以下は、Sander Greenland氏による2019年出版の論文「Valid P-Values Behave Exactly as They Should: Some Misleading Criticisms of P-Values and Their Resolution With S-Values」の抜粋の翻訳です。
https://www.tandfonline.com/doi/full/10.1080/00031305.2018.1529625
(https://www.tandfonline.com/doi/full/10.1080/00031305.2018.1529625)

American Statistician ISSN: 0003-1305 (Print) 1537-2731 (Online) ジャーナルホームページ: www.tandfonline.com/journals/utas20 **有効なP値は期待通りに振る舞う：P値に関するいくつかの誤解を招く批判とS値によるその解決** Sander Greenland この記事を引用するには: Sander Greenland (2019) Valid P-Values Behave Exactly as They Should: Some Misleading Criticisms of P-Values and Their Resolution With S-Values, The American Statistician, 73:sup1, 106-114, DOI: 10.1080/00031305.2018.1529625

この記事へのリンク: https://doi.org/10.1080/00031305.2018.1529625
(https://doi.org/10.1080/00031305.2018.1529625) オンライン公開: 2019年3月20日 ジャーナルへの投稿 記事閲覧数: 43415 関連記事を表示 Crossmarkデータを表示 アクセスと利用に関する完全な利用規約は
https://www.tandfonline.com/action/journalInformation?journalCode=utas20
(https://www.tandfonline.com/action/journalInformation?journalCode=utas20) で確認できます THE AMERICAN STATISTICIAN 2019, VOL. 73, NO. S1, 106–114: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1529625 (https://doi.org/10.1080/00031305.2018.1529625)

有効なP値は期待通りに振る舞う：P値に関するいくつかの誤解を招く批判とS値によるその解決 Sander Greenland カリフォルニア大学ロサンゼルス校疫学部門および統計部門 要旨 本稿は、**「有意水準」と「P値」の定義が権威ある情報源間で衝突していること**、およびそれに続くP値の**誤差確率**としての誤解釈など、P値に対する的外れな批判の原因を探るものである。次に、**P値のいくつかの特性、例えば、サンプル間で極端な変動を示す（したがって「信頼できない」）、効果量とサンプルサイズを混同する、サンプルサイズに敏感である、調査員のサンプリング意図に依存する**といった特性が、致命的な欠陥として提示されていることについて論じる。これらの特性はしばしば尤度やベイズの枠組みから批判されるが、**それらは、有効なP値がその本来の枠組みの中で正しく構築され解釈された場合に、まさに示すべき特性なのである**。その他の一般的な批判は、P値がユーザーに無関係な仮説に焦点を当てさせ、それらの仮説に対するエビデンスを過大に表現するということである。しかし、これらの問題はP値の特性ではなく、**帰無仮説に焦点を当てたり、p = 0.05が仮説を棄却するのに十分なエビデンスであるという誤解に基づいてエビデンスを過大に表現したりする研究者の過失である**。これらの問題は、観察されたP値 p をシャノン情報量（S値または驚き） −log2(p) に変換することで、ベイズの概念を用いることなく容易に理解できる。

記事履歴 2018年3月受理 2018年9月改訂 **キーワード コンパチビリティ(相性の良さ)**; 二分強迫; エビデンス; 情報量; ログワース; ヌルイズム; P値; S値; 有意性検定; 驚き **1. はじめに** 伝統的な検定（P値を決定ルールに固定していた）を批判する理由は数多くあるが、それでもなお、**それらはP値が設計通りに振る舞うことや、その誤用・誤解釈をP値のせいにする根拠を提供するものではない**。以前にも論じられているように、問題はむしろ、**教科書やチュートリアルがP値の推論的意味を正しく記述することに失敗していること、および実践的なニーズに適した検定仮説を記述することに失敗していることにある**。また、これらの問題の責任をユーザーに押し付けるという極端な傾向もあるが、最高レベルの統計学文献でさえ、情報源によって一貫性のない、そして日常言語の意味に反する専門用語に基づいた記述や定義を示している。これらの問題を整理するには、一般のユーザーが合理的に期待できる以上の専門知識が必要であり、Gigerenzer (2004)、Hoekstra et al. (2006)、Hurlbert and Lombardi (2009)、Greenland et al. (2016)、Amrhein et al. (2017)、McShane et al. (2017, 2018)、Wasserstein and Lazar (2016)、その他多くの情報源で論じられているような**混乱と誤解の舞台を設定している**。

本稿では、**権威ある定義における「有意水準」と「P値」がどのように異なり、P値の誤差確率としての誤解釈を招くか**について述べる。次に、**P値に対するいくつかの批判が、むしろ不十分な教育や専門用語の問題であること、およびP値のいくつかの特性が、しばしば致命的な欠陥として提示されるが、有効なP値がどのように振る舞うべきかを反映しているに過ぎないこと**について論じる。根底にある見解は、**P値は確かに範囲が限られており、正しく理解するのが難しいものの、多くの手厳しい批判は、これらの統計量を不適切な**

基準で判断するか、または並行して批判されるべき方法と比較して判断しているということである。特に有害なのは、**決定的な推論への要求が、信頼区間やベイズ統計学を含むいかなる方法をも損なう可能性がある**ことを見過ごしている批判である。 概念は、イブプロフェン（Advil™、Motrin™）を単独で、または他の薬と併用して投与された乳児と、アセトアミノフェン（パラセタモール、Tylenol™）のみを投与された乳児の有害事象発生率を比較したコホート研究で説明される。腎臓の有害事象に関する注目すべき結果は、調整推定率比（RR）が1.84で、95%信頼限界が0.66、5.19であった。これは、アセトアミノフェン単独と比較してイブプロフェンで84%の発生率増加、34%の減少から419%の増加までの区間、そして関連性なし（率比1、0%差）の帰無仮説を検定した場合のP値が0.25に対応する。理論的な記述は、回帰モデルA（検定に使用される背景仮定の集合をエンコードする）に埋め込まれた係数βに関する仮説H、例えばH: β = b（bはβの固定された仮説値）の検定に焦点を当てる。乗法率モデルでは β = ln(RR) である。検定される値 b は通常ゼロであるが、そうである必要はない。議論は、埋め込みモデルAの適合性の検定というより一般的な場合も考慮する。脚注1 簡単のため、本稿ではデータセットが十分に大きいと仮定し、通常のすべての連続 大規模サンプル近似が成り立つとする。最も技術的な補足事項は脚注に入れる。後述するように、P値をエビデンス尺度として理解する上でのいくつかの問題は、観察されたP値 p を、例えばS値 s = log2(1/p) = − log2(p) のような情報尺度に変換することで軽減できる。

