p 値の是非を考える

李　　為

目　　次

1. 問題提起
2. p 値とは何か
3. p 値と仮説検証の起源
4. p 値の呪縛と危機
5. p 値から自由になる

要　旨

p 値は学界で論争されてきた話題で，ロナルド・フィッシャーは p 値を形成することを提案し，0.05 をこの p 値の閾値に設定した後，この話題の論争は中断されたことがない．p 値の廃止を提案した方々は，p 値を過度に重視しすぎたため，研究者は様々な方法で $p<0.05$ を求めるだけに腐心し，実際の効果の大きさを無視していると指摘している．近年，多くの学者が連名で p 値の廃止を呼びかけているが，廃止を支持しない研究者は，p 値が帰無仮説として成立する確率は客観的な評価基準とみなし，p 値を廃止すると論文で結論を判定することが困難であり，様々な無意味な結論に満ちてくるだろうと反論している．現在，様々なフォーラムで，この問題を議論する議題が次々と出ており，議論の内容は素晴らしいが，参加者の見解には大きな隔たりがあり，p 値が廃止されるべきか否かの結論は出ていない．筆者は社会調査データ分析を扱う立場から，本稿で現状に基づいて p 値の是非を考える．

1. 問題提起

現在，世界中の何百万人もの学生が統計学の授業を履修していると思われる．ビッグデータの時代においてデータの量が増えるにつれて，統計学はますます人気のある話題になっている．ほとんどの受講生が統計学の授業から少し記憶に残っているならば，それは「p 値」と「統計的に有意」（statistically significant）の話だろう．筆者自身も現在，担当している「社会調査法」という講義で p 値が 0.05 以下でなければ受け入れられないと教科書的に解説しているが，社会科学における社会調査データから新しい事実を発見するという真意に悖ることも否めない．

「統計的に有意」と「p 値」は，通常，研究結果が偶然に発生したか否かを定量化するために使用される．たとえば，ある会社は二つの異なる広告がホームページに掲載される影響を測定したいと考えている．広報の担当者は，ある広告が 39％のユーザーのクリックを惹きつけ，別の広告が 30％を惹きつけたことを発見した場合，この違いは確かに差の意味を有しているのか，それとも偶然に起こったのかを明らかにするために，統計学的な検証を行い，結果が「有意」であるか否かを調べることができる．p 値が 0.05 より大きければ偶然的だと判断し，そうでなければこの違いは確かに意味があると考えられる．通常，多くのビジネスや医学分野の意思決定は 5％を偶然性の有無を判断

する基準と考えることが多い．つまり，5％を有意水準（significance level）と呼び，偶然に起こる確率が 0.05 以下（p<0.05）であれば，調査の方法などが誤っている可能性も含め，何らかの必然性をもつ結果として捉え，これを有意（significant）と呼ぶ．逆に，偶然に起こる確率が 0.05 以上の場合は偶然であると考え，有意でない（not significant）と呼ぶ．

すなわち，現在 p 値は帰無仮説（null hypothesis）[1] 検証の基本として広く使われ，統計的妥当性を検討するための一つの基準として利用されている．しかし，p 値の利用と，p<0.05 という棄却域，つまり p 値が 0.05 未満の場合，帰無仮説を棄却するという証拠として扱う．ところが，これについて批判的な研究者も多く存在している．p 値の操作が容易であることと，脆弱なデータを支持するために悪用されるというのがその理由である．p 値に対して否定的な研究者は，p 値は信用できない，すなわち再現性が高くないという事実も指摘している．とりわけ，検定結果が「統計的に有意」だったとしても，それは因果関係を意味しないことが自明する必要がある．

2. p 値とは何か

p 値とは，文字通り p は probability，つまり「確率」の略であり，効果が存在しない（null effect）と仮定して，収集したデータおよびより極端なデータを収集した確率を指す．したがって，p 値は理論が存在するか否かの確率を教えることはできない．しかし，この理論が間違っていることを前提に，収集したデータまたは収集できるが収集できていないより極端なデータの確率を教えることができる．

ところが，p 値がいったいどんな事件の確率なのか．p 値を考える前に，学校で学んだ確率の事例を思い出してみよう．たとえば，コインを投げて，表面の確率も 1/2，裏面の確率も 1/2 である．サイコロ 1 個が投げると，1～6 ポイントの確率はそれぞれ 1/6 である．私たちが確率を言うときは，「サイコロが 6 を投げる」，「コインが反対側を投げる」などの「ランダムな事件」の確率を指している．これに対して，医学の分野では患者と正常に対照した CT 値を比較した場合，p 値は確かに確率であるが，どのような事件の確率だろうか．それが患者と対照グループに差のある確率である．では，なぜ p<0.05 であれば差があると言えるのか，それは患者と対照グループに差がない確率を指す．この答えは合理的だったと聞こえるかもしれないが，p 値が小さければ小さいほど差のない確率が小さくなり，差もあるようになる．

しかし，実際にはそんなに簡単ではない．私たちがよく理解している確率の事例について考えてみよう，ランダムな事件の条件（たとえば，サイコロを投げること）をすでに知っていることに気が付くだろう．ランダムな事件の結果（サイコロの点数）を予想し，それぞれの結果が発生する確率を求めていることがわかる．しかし，学問研究の実践ではこれと逆である．つまり，通常，私た

1） 帰無仮説は通常，記号「H_0」として表記する．対立仮説を「H_1」と表記する．

ちはすでに結果を得ており，逆に原因を求めている．たとえば，6 が既知の結果であるが，6 面のサイコロから来たのか，20 面のサイコロから来たのか，スロットや他の何も知らないものから来たのか，6 を実験の結果とすると，その原因をどのように説明するかは，科学研究が直面する問題である．

医学の分野の話に戻れば，研究者が実際に考察しているのは患者の CT 値が健康な人たちとの差があるか否かについて，これは客観的な事実であるが，ランダムな事件ではない．しかし，各個体の CT 値は他の多くの要因からの影響も受けている．研究者がランダムに患者と対照グループを選んだとき，これらの CT 値の違いはランダムな事件の結果になる．したがって，今，直面している問題は，研究者はこのランダムな結果だけで，この数値の変化はサンプルが代表している二つのグループ間の差なのか，それとも他のランダムな要素によるものかが分からない．これらのランダムな要素は完全に集計されるわけではなく，真に患者と対照グループに差がない確率を求めることは不可能なことになってしまう．私たちはサイコロの種類がどれだけあるか分からなければ，既知の結果が 6 であるから，それが 6 面のサイコロから来る確率はどれだけあるかを問うことは不可能で，この確率の計算はできない．したがって，この場合の p 値は仮定条件に基づく確率であることを理解する必要がある．

