頻度(高さ)を伴います。言い換えると、カーブはどの点でも常に同じ高さなので、水平です。 [↑](#footnote-ref-32)
8. 最終稿が必要なので、最終的に掲載された論文に限って出版バイアスの影響を計測できます。 [↑](#footnote-ref-33)
9. 粗い四捨五入で.05が増える影響も考えねばなりません([Kranz and Pütz 2022](#ref-KranzPutz2022))。たとえば、推計値が0.015、標準誤差が0.014の場合、そのまま値を計算すると0.015/0.014=1.07ですが、四捨五入して計算すると0.02/0.01=2.00になります。こうした研究を取り除いてハッキング検定をする必要があるといわれていますが、このような粗い四捨五入をすること自体にハッキングの意図があるとも思えます。よって、粗い四捨五入の研究も含めた図を見るべきだと思います。 [↑](#footnote-ref-38)
10. カーブの傾きが右上がりということだけでは、ハッキングの浸透度は分かりません。このため、著者たちは研究者たちに匿名調査を実施して、どのくらいの割合でQRPがあるのか計測しているのです。 [↑](#footnote-ref-39)
11. 出版バイアス以外のハックする動機としては、経済理論の裏付けに乏しい仮説を扱っていることも考えられます。風が吹けば桶屋が儲かるのように、理論的裏付けの乏しい(疑わしい)仮説は、データで支持されると意外だからこそ注目されます。儲かる「効果なし」だと誰からも注目されず、雑誌で出版することは難しいでしょう。今は存在が疑われている心理学のプライミング仮説も、「効果あり」だったからこそ注目を集めました。検討に値しない仮説を扱うと、無理して「効果あり」を演出する誘因が出てきます。 [↑](#footnote-ref-41)
12. 2次データを使った観察研究では、情報がどの程度豊富か分からないので、事前に決められる内容に限りがあります。 [↑](#footnote-ref-42)