2. 基本的な定義と記述における混乱と認識されていない差異 いくつかの問題は、権威ある情報源間での定義と専門用語の差異に起因しており、それらの情報源自体はその差異について一切言及していない。**「観察されたP値」の定義は、少なくとも2つ広く使用されている**。通常の推論的（フィッシャー流）定義では、P値は、テスト統計量（例えば絶対Zスコアまたはχ2）が、仮説Hと埋め込みモデルAの下で観測された値以上となる確率 p である。フィッシャー(1925)は p を「有意水準」または「Pの値」と呼び、「有意水準」はその後イギリスの情報源で一般的である。意思決定理論的（ネイマン・ピアソン流）定義では、観測された p は、p ≤ α の場合にHを棄却するα水準の決定ルール（ネイマン・ピアソン仮説検定）で棄却を可能にする最小のα水準（検定のカットオフ）として定義されることが多い。この2つの定義は表面的には異なって見えるが、**数学的には等価であり、したがって論理的には単一の定義を2つの異なる方法で述べたものである**。

残念ながら、一部の著者はαを「有意水準」と呼び、pをそう呼ばないことがある。この「有意水準」の2番目の用法は元の用法と矛盾し、pとαの混乱を招くことがあり、しばしば非常に微妙な方法で生じる。このような権威あるテキスト間での inconsistent な用法に直面して、基本的な教科書や研究者が p と α を混同しても驚くべきではない。観察された p は、観測されたデータを p の計算に使用された仮説HとモデルAに関連付ける**サンプル特性である**。具体的には、100p はHとAから計算された分布における観察された検定統計量のパーセンタイル位置であり、この特別な意味で p はデータに関する何かを記述していると見なすことができる。対照的に、αは既知の固定値（0.05など）であり、**データについて何も教えてくれない**。

「P値」が確率であるという一般的な定義にもかかわらず、多くの意思決定理論的著者は**繰り返しサンプリング特性にのみ焦点を当て**、したがって「P値」を観察された値 p ではなく、**所与のサンプルにおけるその値（実現値）が観察された p であるような確率変数 P として定義する**。したがって、我々には「P値」の論理的に異なる2つの定義が存在する。しかし、矛盾する専門用語と同様に、この定義の衝突はめったに言及されないため、基本的なテキストや研究者が確率変数 P と観察された p を混同しても驚くべきではない。これは、変数名と、それを測定した際に観察されるであろう特定されていない値を混同するようなものである（例えば、「体重(kg)」という変数Xと、誰かの体重を測った際に観察されるかもしれない特定されていない値「x kg」を混同するようなもの）。

この2つの定義の区別は重要であり、特に頻度主義者は以下のP値の**妥当性**を定義しているからである。 **P は、HとAが正しい場合に一様分布する場合、HとAの検定において妥当である（または適切に較正されている、あるいはU値である）**。その場合、すべてのαに対して、ルール「p ≤ α のときにHを棄却する」は、Hを誤って棄却する頻度が α となる。脚注2 もちろん、P値の計算に使用される検定統計量の選択にはパワーの問題が入ってくるが、それらの問題は現在の範囲外である。

観察されたP値 p と確率変数 P の混同は、後述および以前に説明されたいくつかの誤謬の主な原因である可能性がある。さらに悪いことに、**観察された p の尾部確率定義は、**「pは単なる偶然によって関連性が生じる確率である」**のような全く誤った記述と同一視されたり、置き換えられたりすることが多い**。これは単なる専門用語を超えた混乱を反映している（単なる偶然の確率は、HとAが正しい確率に他ならないため）。

このような明白な誤りを別にしても、**観察された p がHに対するエビデンスを測定すると記述される場合に、より微妙な問題が発生する**。なぜなら、pは H に対するエビデンスと逆相関しているからである。**より小さい p の値は、データにおける H に対するより多くのエビデンスに対応する**。

誤解釈を防ぐために、p は、H とデータ（モデル A を所与として）との間の**統計的コンパチビリティ(相性の良さ)**の度合いを測定するものとして記述することができる。これは 0 = 完全な**インコンパチビリティ(相性の悪さ)**（HとAの下ではデータは不可能）から 1 = 検定からは明らかな**インコンパチビリティ(相性の悪さ)**がない（HとAの下ではデータと適合している）までの範囲に限定される。同様に、A の適合性の検定における結果として得られる p は、A とデータとの間の**コンパチビリティ(相性の良さ)**の尺度として解釈することができる。脚注3 しかし、尺度としての p のスケーリングは貧弱である。例えば、0.01 と 0.10 の差は、0.90 と 0.99 の差よりも幾何学的にかなり大きい。例えば、H と A の下で平均ゼロ、標準偏差（SD）1 の正規分布に従う検定統計量を使用した場合、p が 0.01 vs 0.10 は統計量で約 1 SD の差に対応するのに対し、p が 0.90 vs 0.99 は約 0.1 SD の差に対応する。

方向性とスケーリングの両方の問題に対する1つの解決策は、方向を逆にして、P値を負の底2ログに再スケーリングすることである。これにより、S値 s = − log2(p) が得られる。**s の値が大きいほど、H に対するより多くのエビデンスに対応する**。以下で議論するように、これはS値をAを所与としたHに対する（あるいは p がAの適合性検定から得られた場合はAに対する）エビデンスの尺度として使用することにつながる。 **3. P値は誤差確率ではない（無意味な仮想的な意味の場合を除く）** 観察されたP値 p を最小α水準として記述または定義し、pとαの両方を「有意水準」として参照することは、**p が実際の決定問題における誤差確率として誤解釈されることにつながったようである**。後者の解釈は単なる間違いである。α水準仮説検定の理論では、分析者がデータを見る前に、モデルから生成される一連のデータ全体にわたる**棄却ルールのエラー率の望ましい上限**として、定数（0.05など）であるαを指定する必要がある。具体的には、αはデータのコストに基づき、データのコストとは独立して決定される。偽陽性の結果が軽微な結果をもたらす人々はより高いαを使用し、偽陽性の結果が重大な結果をもたらす人々はより低いαを使用するであろう。同じ設定でも利害関係者によって意思決定の結果が異なる可能性があるため、同じ研究報告を読む異なる読者は、偽陽性と偽陰性による損失が異なり、したがって異なるα水準を持つことになり、同じデータを見て同じ観察されたP値 p に基づいても**矛盾する意思決定**につながる可能性がある。