いまのべてきたように，簡単なランダム実験で客観的な法則を説明することは，確かに想像以上に簡単ではないことがわかった．とりわけ，患者と対照グループに差がない確率を教えてくれないが，実験で少しでも結論を得ることはできないだろうかと統計学者は考える．もちろん，サイコロのシーンに戻ると，ランダムな結果 6 が知られている．まず考えてみよう．たとえば，スロットマシンは，数字 7（他は果物の模様）しかランダムに出ないようであれば，明らかにスロットマシンからの確率は 0 である．20 面のサイコロの場合，すべての数字を投げる確率は 1/20 で，6 がここから来る可能性も高くないようである．6 面サイコロの確率は 1/6 で，ここから来る可能性が少し高くなったような気がする．したがって，確率を正確に計算するのは難しいが，これらの要素の中から不可能なものを排除するのはそれほど難しくないようである．

このような考え方が，実は複雑な結果を原因の問題に押し返し，再び普通の確率問題に戻ることを発見したかもしれない．その論理は，ある仮定条件でランダムに既知の結果を得る確率が低い場合，その結果が他の要因から来る傾向が強いだろう．仮定条件である結果が得られる確率は明らかに計算できれば，結果の原因を知らない問題よりずっと簡単になる．これが仮説検証の原理である．仮説検証を行う前に，もう一度思考を逆転させる必要がある．この方法はある要素を特定するのではなく，ある要素を排除する傾向があるからである．まず患者と対照グループに差がないと仮定しておけば，この可能性を排除すべきかどうかを見て，「差があるかどうか」の検討に戻ることができないだろうか，仮説を立てた後，差のない二つのグループの中でランダムに標本を抽出し，実験結果を得る確率を計算することは，確かに定量化ができる．

p 値とはいったい何かというに問いに戻ると，それが仮定条件に基づいて得られた実験サンプルの確率であると定義できるだろう．これは統計学者が人間にランダムなサンプル実験から客観的な

法則をまとめて探索するための重要なツールであり，これは統計学の基礎であるにもかかわらず，その背後にある意義，推論，創造という方法の考え方を実現させるために，多くの人が体験できる価値があるに違いない．しかし，「統計的に有意」な p 値について理解できたとしても，p>0.05 の場合，「二つのグループに差はない」についてどのように考えればよいだろうか．

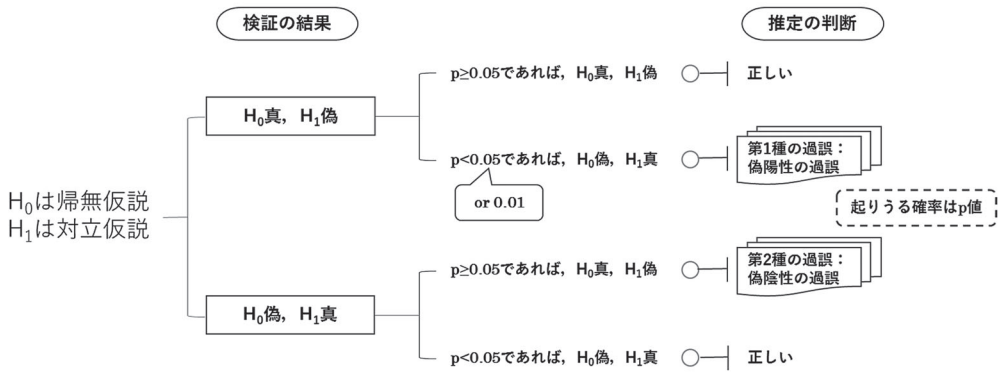
「統計的に有意」は仮説検証，正規分布，p 値の 3 つの概念に基づいている．仮説検証は，サンプルデータを用いて得られた帰無仮説が全体的な特徴に合っているか否かをテストするために用いられる．対立仮説（alternative hypothesis）は，本来の仮定が間違っていると思われるときに受け入れるべき仮説である．言い換えれば，まず帰無仮説を作成し，その後サンプルデータを用いて帰無仮説が成立するか否かを検証する．もし成立しなければ，対立仮説を受け入れる．そこで帰無仮説が成立するか否かを判断するためには，p 値を用いてその統計的に有意を測定する必要がある．とりわけ，データが対立仮説をサポートする傾向がある場合は，帰無仮説を拒否し，対立仮説を受け入れる．

仮説検証の狙いは，小さな確率の反証法であり，仮説のもとで，ある事件が起こる可能性（確率）を推定し，その事件が小さな確率の事件であれば，本来起こり得ないはずだが，今それが起こると，帰無仮説（H_0：両者に差がない）を棄却する理由が成立し，対立仮説（H_1：両者に差がある）を受け入れる．小さな確率の事件でなければ，帰無仮説を棄却することができない．

しかし，p 値の欠点は，p 値と統計分析の前提とする良好な実験（調査）設計とデータ収集のプロセスおよびサンプルデータの代表性を担保しなければならないことが忘れがちである．それによって引き起こした異議が確かに存在している．ここでいう統計解析とは，記述統計と統計推定の両方を含む．前者は統計量，統計表，統計図などのデータ特徴またはデータ分布の特徴であり，後者はパラメータ推定（サンプルによるパラメータ全体を推定する）と仮説検証（サンプルデータがパラメータ全体または分布の仮説を検証し，仮説が成立するか否かを推定する）に分けることができる．

図 1

統計推定の二つの過誤



（筆者作成）

仮説検証と統計推定は異なっているが，図 1 で示しているように推定であれば誤る可能性を否めない以上，小さな確率の事件であってもその可能性がある．仮説検証で起りうる二つの過誤を知る必要がある．