それにもかかわらず、P値の定義や提示にそのような α は必要なく、また p と α を比較することが p をエラー率にするわけでもない。観察された p に等しいエラー率 α を持つ決定ルールを想像することはできるが、α と p は概念的に異なる量であり、p はサンプルによって変動するため事前に指定することはできない。例では、腎臓の有害事象に関する率比1.84は、指数率モデルにおける推定イブプロフェン係数 $\hat{\beta} = \ln(1.84) = 0.610$ に対応し、標準誤差 $\hat{\sigma} = \ln(5.19/0.66)/2(1.96) = 0.526$、H: β = 0 の検定統計量 $\hat{\beta}/\hat{\sigma} = 1.159$、そしてP値 p が0.25となる。我々は「P値が α = 0.25 未満であれば β = 0 を棄却する」というルールを想像することはできるが、データを見る前にそのルールを述べた人はおらず、ましてやエラーコストから導き出した人はいない。したがって、観察された p を25%の第一種の過誤率と記述することは、せいぜい偶然にもP値が0.25に等しくなった事前に指定された α が0.25であった誰かに関する記述である。例にはそのような人は確かに存在しないため、この記述は全く不必要な注意散漫である。**検定結果の正確な記述は、単に β = 0 の P値が0.25であったということである**。そのP値のみに基づいて決定を下したい読者は、0.25が自身のカットオフ α を上回っているか下回っているかをすぐに確認できた。

関連する小さな点は、α水準 α と、Hが真である場合に検定がHを棄却する**実際の確率**（検定の第一種の過誤率、または検定サイズ）の一般的な混同である。仮説Hおよび仮定Aを所与とした検定の実際の第一種の過誤率は、Aの欠陥やデータの離散性によりαから逸脱する可能性があるため、調査者にはしばしば不明である。脚注4 対照的に、αは調査者によって設定される許容される最大第一種の過誤率として定義され、したがって既知である。pは、対応する確率変数 P が妥当である（これによりαが第一種の過誤率と等しくなる）という仮定の下で、意思決定のためにこのαと比較される。

4. P値に関する誤解を招く批判 4.1. P値はユーザーに帰無仮説に焦点を強制しない — しかしヌルイズムの専門用語は強制する 統計的検定に対する一般的な批判は、それがユーザーに無関係な帰無仮説に焦点を強制することである。多くの帰無仮説が確かに科学的に無関係であることは疑いようがない。しかし、この無関係性の問題は、P値の欠陥ではなく、**ユーザーにそのような仮説に焦点を当てさせる伝統的な訓練と学術環境**

の産物である。誤解を招く「有意性」を含む専門用語を廃止する継続的な呼びかけがある一方で、**フィッシャーが「帰無仮説」をあらゆる検定される仮説Hに使用した間違いを訂正する試みはほとんど行われていない**。これは、日常英語で「null」がゼロまたは何も意味しないことを無視している。脚注5 この伝統により、ユーザーは統計科学が「帰無仮説」の検定のみに関するものだと考えるようになり、特定の専門家はそう主張するに至った。「null」が「関連性なし」または「効果なし」を意味し、重要または懸念されるあらゆる仮説すべてを意味しないという誤解が生まれた。

フィッシャーの誤解を招く専門用語から脱却し、ネイマン(1977)は H を標的または検定される仮説と呼んだ。しかし、フィッシャーの専門用語が、「null」は検定によって「無効にされる」べき仮説であるという後付けの合理化によって広まり、β = 0 のような仮説を「nil hypotheses」として区別する試みによって維持されてきた。したがって、フィッシャーの用語の使用の悪影響が、**正当に批判されている帰無仮説有意性検定（NHST）の崇拝において見られ続けている**ことは驚くべきことではない。NHSTでは、P値は効果なしの仮説についてのみ計算されるが、関連する代替仮説（例えば、資金申請の際にパワーを計算するために使用される仮説）についても提供されるべきである。

より技術的に複雑な問題は、**点仮説への過度な焦点**であり、これについてはサンプルサイズへの感度に関する箇所で後述する。 **4.2. P値は「信頼できない」 — まさにそうあるべきである** P値は、理想的な実験の繰り返しであっても、サンプルごとに劇的に変動することがしばしば指摘されている。その結果、一部の研究者はP値を「再現できない」「信頼できない」「ノイズが多い」と批判したり、P値がパラメータを推定しているか、「統計的有意性」が自然の状態であるかのように、繰り返しやサンプルサイズの増加に伴って何らかの定数に収束しないことを批判したりする。

しかし、**パラメータの測定は、P値ではなく、推定量 β̂ によって行われる**。この推定量とその標準誤差は、モデル A における β によって表現されるデータの変動の系統的な成分に関する情報を抽出する。観察された p は、H が正しい場合の A のサブモデル（β = b）を、β に制約のない埋め込みモデル A と比較するための標準化された残差の要約と見なすことができる、H: β = b の検定のための絶対 Z スコア |β̂−b|/σ̂ によって捉えられる残りの変動から計算される。**P値が、H と A の下で純粋なランダムエラーであるはずの関数の関数であることを認識すれば、それは定義上、パラメータや集団量などを測定するものではない — まったくその逆である。P値は、H と A を所与とした、推定値におけるランダムエラーの尺度であり、一様分布に再スケーリング（標準化）されている**。同様に、埋め込みモデル A の適合性の検定の P値は、モデル A が信号または系統的な変動であると示しているもの（ノイズが除去された後に残る信号である、フィットされた回帰方程式またはフィットされた値の表）を抽出した後に残るノイズまたはランダムな エラーの再スケーリングされた尺度である残差平方和に基づいている場合がある。

したがって、H と A が正しい場合、確率変数 P は（有効であれば）単位区間全体で一様分布に従って変動し、時間の約 100α% で p ≤ α を生成する。**H と A が正しい場合、P は純粋な一様ノイズ（ランダムエラー）であるべきであり、それは研究ごとに予測不能に変動することになっているため、これはまさに P が行うべきことである**。データ収集と分析の繰り返しにおいて P が一様ノイズのように見えない場合、少なくとも H または A の一方が正しくないこと、すなわち、推論（信号抽出）または決定に間違った仮説または間違ったモデルを使用していることを警告される。それが統計的検定が目的とするところである。代わりに P が一様である場合、この特定の検定が H または A のどのような違反が存在してもそれに insensitive であるとしか言えない（意思決定理論の観点からは、これらの違反に対して、P に基づく検定のパワーはその第一種の過誤率と等しく、それらの違反に関する情報は何もないことを意味する）。したがって、**P値は仮説またはモデルの問題に対する診断または警告メカニズムを提供し、そのようなすべてのメカニズムと同様に誤りやすい**。

4.3. P値は効果量とサンプルサイズを混同する — まさにそうあるべきである 「統計的コンパチビリティ(相性の良さ)」の概念は、P値の適切な理解と解釈の主要な障害となっている。この概念は、観測値と期待値の間の乖離の大きさを、それらの乖離における推定されるランダム変動と比較することを含む。実際、**P値が効果量とサンプルサイズを混同することはしばしば嘆かれており、この理由でP値が禁止されることさえある**。これは皮肉なことである。なぜなら、**この「混同」はP値がその仕事を正しく行っていることを反映しているからである**。推定値とモデル予測からの距離（例えば）は、モデルに対するエビデンスを構成するべきであるが、その距離がどれだけのエビデンスに対応するかは、推定値の精度に依存するべきである。

特に、モデルAを所与としたH: β = bに対する情報は、**絶対的な偏差（推定値と仮説値からの距離）|β̂ − b| と標準誤差 σ̂ の両方の関数である**。β̂ と σ̂ はAにおける補助的な（デコードする）仮定に基づいた信号とノイズレベルに対応する。これらの量は、Aを所与としたHの検定統計量 |β̂ − b|/σ̂ に結合される。β̂ の精度 σ̂−2 はサンプルサイズ n に比例するため、検定統計量は n と推定値 β̂ から仮説値 b までの距離 |β̂ − b| に直接依存する。したがって、p は n と |β̂ − b| の両方に逆相関しており、**β = b に対する一貫性のあるエビデンス尺度としてまさにそうあるべきである**。例では、H: β = 0 に対する距離は |β̂−0| = 0.610 であり、P値 p は再び0.25である。しかし、もし研究の精度が4倍になった場合（例えば、コホートのサイズ n を4倍にした場合）、σ̂ は半分になり、同じ距離0.610でもP値は0.02となる。これは、**与えられた偏差によって表されるHに対するエビデンスが n とともに増加するべきである**という事実を反映している。

4.4. P値は仮説に対するエビデンスを過大に表現しない — 人々がそうする P値に対する公正な批判の中には、P値が**歪曲されたエビデンス尺度であり、適切に評価するために対数変換が必要である**というものがある。したがって、観察されたP値のシャノン情報量変換またはS値 s = − log2(p) を再び考える。これは、**検定統計量にエンコードされたHに対する情報量（モデルAを所与とした検定によって提供される反証情報）の尺度である**。脚注6 確率の負の対数変換は、その確率 p を持つイベントを観測することに対する**自己情報量**または**驚き (surprisal)** としても知られている。底2の対数では、この情報を測定する単位はビット（バイナリ数字）である。s より大きい最初の整数は、p をエンコードするために必要なバイナリ数字（指示変数）の数である。脚注7

統計的情報の他の正式な定義も存在するが、**S値 s は、H と A を指定するために使用されたもの以外の文脈の詳細を参照することなく、p が伝える情報を理解するための単純な認知デバイスである**。特に、情報量尺度 s は観察された尾部確率 p を参照し（確率変数 P ではなく）、したがって観察されたデータと H を A と組み合わせることで形成されたモデルとの関係を表す。対数スケーリングの有用な結果は、独立した検定間で情報量を加算できることであり、これはメタアナリティックP値を作成するために使用される事実である。s が伝える情報の直感的な解釈を提供するために、k を s に最も近い整数とする。我々は、p は、k 回の独立なコイン投げですべて表が出ることと比較して、「公平である」（各投げが独立で表の確率は1/2）という仮説に対する、または表に偏っているという仮説に対する、同じくらいの情報またはエビデンスを、検定される仮説 H に対して A を所与として伝えると言うことができる。このイベントを表すには、すべて1に等しい k 個の指示変数が必要となる。

例として、4回の公平なコイン投げですべて表が出る確率は 1/24 = 0.0625 である。したがって、モデルAの下で、0.05のP値を観測することは、H: β = b に対して s = − log2(0.05)= 4.3 ビットの情報しか伝えず、これは4回の公平なコイン投げですべて表が出るよりも**ほとんど驚きではない**。イブプロフェンの例では、0.25のP値は、β = 0 の場合よりも2回の公平なコイン投げで2回とも表が出る場合の方が驚きではない。なぜなら s = − log2(0.25)= 2 だからである。対照的に、β = ln(2) は、イブプロフェンによる有害事象発生率の倍増に対応する。検定統計量 |β̂ − ln(2)|/σ̂ = 0.158 からのP値は0.87であり、s = − log2(0.87)= 0.19 である。これは、イブプロフェンによる有害事象発生率の倍増に対する情報量が、差がない場合（β = 0）に対する情報量よりもさらに少ないことを示している。