第一種の過誤は，帰無仮説が真（すなわち，二つのサンプルの平均に差がない）であっても，サンプルの抽出，統計量の算出，検証により $p<\alpha$ が得られたら，帰無仮説を棄却し，対立仮説の結論を受け入れる「二つのサンプルの平均に差がある」という推定の結論となる．この時，犯した誤りを「偽陽性の過誤」と呼び，真を棄却する過誤とも呼ぶ．過誤を犯す確率 α は事前に設定した有意水準であるため，通常は 0.05 または 0.01 であるが，他の数値はだめだろうか．たとえば 0.03，0.01 は如何だろう，実際には可能である（本稿で述べている p 値の基準は必要に応じて緩やかまたは厳しめに設定する）が，現在 $p<0.05$ または 0.01 は通常，約束事としての小さな確率の事件発生率である．

第二種の過誤は，もし帰無仮説が偽（つまり，実際の両サンプルの平均に差がある）であっても，サンプル抽出，統計量の算出，検証により $p \geq \alpha$ が得られたら，帰無仮説を受け入れ，対立仮説を棄却する「二つのサンプルに差があると考えられる理由はない」という推定の結論となる．ただし，通常，二つのサンプルの平均に差がない，あるいは二つのサンプルの平均が同じであるような言い方はしない．検証する前に β 値（第 2 種の過誤を犯す確率や偏回帰係数）が設定されていないため，この時，犯した誤りを偽陰性の過誤と呼ぶ．

3．p 値と仮説検証の起源

「統計的に有意」という言葉は 19 世紀 80 年代に最初に見られ，英国の経済学者や統計学者フランシス・イシドロ・エッジワース（Francis Ysidro Edgeworth）は統計検定で初めて使われた．当時この言葉の使い方は今日とはかなり違っていた．エッジワースは，この言葉がどのくらいの確率でマークされているかという「意味のある差」を議論した．当時エッジワースは発見を「有意的な可能性がある」または「一定の有意性がある」と呼んでいた．しかしその後，p 値は学界での論争はやむことがなかった．

他方，p 値に関する文献ではフィッシャーによって発表されたと誤解している場合も多い．実は，はじめて文献で p 値とその計算を正式に述べたのは統計学者のカール・ピアソン（Karl Pearson）で，彼の名を知らない人もいるかもしれないが，Pearson カイ二乗検定については知っているに違いない．カイ二乗検定に関する論文は 1900 年 *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*「哲学の雑誌」に発表され，カイ二乗検定と合わせて紹介されたのは「p 値」と呼ばれたものである（図 2）．

図2 ピアソンの論文資料

158 Prof. Karl Pearson *on Deviations from the*

that denoted by χ is given by

$$P=\frac{\left[\iint\ldots e^{-\frac{1}{2}x^2}\,dX_1\,dX_2\ldots dX_n\right]_{\chi}^{\infty}}{\left[\iint\ldots e^{-\frac{1}{2}x^2}\,dX_1\,dX_2\ldots dX_n\right]_{0}^{\infty}},$$

the numerator being an n-fold integral from the ellipsoid χ to the ellipsoid ∞, and the denominator an n-fold integral from the ellipsoid 0 to the ellipsoid ∞. A common constant factor divides out. Now suppose a transformation of coordinates to generalized polar coordinates, in which χ may be treated as the ray, then the numerator and denominator will have common integral factors really representing the generalized "solid angles" and having identical limits. Thus we shall reduce our result to

$$P=\frac{\int_{\chi}^{\infty} e^{-\frac{1}{2}\chi^2}\,\chi^{n-1}\,d\chi}{\int_{0}^{\infty} e^{-\frac{1}{2}\chi^2}\,\chi^{n-1}\,d\chi}. \qquad \ldots\ldots \text{(iii.)}$$

出所：The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science Series 5, vol. 50, n° 302, 1900

p 値を学界で風靡させたのは，英国の遺伝学者，統計学者であるサー・ロナルド・エイルマー・フィッシャー（Sir Ronald Aylmer Fisher）である．以来，広く応用され，p 値は研究者たちの未知の探索プロセスを伴ってきた．統計学の仮説検証はすでに臨床命題を検証する重要な手段になっている．フィッシャーに代表されているのは p 値のほか，「フィッシャー学派」（Fisherian）と呼ばれる仮説検証の思想である．彼は 1925 年に *Statistical Methods for Research Workers*（『研究者のための統計的方法』）という本を出版した．この本によって彼を現代の推計統計学の確立者として高く評価されている．彼は本の中で，研究者がどのように統計検定の理論を実際のデータに適用して，データに基づいて発見した結論を得るべきかを重点的に述べた．ある統計的仮説を用いて検定を行う場合，この検定はデータとその仮説モデルとの関連性を概括し，p 値を生成することができると提案した．

フィッシャーは，便宜上，p 値を 0.05 に設定することの検討を推奨した．これについては，あるばらつきが目立つと思われるべきか否かを判断する際に，この閾値を判断基準にするのが便利だという意味である．彼はまた p 値がこの閾値を下回る結論は信頼できるので，この閾値を超える統計的結論に時間をかけるべきではないと提案した．その後，フィッシャーのこの提案はますます多くの人に受け入れられ，$p<0.05$ は次第に「統計的に有意」になり，統計学的定義となった．

20 世紀半ばまでに，研究者はある結果を「統計的に有意」または「統計的に有意ではない」と呼び始めた．「有意」という言葉は，判断ではなく一種の提案として捉えた．その後，統計的に有意と p 値はその基準が認められ，計算が容易になるにつれて，科学研究の信頼性を測定する重要な判断基準となってきた．

仮説検証では，p 値は，帰無仮説が真であると仮定した場合に得られる観測結果が現れたり，既

存の観測結果よりも極端な場合が現れたりする確率を表すため，確率値自体は連続的に変化する数値である．しかし，応用には結論を出すための境界値が必要である．0.05 は長い間で応用の境界値（検証水準）として，また受け入れられる偽陽性水準として，研究者の心の中にある「ものさし」となっている．0.05 という境界値が現れたからこそ，もともと連続していた p 値には全く異なる「運命」がもたされた．さらに，統計学的検証は科学問題の検証にしばしば核心的な地位になっているため，p 値が 0.05 以下であるか否かは，研究の成否を決める重要な物差しとなっている．すると，p 値自体の意味はそれほど重要ではないように思われつつ，$p<0.05$ が研究者の目指す目標となった．目標を達成するために，p 値の誤解や誤用も台頭しはじめた．文献レビューで偽陽性率が設定レベルをはるかに上回る可能性があることがわかった場合，偽陽性率を制御するために検査レベルを下げるという発想が出てくるのもその操作にあたる．