この尺度によれば、特定の解説（例: Goodman 1999; Sellke et al. 2001; Hubbard and Lindsay 2008）とは対照的に、**P値は仮説やモデルに対するエビデンスを過大に表現しない**。観察された p は、HとAから計算された**単なる仮説上の確率またはパーセンタイル**である。p が伝えるエビデンスの過大評価は、（自動化された決定ルールにおけるα = 0.05の定着に基づき）**p = 0.05が検定された仮説またはモデルを棄却するのに十分なエビデンスであると誤って考える人々によるものである**。さらに、この解釈の下で、埋め込みモデルまたは検定された仮説を変更した結果としてP値 p およびしたがってS値 s の変化は（例えば、単一比較から多重比較への変更から）、異なるモデルまたは仮説に対してデータが提供する情報量の違いを正しく反映している。それでもなお、S値は p = 0.05 がわずか4.3ビットの反証情報にマッピングされることを明らかにすることにより、解説を裏付けている。これは、**p = 0.05を「境界線上のエビデンス」と解釈することが悪い文化的な習慣に過ぎない**ことを示している。これは、事前確率にスパイクを置くことやベイズ因子のようなメカニズムを使用することなく理解できる根底にある認知の問題である。そのようなメカニズムは、ベイズ派の擁護者でさえ異議を唱えるものがある。

4.5. P値はサンプルサイズに敏感である — まさにそうあるべきである 検定統計量における情報量を測定するために S値 s = − log2(p) を使用すると、P値はP仮説（より不正確には「α仮説」と呼ばれる）の反証バージョンに従う。**等しいP値は、検定された仮説 H または検定されたモデル A に対する等しい反証情報量に対応する**。この特性はしばしば批判される。例えば、同じ検定仮説に対して2つの研究が同じP値を提供するかもしれないが、非常に異なる観察された関連性を示す可能性があるからである。この批判は、**H のどのような違反が実践的に重要であるかについての追加の背景（文脈的）情報を考慮する必要性**を指摘している。関連する批判は、埋め込みモデル A が、効果の均質性やランダムサンプリングまたは治療のランダム化などの明示的な仮定、およびデータベースエラーや選択的分析報告などのあまり述べられない仮定を含む、推論に使用されるすべての補助的な仮定で構成されているということである。そのため、A は決して完全に正しくない。したがって、十分な大きさのサンプルでは、p は非常に小さくなり、したがって H が正しい場合でも s は非常に大きくなる。

したがって、**検定は、H と A の両方が実用的な目的には十分であったとしても、少なくとも H または A の一方または両方が厳密には偽であると示す可能性がある**。したがって、批判は「大きなサンプルでは、P値は H または A からの小さな逸脱に対して敏感になりすぎる」と再構成することができる。ベイズ派は、H に事前確率の点質量を置き、残りを代替仮説の制限されたファミリー（通常は同じ埋め込みモデル A において β が既知の定数であるすべてのモデルのファミリー）に広げるという人為的な手段によってこの批判を拡大する。この人為的な手段の結果は、Pまたは-log(α)の基準に従ってHに対してかなりの情報を提供するデータが、依然としてAを所与としたHの事後確率を増加させる可能性があるということである（ジェフリー・リンドレーのパラドックス）。

この大規模サンプルのP値が重要でない不一致に対して敏感であることの粗雑な修正策は、サンプルサイズが増加するにつれてHを棄却するためのα水準を下げることである。しかし、この修正策は、サンプルサイズが増加するにつれてHまたはAを棄却するためにより多くの情報を要求するだけであり、**αが基づくべき実際の誤差コストを無視している**。さらに、これらのコストと要求は、検定によって与えられる統計的情報の測定とは無関係である。実際、**どのような有効で効率的な検定も、十分なデータ情報が与えられれば、たとえそれが実用的な結果をもたらさないものであっても、モデルの不完全性を検出するべきである**。むしろ欠陥は にあり、観測された不一致 |β̂ − b| の大きさに実用的な観点から十分な注意を払わないことにある。

実践的な重要性を考慮するには、**不完全性に対する許容範囲を指定する必要がある**。例えば、実際の不一致 |β − b| に対する最大許容範囲などである。これを行う1つの方法は、H: β = b のような点目標を、H: |β − b| ≤ c のような区間目標に置き換えることである。ここで、[-c, c] は実際には不一致なし（β = b）と同等と見なされる区間を表す。コンテキストによっては、検定と代替仮説の役割を逆転させる方がより合理的である場合もある。その場合、検定される H は |β − b| ≥ c となり、**同等性検定**のようになる。脚注8 役割逆転の他の例としては、リスク制限監査がある。これらのすべての場合において、観察されたP値は、Aを所与とした複合仮説Hとの**データコンパチビリティ(相性の良さ)**を測定し、S値はAを所与としたそのHに対する情報量を測定する。これらの仮説が β = b よりも文脈的により関連性があることを認識すると、サンプルサイズへの感度はP値に対する有効な反論ではなくなる。それはむしろ**点検定仮説に対する反論であり、我々のすべてのモデルには十分なデータがあれば気づかれる不完全性がある**ことを思い出させるものである。P値とS値がこの健全な直感に適合することは心強いはずである。

反対の異議は、P値やS値が、モデルに対するエビデンスと見なされうるデータの特性（例えば、ありそうもない推定値など）を無視する限り、**無感度**であるということである。これは真実であるが、これは**点推定値や信頼限界のような単一の数字の限られた情報容量**を反映しているに過ぎない。多次元情報の一次元的な要約は、推論のための**十分統計量**ではない。したがって、 **sensible な推論を行うには、さらなる情報（例えば、残差プロットや代替仮説またはモデルのP値など）も必要になる**。小さな p は、仮説 H または埋め込みモデル A の何かが間違っている可能性があることを警告するだけで、何が間違っているのか、またはそれらが安全に使用できないということまでは警告しない。初期チェックに続くモデル更新の種類は、集中的な事前情報、診断、およびモデル違反に関する疑念に依存する。例えば、埋め込みモデル A の不十分な適合は、検出された不完全性に対処するためにそのモデルを緩和（拡張）すること（例えば、高次項を追加すること）によってしばしば対処される。

逆に、**大きな p は、検定が検定された仮説またはモデルに対して多くの情報を提供しなかったことを意味するだけであり、「安全信号」ではない。**S値は、95%信頼区間内のいかなるパラメータ値も、それに対して4.3ビット以上の情報を持っていないことを明らかにしている。これは、**区間内部をデータと非常にコンパチブル(相性が良い)な仮説の領域と見なす推奨**を裏付けている。区間外部をデータによって除外された領域と過度に自信を持って見なすのではなく、**区間内部をデータと非常にコンパチブル(相性が良い)な仮説の領域と見なす**推奨を裏付けている。イブプロフェンの例では、率比exp(β)のこの**高いコンパチビリティ(相性の良さ)**の領域は0.66から5.19に及び、この研究が（たとえそれ以外は完璧であったとしても）、腎臓の結果に対するイブプロフェンの考えられる有害効果について実際には**実用的に情報量が少ない**ことを明らかにしている。なぜなら、1/3の発生率減少（RR = 1/3）から5倍の発生率増加（RR = 5）までの anything が、それに対して4.3ビット未満の情報しか持っていないからである。しかし、研究の要約は、乳児における腎臓の有害事象発生率は「イブプロフェン（+/-アセトアミノフェン）群とアセトアミノフェン単独群の間で異ならなかった」と結論付けており、これは引用された結果の実用的な情報量の少なさに対する**ヌルイズム的な認知盲目**の一般的なタイプを示している。対照的に、正確な報告は「我々の研究は、イブプロフェンとアセトアミノフェン単独を比較した腎臓の有害事象に関する有用な推論に到達するための十分な情報が不足していた。イブプロフェンの安全性の懸念に対処するには、はるかに多くのデータが必要だろう」と述べるべきであった — もっとも、現在のジャーナル出版基準の下では、そのような正直な結論は出版を困難にするであろうが。さらに、2/3の発生率減少（RR = 1/3）と10倍の発生率増加（RR = 10）に対して、それぞれp ≈ 2−10 のほぼ10ビットの情報があると言うこともできるが、それは安全性の保証に関する研究の不十分さから注意をそらすだけである。

4.6. P値は調査員のサンプリング意図に依存する — まさにそうあるべきである 統計的原則は、意図または分析計画（プロトコル）の推論における役割において、これほどまでに **persistently に衝突するようには見えない。「証拠を測定する際に短期的な視点に固執するならば、同一のデータは実験者の意図に関わらず同一の証拠を生み出す」**という主張を考えてみよう。この主張は、 **possibly questionabl な限定を伴わなければ、偽である。**一つの実験（非常に短い実行である）が、H: RR = 1 に対して p = 0.004、RR の点推定値 R̂R = exp(β̂)= 1.8、そして RR = 2 と RR = 1 を比較する尤度比が50を超えるという結果を報告したとしよう。もし実験者が最善のプラクティスに従うことを意図し、そうしたと報告していれば、我々はその結果を H に対するいくつかの信頼できるエビデンス（s = − log2(0.004)= 8ビット）を提供するものと見なすかもしれない。しかし、実験者の意図が、必要であれば事後ランダム化の再配分と除外によって p < 0.005 となるデータを作成することであったと判明した場合、我々は得られたデータを H または一般的な RR に関するエビデンスをほとんどまたは全く提供しないものと見なすべきである。

実験者の意図がこの例で重要であると言うことができる。なぜなら、最初のケースでは、意図は初期の有効な研究デザインによってのみ制限された標本空間を残すのに対し、2番目のケースでは、それらは標本空間を p < 0.005 のサンプルにさらに制限するからである。データのエビデンスを測定する際にこれらの意図を合理的に無視できる唯一の方法は、意図がサンプルを変更する行為を引き起こさなかったことを確実に知っている場合である。例えば、初期の治療割り当てが偶然 p = 0.004 を与え、 ill-motivated な実験者による操作にはつながらなかったことを知っている場合である。しかし、そのような情報がなかった場合、操作が不要になる確率に依存する必要がある。これは H が正しい場合0.005となる。その場合、p = 0.004（または H に対する対応する小さなベイズ因子）を見ることは、「H が正しく、かつ 実験者の行動と報告を介してその影響が mediated されるべきであるにもかかわらず、操作は行われなかった」という複合仮説に対するエビデンスを提供するだけである。

先ほど説明した種類の欺瞞的な意図を発見した場合、正直な意図の下での形から RR の尤度関数を変更する方法で埋め込みモデルを変更するべきである（例えば、RR=1のときのR̂R=1.8 の確率は、正直な意図よりも欺瞞的な意図の方がはるかに高くなるであろう）。したがって、そのような意図を考慮に入れることは論争を招く必要はない。**現在の論争は、オプショナル・ストッピング問題のように、 p を変更するが尤度関数は変更しない意図に関するものである。**これは広範なトピックであるが、簡単に言えば、**P値が実際には観測されていないサンプル（反事実データ）に依存することから、一部の人々は純粋な尤度またはベイズ統計量を支持してそれをエビデンス尺度として拒否する。**しかし、多くの統計学者にとって、**完全なサンプリング分布（反事実データセットを含む）は、データ生成メカニズムに関する重要な情報をエンコードする。**脚注9 最終結果が事後確率として報告される可能性がある場合でも、**P値は仮定をチェックするための基本的なツールとして残る。**

P値が反事実データに依存することは、したがって、すべての代替案を指定することなく（予見されない種類のモデル違反を考慮に入れる適合性検定のように）仮定の不確実性を考慮に入れる手段と見なされている。この観点から、**尤度比の反事実データに対する無感度は情報損失を反映しており、高次元問題ではそれが極端になる可能性がある。**高次元問題では、尤度ベースの手続きは一貫した推論を提供しない可能性があるにもかかわらず、完全なサンプリング分布から有効なP値を構築することができる。

5. 結論 P値は、正しい解釈の下で持つべき特性を持っていること、およびデータの誤用と誤解釈を助長することについて、しばしば批判される。P値が、正しい教育、記述、および使用にとって問題があることは間違いなく、多くの一般的な誤解がその証拠である。これらの問題は、**P値の歪んだスケーリングと不当な二分化、その定義と記述を取り巻く誤解を招く矛盾した専門用語、およびその使用と解釈を取り巻く不当な伝統をP値のせいにする的外れな批判**によって悪化する。これらの問題は、以下の方法で軽減できる。