上述のように，p 値の起源とその発想の原点を振り返ってみれば，p 値が検証プロセスのすべてではなく，科学問題を探索して論理的なつながりを検証する一環にすぎないことを思い出させる．p 値の大きさと，それが検定限界値に達したか否かだけで結論を出すのは合理的ではない．一つの科学問題の検証プロセスにおける科学的であるか否かは研究設計の段階から始めたのである．サンプルの代表性，合理的な実験措置，信頼できる観察手段とデータ収集の記録措置を選択すれば，偏りを最大限に防止し，データの信憑性を担保することが期待できる．それによって，妥当な統計モデルを選択して検証へ進めば，最後に p 値が働く瞬間となる．しかし，それまでのすべての努力が，最終的に p 値によって規定されることは否めない．しかも私たちが見た割合の高い偽陽性率は，その原因が決して p 値とその検査限界値によってすべて賄うわけではない．直感的には，閾値を変えることは偽陽性率を下げるに違いないが，研究プロセスにおけるほかの起こりうる問題を変えることと，根本的に研究の質を向上させることも期待できない．さらに限界値を変えることは，直面しなければならない人的，財的，物の投入の大幅な増加をもたらす．標本サイズは検査レベルと密接に関係しており，同様の効果で推定値に対して，検査レベルの向上は標本サイズの大幅な増加につながる．

4. p 値の呪縛と危機

(1) p 値の呪縛

医学研究や社会調査研究において，統計方法に関する研究は，理論的には研究者が恐る恐る仕事をしていても，データ分析における論文の結果が無効になる可能性が高い．これは研究者たちが喜んで偽を作るということではなく，統計的な方法を使うと，p 値の呪縛を避けることができないということである．

たとえば，医学研究ではよく見られる二重盲検法（Double blind test）を例に挙げて，p 値とは何かを説明すれば分かりやすいかもしれない．今，ある医薬会社が新薬を開発したとしたら，どのよ

うに有効性を判断するのかについて，二重盲検法が使われている．病状が似た 100 人の患者をランダムに抽出し，ランダムに 2 つのグループに分け，各グループに 50 人ずつ振り分けておく．一方は開発した新薬を服用し，もう一方はプラセボ（偽薬）を服用する．しかし，各患者は自分がどのグループにいるのか，何を服用しているかは告知されない．それだけでなく，介護している患者が何を服用しているか医療従事者にも知らせず，他のすべての条件も同じに設定する．理想的には，新薬を服用したグループが生きており，プラセボを服用したグループでは死者が生じたら，薬効は良好であると考えられる．しかし，実際はこのように理想的ではなく，たとえば，薬を服用したグループは 22 人が良くなり，死者が 4 人出たことと，プラセボを服用したグループは 15 人だけが発病したが，3 人しか死亡しなかったら，非常に気まずいことになる．新薬は有効なのだろうか，それとも無効なのだろうか，非常に判断しにくい．この場合，研究者は p 値を利用して帰無仮説を立てる．薬の効用がないならこの病気の死亡率が 10％であれば，患者が生きる確率は 90％であるが，50 人のグループが生きる確率は 0.9 の 50 乗で，0.00515 に等しいことを意味する．この値が p 値である，帰無仮説が成立しない場合，薬が有効であり，偶然ではない確率は 1－p=0.99485 である．この時，研究者の論文は新薬が有効な実験結果は偶然ではなく，薬物が有効である可能性は 99.485％に達したと結論付ける．これは p 値が 0.00515 に等しいという条件で，第一グループの中で多く死亡した人は偶然であるかもしれないが，第一グループは第二グループより治癒率が高い確率は本当であるといえるようになる．

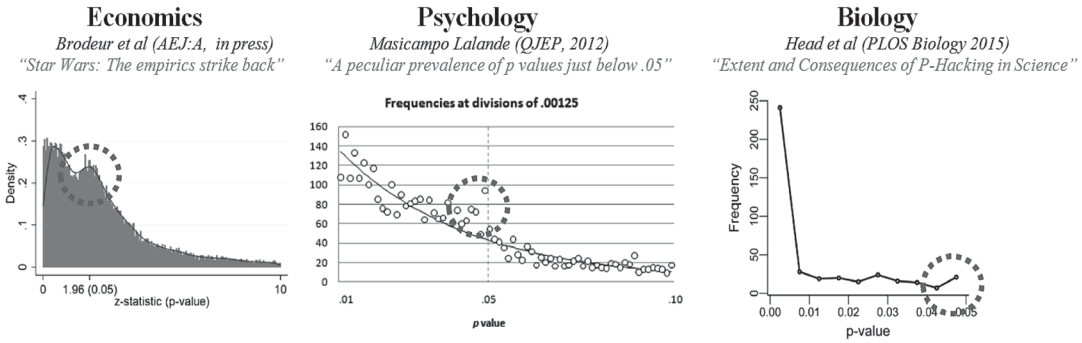
この結論は一見もっともらしい，すきもない結果である．しかし，p 値の取捨基準に注目すると，新薬が有効である可能性の 99.485％は間違っていることがわかる．p 値の本当の意味は，死亡率が 10％という帰無仮説に対して，実験結果が完全に偶然的である確率になる．しかし，なぜ帰無仮説として 10％を選ぶのかについて科学的な論証がなく，研究者の主観的な選択にすぎない．万が一この選択が理にかなっていない場合，さらには間違った判断をすると，それは人の命に係わる問題になる．