1. α水準や決定を参照しないP値の定義に焦点を当てること。その定義では、**観察された p は尾部確率であり、その対応する確率変数 P は、その構築に使用された仮説とモデルの下で一様変数である。**
2. 零（無効または「効果なし」）仮説に対する、**文脈的に重要な代替仮説**（例えば、最小重要差）についてP値を計算すること。
3. **p をシャノン情報量（S値） s = − log2(p) に再スケーリングして、検定が仮説に対して提供する情報量のより良い尺度を提供すること。** そして最も重要なのは
4. **p の二分化や p と α水準の比較を避けること。**

例外として、例えば分析の中心目的が、関連する代替仮説の下での偽受容（第二種）エラーコストと偽棄却（第一種）エラーコストの両方を含む明示的な正当性が与えられている決定ルール（例えば、品質管理アプリケーションのような場合）にのみ基づいて、 p を α と比較して決定を下すことである場合、(4)の例外が生じると主張されるかもしれない。しかし、**そのような徹底的に正当化されたアプリケーションは、観察研究の設定では一般的ではない。**

そうすると、**厳格なα水準仮説検定に対して熱心に運動してきた多くの著者が、その代わりに95%信頼区間の形でその一種を推進している**ことは皮肉に思えるかもしれない。5% α水準から派生しているように、慣習的な95%信頼水準はエラーコストの考慮から切り離されており、得られる区間は典型的には、β = b の検定が p > 0.05 を生み出すすべての β に対する b の集合に過ぎない。したがって、**信頼区間が期待された検定乱用の治療法を提供しておらず、代わりに研究報告を悩ませる二分強迫と過剰な確実性を永続させている**ことは驚くべきではない。そして、ベイズ検定や事後区間がより賢明に扱われると期待する根拠もない。

せいぜい、95%信頼区間は、例えば区間内の値に対して4.3ビット未満の情報しかないと判断される、**与えられたモデル内のデータと可能なパラメータ値との間の高いコンパチビリティ(相性の良さ)の領域全体を大まかに示している。**したがって、これらの区間を信頼区間ではなく**コンパチビリティ(相性の良さ)区間**と呼ぶ方が誤解を招かないであろう。ここで「**コンパチビリティ(相性の良さ)**」とは、データが区間内のパラメータ値に対して（区間を構築するために使用されたモデルの下で）限定的な情報しか提供しないこと（5回のコイントス分にも満たない）を意味するだけである。事後確率区間にも同様の注意が適用され、「**コンパチビリティ(相性の良さ)**」とは、データが区間内のパラメータ値に対して（区間を構築するために使用されたモデルと事前分布の下で）限定的な情報しか提供しないことを意味するという言い換えがなされる。

この制限は、**区間内の値を支持する決定的なエビデンスや、区間外の値を反証するエビデンスではなく、データとモデルにおけるパラメータに関する情報の不足を表している**ことを強調する必要がある。特に、95%区間から遠くない値も限定的な 情報しか持たず、別の、同様に妥当なモデルから生成される可能性がある。このような注意は、区間推定法の根底にあるモデル（いわゆるロバストな方法を含む）が、モデルで説明されていない方法（例えば、モデル化されていないまたは誤ってモデル化された測定誤差など）で間違っている可能性がある場合に特に重要である。そのような場合（健康科学および社会科学ではこれが標準である）、95%のカバレッジまたは事後確率記述は、モデルの不確実性を適切に組み込んでおらず、したがってこのように記述された区間は**過信区間**となる。これらの問題に関するより完全な議論は Greenland (2018) に記載されている。

謝辞 この論文の初期の草稿に対して有益なコメントを提供してくれた Valentin Amrhein, Sameera Daniels, Michael Fay, Lawrence McCandless, Mohammad Mansournia, Keith O'Rourke, Philip Stark, David Trafimow, および匿名の査読者に感謝する。また、元のバージョンおよび改訂バージョンに関する非常に詳細なコメント、修正、提案をしてくれた Allen Schirm に特別な感謝を捧げる。

脚注:

1. β = b が A を所与として成り立つ場合の検定は、β が制約されていないが埋め込みモデル A は成り立つモデルと比較して、β = b が成り立つモデルの適合性の検定として扱うことができるため、この場合はより一般的である。
2. 同様に、HとAが正しい場合に「p ≤ α のときにHを棄却する」というルールがHを棄却する頻度がα以下である場合、Pは保守的に妥当である。保守的妥当性は、離散データで通常できる最善のことである。
3. 主語が A の場合は「p はデータとの A の**コンパチビリティ(相性の良さ)**を測定する」と言うこともできるし、主語が H と A の場合は「p は H と A を所与としたデータとの**コンパチビリティ(相性の良さ)**を測定する」と言うこともできる。
4. 検定がモデル全体（仮定
5. 例として、Oxford 2017 の「null」を参照のこと。形容詞 2. 値ゼロを持つか、関連する。名詞 1. ゼロ。Merriam-Webster 2017 の「null」を参照のこと。形容詞 6. の、である、または関連する。
6. Good (1956, pp. 1132; 1983, 146) は、この尺度をその平均 E{− log2(P)}、つまり P のシャノンエントロピー（有効な P の場合は、H と A が正しいときにそれ自体が最大になる）を中心として使用することを提案した。
7. 底10の対数では、単位は Hartleys と呼ばれる。H: β = 0 の場合、この尺度は β のログワースと呼ばれることがある。− log10(p) より大きい最初の整数は
8. P値を複合仮説 H に拡張するには、複雑な論理的および技術的な問題がある。例えば、Berger and Boos 1994; Berger and Hsu 1996; Bayarri and Berger 1999, 2000, 2004 を参照。一部の拡張は論理的な非整合性につながるという理由で拒否される可能性があるが、Schervish (1996) は H 上で p を最大化することに基づいた整合的な拡張も拒否している。ただし、その上限の負の対数は、
9. この文献で使用される用語では、P値は尤度原理（比例する尤度関数は同じ推論を生み出すべきであるとする原理）に違反する可能性がある。しかし、この原理に反対する人々は、いくつかの狭い 範囲外では、尤度関数を実際には十分統計量と見なさない。