現在の学界で統計学的に関与しているすべての研究には，p 値が 0.05 未満であるという p 値が存在しており，科学研究の p 値が 0.05 以上であれば，研究結果は偶然であり，信頼できないと考えているが，0.05 未満であれば，科学研究の成果は「有意性」があると捉えて信頼できる．しかし，このような結論の信憑性が乏しい．なぜ 0.05 を選ばなければならないのかは，科学界の「約束事」である．フィッシャーが考えた意味は，p 値が 0.05 未満の結果こそ「見る価値」があるが，p 値が 0.01 未満の結論を満たすことも受け入れられる．問題は，p 値が 0.01 以下を選ぶと，科学研究コストが高すぎることである．一つの科学実験では多くの被験者を探す必要があり，コストが高くて耐えられない．そこで科学界では普遍的に妥協して次項を求め，0.05 を選んだわけである．実際，この基準のハートルも低くない，多くの若手研究者がこの 0.05 のために時間を浪費しているかもしれない．筆者もかつて 0.05 を求め続けてきた．しかし，仮に p 値が 0.05 未満になっても，研究の結果が妥当だとは言えない．つまり，p 値が 0.05 に等しいということは 20 分の 1 の可能性を意味することで，

帰無仮説の任意性などを考慮すると，p 値が 0.01 に等しい論文であっても，完全な偶然の可能性もあり得る．

これはまだ最悪の状況ではない．さらに深刻なことに，一部の研究者は深刻な「動機的推論」を持っている．簡単に言うと，p 値が 0.05 未満になるために「手段を選ばない」ということである．たとえば「データ収集」の手段として，英語では「cherry-picking」と呼ばれる方法で p 値を下げる．Data Colada に紹介されている *Falsely Reassuring: Analyses of ALL p-values* という内容（図 3）は，ここ数年，経済学，心理学，生物学の論文では，p 値の分布は明らかに p 値が 0.05 に等しいところに，明白な突起があることを示している．唯一のできる解釈は，p の値を故意に 0.05 以下に操作した論文が沢山存在していることである．

図 3　Data Colada で指摘された P-hacking



出所：Data Colada（http://datacolada.org/41）

研究者たちがこのように実験データを操作している以上，人びとは科学を信頼することはできないだろう．しかし，私たちは科学を信じる必要がないだろうか．同時に，科学を信じていなければ，まだ何を信じることができるのか．今日では p 値が科学にもたらす呪縛を積極的に反省しないかぎり自由にはなれないだろう．人間の世界では，美しいものはそれほど真実ではないだろうが，真実もまた不完全である．本当の世界は，おそらく私たちが最初に思ったほど美しくはないが，今や生きていく価値がある．現に認めざる得ないところは私たちが思っているより面白い人間の世界である．

（2）p 値の危機

p 値という科学研究分野における不思議な数値は，多くの人を歓ばせ，また悲しみを与えている．p 値が統計の真実性を測る「基準」になると，それを小さくしようとする沢山の方法が存在している．このような p 値至上のやり方は確かに現在の科学論文が非難される原因の一つであり，今では多くの統計学者に指摘されているように，p 値は想像しているほど信頼できず，明らかに統計学における p 値の地位が危機に瀕している．

90 年前に p 値は誕生してから論争が続けており，嫌われても追いつけない存在だと言う人もいれば，明らかな問題があるが誰もが見てみない皇帝の新しい着物だと揶揄する人もいる．しかし，いまのべてきたように，フィッシャーが初めて p 値を導入した際，得られた実験データが伝統的な意味で有意か否かを判断する非公式の方法として，決定的な検証方法ではなかった．p 値の世界的流行を真に推進していたのはフィッシャーの競合者であるポーランドの数学者イェジ・ネイマン（Jerzy Neyman）とイギリスの統計学者カール・ピアソン（Karl Pearson）で，当時彼らは統計的有効性，偽陽性，偽陰性などの概念を含む代替的なデータ分析方法を提案した．この方法は p 値という指標を直接無視していたことも一つの要因であるかもしれない．

これまで二つの派閥は争ってきたが，他の非統計学者の忍耐力も限界にきたところ，どちらのアルゴリズムにも完全な理解のない研究者は，ネイマンとピアソンが作成した統計システムに p 値を大まかに統合し，p 値 <0.05 の場合，統計結果が有意であるとみなされる新しい混合統計手法を利用した．それゆえ，現在，多くの研究分野における研究結果の意義は p 値によって判断されている．これらは帰無仮説を支持するか，または棄却するかのために利用されている．通常，統計検定された効果は存在しないと仮定する．p 値が小さいほど，結果は純粋な偶然によるものではない．そこで，いわゆる有意な結果を苦心して追求するために，研究者は実験データを収集する際，帰無仮説を仮定せずに意図的に実験結果の中で p 値を発表できる程度にする「p 値操作（P-hacking）」を行うが，これもいくつかの探索的な研究結果をもたらし，実際には重複しにくいように見せる．そのため，多くの研究者は p 値が偽陽性結果を生むことを懸念しているなか，世界で権威のある学会と学会誌では p 値を糾弾している様子を伺うことができる．

① ASA の声明

そこで p 値の乱用については，2016 年 3 月に米国統計学会（ASA）が，p 値に関する声明を発表した．p 値に基づいて科学的な結論や政策的な決定を下すのではなく，統計的に有意な結果をデータ分析して記述し，すべての統計実験や計算における選択について合理的に解釈するよう研究者に呼びかけた（図 4）．

図 4　ASA の六原則

The statement's six principles, many of which address misconceptions and misuse of the *p*-value, are the following:

1. *P-values can indicate how incompatible the data are with a specified statistical model.*
2. *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*
3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*
4. *Proper inference requires full reporting and transparency.*
5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*
6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*

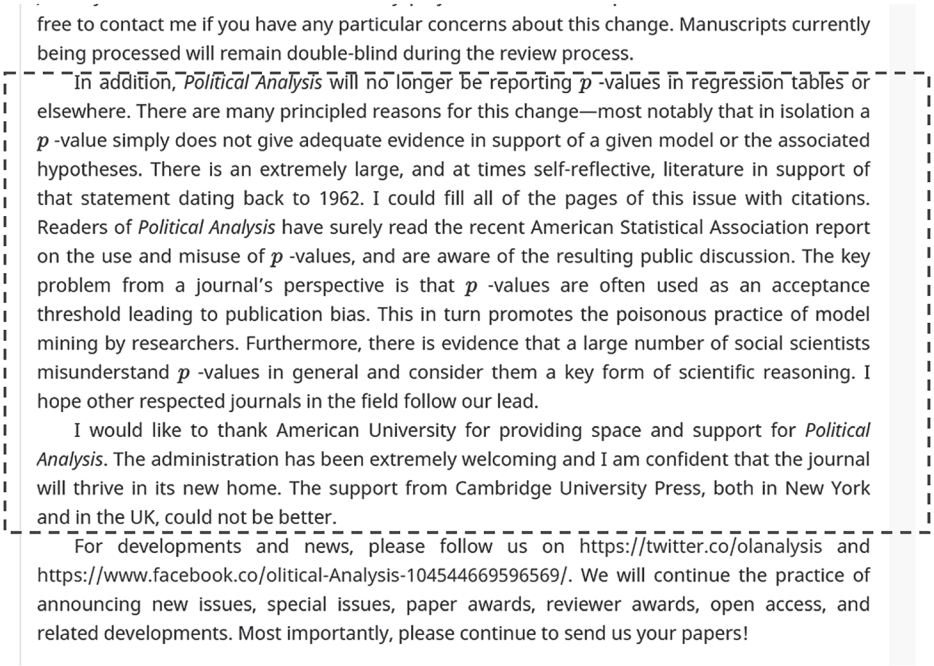
出所：http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108

図 4 に示している英文は，米国統計学会（ASA）は 2016 年に統計的意義と p 値に関して発表した声明であり，p 値の使用と解釈について 6 つの原則を提示している．残念ながら，この声明はより多くの「原則」レベルで解読され，「操作」レベルで実行可能な方法はなかった．

② Political Analysis の新しい編集方針

アメリカ政治学の学術雑誌『政治分析』（Political Analysis）は 2018 年 1 月 29 日に公式 HP で新しい編集方針に関するコメントを発表した．同誌のコメントによると，第 26 号から p 値を無効にする（図 5）．多くの原則的な要因に立脚しながら判断したと説明している．特に p 値は，特定のモデルまたは関連する仮説を支持する十分な証拠を単に与えていないのは問題である．この考えを支持する声は 1962 年に遡ることができるが，大事なのは，p 値がよく閾値として使用されることで出版のバイアスをもたらした．多くの社会科学者が p 値を誤解し，科学的推論の重要な形として捉えていることは原因であると指摘している．

図 5　Political Analysis の編集方針



free to contact me if you have any particular concerns about this change. Manuscripts currently being processed will remain double-blind during the review process.

In addition, *Political Analysis* will no longer be reporting p -values in regression tables or elsewhere. There are many principled reasons for this change—most notably that in isolation a p -value simply does not give adequate evidence in support of a given model or the associated hypotheses. There is an extremely large, and at times self-reflective, literature in support of that statement dating back to 1962. I could fill all of the pages of this issue with citations. Readers of *Political Analysis* have surely read the recent American Statistical Association report on the use and misuse of p -values, and are aware of the resulting public discussion. The key problem from a journal's perspective is that p -values are often used as an acceptance threshold leading to publication bias. This in turn promotes the poisonous practice of model mining by researchers. Furthermore, there is evidence that a large number of social scientists misunderstand p -values in general and consider them a key form of scientific reasoning. I hope other respected journals in the field follow our lead.

I would like to thank American University for providing space and support for *Political Analysis*. The administration has been extremely welcoming and I am confident that the journal will thrive in its new home. The support from Cambridge University Press, both in New York and in the UK, could not be better.

For developments and news, please follow us on https://twitter.co/olanalysis and https://www.facebook.co/olitical-Analysis-104544669596569/. We will continue the practice of announcing new issues, special issues, paper awards, reviewer awards, open access, and related developments. Most importantly, please continue to send us your papers!

出所：https://www.cambridge.org/core/journals/political-analysis/article/comments-from-the-new-editor/3BE4074AAFA13AADEFF3CF7458C9F77E#

③ JAMA 誌の精度アップ

2018 年 4 月 10 日，米国医学会誌（JAMA）は Ioannidis 氏の *The Proposal to Lower P Value Thresholds to .005* が掲載され，検査水準を 0.05 から 0.005 に水準を高めるべきだと提案した．Ioannidis 氏はほとんどの生物医学研究論文が $p<0.05$ の結果を論文で報告していると指摘したが，論

文に強調している論点は間違った可能性がある．多くの論文に p 値が検査水準より低い場合を，誤って「研究結果（あるいは発見）が真実で有効である」と同等に扱った．このような状況は，p 値の誤解，過度の信頼，誤用に由来する．この問題を解決するために，また偽陽性結果の出現を抑制するためにも検査水準を 0.05 から 0.005 に水準を高めるべきであると主張した（図 6）．Ioannidis 氏の見解は決して孤立しているわけではない．Benjemin 氏を含む 72 人の科学者が 2018 年初頭に Nature Human Behaviour 誌で発表した *Redefine statistical significance*（統計的有意性の再定義）においても，検査レベルを 0.05 から 0.005 に変更すべきだと提案した（図 7）．

図 6　JAMA 誌の HP

出所：https://jamanetwork.com/journals/jama/article-abstract/2676503

図 7　Nature Human Behaviour 誌の HP

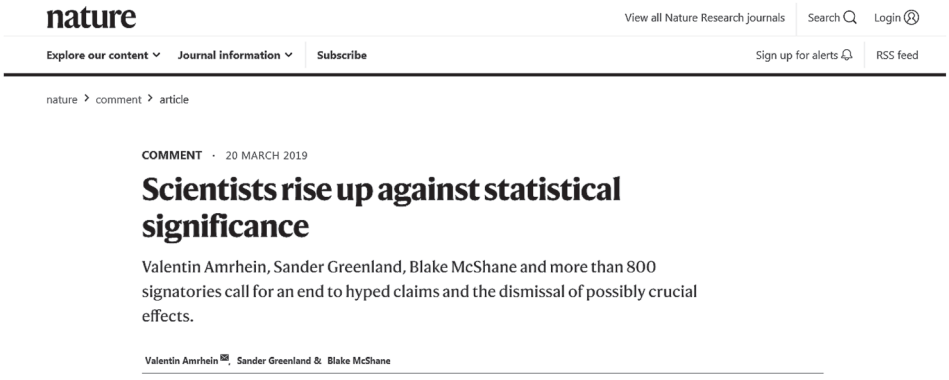
出所：https://www.nature.com/articles/s41562-017-0189-z

もちろん反対意見を持つ研究者も少なくない．たとえば，検査レベルを 0.005 に下げる本来の目的は偽陽性率を下げることに由来しているが，p 値の応用に対する誤解を招き，革新に支障をきたす可能性があると主張している人もいる．多くの要因が偽陽性率の理想と現実の違いをもたらし，単純に閾値を下げてもこれらの問題は解決されず，偽陰性リスクが増加する．この問題を解決するには，研究設計の段階から始め，潜在的な偏りを十分に考慮しなければならない．また，単純に検査水準を高めることよりも，繰り返し試験を行う方が好ましいかもしれない．1 回目の試験の検査レベルを 0.05 のままに据え置き，2 回目の試験の偽陽性リスクは 0.05×0.05=0.0025 を超えないようにすればよい．同時に，研究プロセスの管理と生データの公開を強化し，p 値の定義とその検査限界値を客観的に扱うべきである．しかし検査レベルが 0.005 まで引き上げられたら，研究コストの増加につながると考えられる．これは公共基金の助成金による研究にとって，より研究価値のあるプロジェクトが助成金を受けられない可能性がある．新薬登録を支援する研究にとっても，より大きな財務リスクによって新しい治療法研究が遅延する可能性がある．さらに募集の難しさが高まることで，まれな病気に対する研究設計もさらに困難になると指摘されている．

④ Nature の p 値反対の呼びかけ

事態がさらに進展し，2019 年 3 月に研究者の Valentin Amrhein，Sander Greenland，Blake McShane 三名の研究者は，「統計的に有意」の概念がなければもっと良くなるかもしれないと提案した．彼らは「統計的に有意」という概念が歴史の舞台から去るべきであることを望んでおり，彼らの考えは多くの研究者から支持を得ている．『ネイチャー』誌に「統計的に有意」という言葉を統計学から取り除くことを呼びかけ，800 人以上の学者の署名支持を得ており，その中には定量化と統計学分野の重鎮もいた．彼らの呼びかけは *scientists rise up against statistical significance*（「科学者は統計的有意性を反対する」）というタイトルで掲載された（図 8）．

図 8　Nature の p 値反対の呼びかけ



出所：https://www.nature.com/articles/d41586-019-00857-9

この呼びかけのタイトルは戦闘檄文のように胸が躍る．発表されてから 24 時間もたたないうちに

250 人以上の署名が得られ，1 週間で 800 人以上の研究者も署名に参加した．しかし，大学でやっと分かった統計学は，意味のないことになってしまうだろうか．

なぜ「統計的に有意」という概念を放棄しようとするのか．研究者たちは「統計的に有意」vs「統計的に有意でない」，いわゆる「統計的に有意」に基づいて結論を出す二者択一の思考を捨てるよう呼びかけている．実は，フィッシャーが 1925 年に p 値を触れ，0.05 をこの p 値の閾値に設定することを提案して以来，この話題の論争は中断されておらず，反対する側は p 値を重視しすぎて，様々な方法で $p<0.05$ を求めているだけに腐心し，実際の効果の大きさへの吟味を軽視しており，近年多くの研究者が p 値の廃止を求めているようになった．廃止を反対する側は「統計的に有意」という境界値を設けなかったら，ほとんどの結果が好き放題で発表され，反論もできないナンセンスが支配的になってしまうだろうと反論している．

5. p 値から自由になる

いまのべてきたように，p 値による天下の苦しみはしばらく続けていくだろう．p 値の問題は代替案の進展が見られない今，たとえデータ統計学では徹底的な変革が必要であれば，統計学の教授方法，データ分析の方法，結果の提示と解釈の方法などの新しい統計学的な基準の開発と一連の研究インセンティブをかえなければならないだろうと思う．幸いなこと，現在，日本の社会科学の分野において，上述のような p 値を排除しようという議論が見当たらない．

p 値は私たちに効果や理論が存在するかどうかを教えることができない以上，p 値はいくらか，統計的に有意であるか否かは，私たちに証拠を提供し何も認めたり覆したりすることはできないのである．こうなると，「統計的に有意」という結論は学術的に重要だと考えたりすることが誤ってします．p 値の呪縛から自由であるために，誤る可能性のあるポイントを知っておく必要がある．

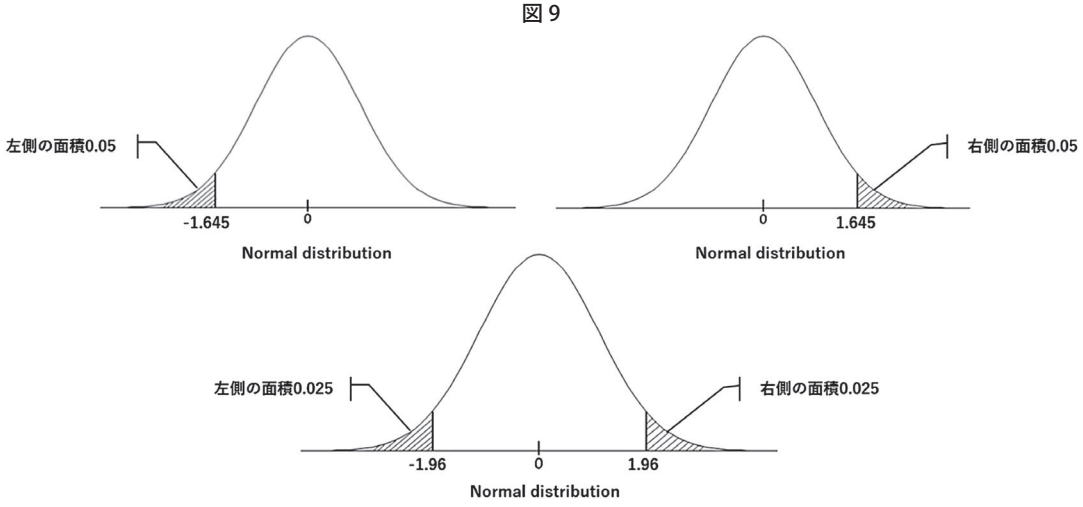
まず 2 つの実験のデータが，1 つは $p<0.05$，もう 1 つは $p>0.05$ であれば，この 2 つの実験の結論は相反していると考えてはならない．さらに 2 つの実験のデータが同じ p 値を得られたら，2 つの実験が同じ有力な証拠を持って帰無仮説を棄却できるとは言えない．つまり，ベイジアン（Bayesian）のデータ統計と比較して，p 値は証拠を定量化する（quantify evidence）ことができない．p 値はいわゆる「統計的に有意」である効果はどれだけ大きいか，どれだけ小さいかについて語ることができない．信頼区間（Confidence interval, CI）であればできるかもしれないが，頻度論者（Frequentist）に基づいた信頼区間も信頼できる保証を提供していない．したがって，有意な差は非常に小さな差であっても，あまり重要な意味はなさそうである．同じ効果であれば相反する p 値を持つが，同じ p 値から異なる効果を得ることもある．

$p<0.05$ は，帰無仮説が成立したことを前提に，データを収集する機会が 5％しかないことを意味すると解釈してしまう場合がある．しかし，これも正しくないと思う．本稿で述べているように，p 値の定義は効果が存在しない（null effect）と仮定して，私たちが収集したデータとより極端なデー

タを収集した確率はどれくらいであるか，実際に極端なデータは存在しなくても，私たちはそれが存在すると仮定しておく．これらの存在しない極端なデータは，私たちが想像しているだけで，やりたがっているがやれない実験である．したがって，原則として，二つの実験のデータは同じデータ，同じ被験者を持つことができる．想像上の実験が異なると，p 値も異なる．これによって，p 値を利用して「カンニング」するという操作性が生まれる．たとえば，文献やデータではサポートできない標本サイズを利用して，p 値が 0.05 に辿りついた時点を見計らってデータ収集を止めてしまう（unjustifiable stopping rule）ことはこれに該当する．

さらに p=0.05 と p≤0.05 はそっくりだと思ってはならない．何故なら p 値と私たちの数学の常識から考えても違うはずである．そのため，そっくりだと考えたり，そっくりのように解釈したりすることは正しくない．他方，p 値を表現する際，たとえば，p=0.049 を p≤0.05 と表現すると，読者を誘導してしまう恐れがある．私たちは通常，p 値と数字を比較しようとする．たとえば，最もよく使われるのは 0.05 である．これは 0.05 を表す一種の誤りであり，私たちが誤って陰性の結果が陽性である確率を受け入れたことを意味する．または，法廷では，すべての人が有罪判決を受ける前に潔白であるということは，有罪判決を受けるには強力な証拠が必要である．p 値は確かに証拠の多寡を定量化することができないが，p=0.049 と p≤0.05 も決して同じことではない．p=0.049 を p≤0.05 と表現するのは人を惑わすだけである．

図 9



（筆者作成）

最後に，片側の結果を気にしなければ，または片側の結論が成立しなければ，反対側の p 値（one-sided p value）を使用してもよいという操作も正しくない．その理由は非常に簡単である．2 つのグループ（a と b）のデータの有意な差があるか否かを比較する場合，二つの仮説を立てることができる．対立仮説：2 つのグループのデータの有意な差がある，帰無仮説：2 つのグループのデータの有意な

差がないという 2 つの仮定から出発する．しかし，対立仮説の場合は a>b と b>a の 2 つの可能性があると考えられる．両側の p 値を以て仮説を検証すると，両方の可能性を同時に考慮していることを意味する．この際，設定した有意水準 0.05 は両側にあり，それぞれ 0.025 である．そのため，帰無仮説を棄却するには 0.025 未満の p 値が必要である．しかし，それも問題である．分析者の主観的推測に基づいており，客観的なデータの支持はなく，統計的に有意な差が検証される可能性を高めるために，a>b, b>a の 2 つの可能性のいずれか一方のみを取り上げるからである．正規分布（図 9）でいえば，左の片側か右の片側か一方のみを取り上げることである．ここまで言うと，読者は p 値を再び疑問視するかもしれない．しかし，筆者が思うのは，p 値の誤りに騙されないように騙す方法を知る必要がある．データ統計はすべてツールであるから，その役割と限界を知っていれば，p 値の呪縛と危機から自由になる．

参考文献

1. Fisher, R.A.（1925）, Statistical Methods for Research Workers, Edinburgh: Oliver and Boyd.
2. 安藤洋美『統計学けんか物語：カール・ピアソン一代記』海鳴社，1989 年.
3. Gigerenzer, G.（2004）. Mindless statistics. The Journal of Socio-Economics, 33, 587-606.
4. Nuzzo, R.（2014）. Statistical errors. Nature, 506（7487）, 150-152.
5. https://www.nature.com/articles/d41586-019-00857-9.（2020 年 12 月 13 日）
6. https://www.nature.com/news/scientific-method-statistical-errors-1.14700.（2020 年 12 月 13 日）
7. https://osf.io/preprints/psyarxiv/mky9j/（2020 年 12 月 20 日）
8. https://qz.com/638059/many-scientific-truths-are-in-fact-false/（2020 年 12 月 20 日）
9. https://statmodeling.stat.columbia.edu/2019/03/20/retire-statistical-significance-the-discussion/（2020 年 12 月 25 日）
10. https://qz.com/1729049/the-origins-of-the-concept-of-statistical-significance/（2020 年 12 月 25 日）

Consider the pros and don’ts of p-value

Wei LEE

ABSTRACT

P-value is a controversial topic in academic circles. After Ronald Fisher proposed to form p-value and set 0.05 as the threshold of this P-value, the debate on this topic has never been interrupted. Scientists who proposed to abolish the p-value pointed out that due to paying too much attention to the p-value, researchers used various methods to find $p<0.05$, while ignoring the actual effect. In recent years, many scholars have jointly called for the abolition of p-value, but those who do not support the abolition of p-value think that the probability that p-value holds as the null hypothesis is an objective evaluation standard. If p-value is abolished, it will be difficult to judge conclusions in the paper, which will be full of meaningless conclusions. Now, in various forums, there are many topics to discuss this issue, and the content of the discussion is excellent, but the participants' opinions are very different, so far, no conclusion has been drawn whether the p-value should be abolished. Based on the current situation, the author discusses the advantages and disadvantages of p-value from the standpoint of processing social survey data analysis.