参考文献

- Amrhein, V., Korner-Nievergelt, F., and Roth, T. (2017), "The Earth is Flat (p > 0.05): Significance Thresholds and the Crisis of Unreplicable Research," Peer J, 5, e3544.
- Amrhein, V., Trafimow, D., and Greenland, S. (2018), "Inferential Statistics are Descriptive Statistics," The American Statistician, this issue.
- Bayarri, M. J., and Berger, J. O. (1999), "Quantifying Surprise in the Data and Model Verification," in Bayesian Statistics 6, eds. J. M. Bernardo, J.O. Berger, A.P. Dawid, and A. F. M. Smith, Oxford, UK: Oxford University Press, pp. 53–82.
- Bayarri, M. J., and Berger, J. O. (2000), "Values for Composite Null Models," Journal of the American Statistical Association, 95, 1127–1142.
- Bayarri, M. J., and Berger, J. O. (2004), "The Interplay of Bayesian and Frequentist Analysis," Statistical Science, 19, 58–80.
- Benjamini, Y. (2016), "It's Not the P-values' Fault," The American Statistician, Online Supplement to ASA Statement on P-values. 70, online supplement 1, available at http://amstat.tandfonline.com/doi/suppl/10 (http://amstat.tandfonline.com/doi/suppl/10). 1080/00031305.2016.1154108/suppl_file/xxxx.
- Berger, J. O., and Sellke, T. M. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of P-values and Evidence" (with discussion), Journal of the American Statistical Association, 82, 112–139.
- Berger, J. O., and Wolpert, R. L. (1988), "The Likelihood Principle" (with discussion) (2nd ed.), IMS Lecture Notes-Monograph Series, 6, 1–199.
- Berger, R. L., and Boos, D. D. (1994), "P Values Maximized Over a Confidence Set for the Nuisance Parameter," Journal of the American Statistical Association, 89, 1012–1016.

- Berger, R. L., and Hsu, J. C. (1996), “Bioequivalence Trials, Intersection-Union Tests, and Equivalence Confidence Sets,” Statistical Science, 11, 283–319.
- Boos, D. D., and Stefanski, L. A. (2011), “P-Value Precision and Reproducibility,” The American Statistician, 65, 213–221.
- Box, G. E. P. (1980), “Sampling and Bayes Inference in Scientific Modeling and Robustness,” Journal of the Royal Statistical Society, Series A, 143, 383–430.
- Casella, G., and Berger, R. L. (1987), “Reconciling Bayesian and Frequentist Evidence in the 1-sided Testing Problem” (with discussion), Journal of the American Statistical Association, 82, 106–135.
- Casella, G., and Berger, R. L. (1987), “Comment,” Statistical Science, 2, 344–417.
- Cohen, J. (1994), “The Earth is Round ($p < 0.05$),” American Psychology, 47.
- Cox, D. R., and Donnelly, C. A. (2011), Principle of Applied Statistics, Cambridge, UK: Cambridge University Press.
- Cox, D. R., and Hinkley, D. V. (1974), Theoretical Statistics, New York: Chapman and Hall.
- Edwards, A. W. F. (1992), Likelihood (2nd ed.), Baltimore, MD: Johns Hopkins University Press.
- Fisher, R. A. (1925), Statistical Methods for Research Workers, Edinburgh, UK: Oliver and Boyd.
- Fraundorf, P. (2017), “Examples of Surprisal,” available at http://www.umsl (http://www.umsl). edu/~fraundorfp/egsurpri.html.
- Gelman, A. (2013), “P Values and Statistical Practice,” Epidemiology, 24, 69–72.
- Gelman, A., and Stern, H. (2006), “The Difference Between ‘Significant’ and ‘Not Significant’ is not Itself Statistically Significant,” The American Statistician, 60, 328–331.
- Gigerenzer, G. (2004), “Mindless Statistics,” Journal of Socio-Economics, 33, 587–606.
- Good, I. J. (1956), “The Surprise Index for the Multivariate Normal Distribution,” The Annals of Mathematical Statistics, 27, 1130–1135.
- Good, I. J. (1983), “Some Logic and History of Hypothesis Testing,” in Philosophical Foundations of Economics, ed. J. C. Pitt, Dordrecht: D. Reidel, pp. 149–174. Reprinted as Ch. 14 in Good, I.J. (1983), Good Thinking, Minneapolis, MN: University of Minnesota Press, pp. 129–148.
- Goodman, S. N. (1992), “A Comment on Replication, p-values and Evidence,” Statistics in Medicine, 11, 875–879.
- Goodman, S. N. (1999), “Towards Evidence-Based Medical Statistics, I: The P-value Fallacy,” Annals of Internal Medicine, 130, 995–1004.
- Greenland, S. (2004), “The Need for Critical Appraisal of Expert Witnesses in Epidemiology and Statistics,” Wake Forest Law Review, 39, 291–310.
- Greenland, S. (2017), “The Need for Cognitive Science in Methodology,” American Journal of Epidemiology, 186, 639–645.
- (2018), “The Unconditional Information in P-values, and Its Refutational Interpretation via S-values,” manuscript.
- Greenland, S., and Poole, C. (2013), “Living with Statistics in Observational Research,” Epidemiology (Cambridge, Mass.), 24, 73–78.
- Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.C., Poole, C., Goodman, S.N., and Altman, D.G. (2016), “Statistical Tests, Confidence Intervals, and Power: A Guide to Misinterpretations,” The American Statistician, 70, online supplement 1, available at http://amstat (http://amstat). tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108/suppl_file/ utas_a_1154108_sm5368.pdf ; reprinted in the European Journal of Epidemiology, 31, 337–350.
- Hoekstra, R., Finch, S., Kiers, H. A. L., and Johnson, A. (2006), “Probability as Certainty: Dichotomous Thinking and the Misuse of p-values,” Psychonomic Bulletin & Review, 13, 1033–1037.
- Hubbard, R., and Bayarri, M. J. (2003), “Confusion Over Measures of Evidence (p’s) Versus Errors (α’s) in Classical Statistical Testing,” The American Statistician, 57, 171–177.
- Hubbard, R., and Lindsay, R. M. (2008), “Why P Values Are Not a Useful Measure of Evidence in Statistical Significance Testing,” Theory & Psychology, 18, 69–88.
- Hurlbert, S. H., and Lombardi, C. M. (2009), “Final Collapse of the Neyman–Pearson Decision Theoretic Framework and Rise of the neoFisherian,” Annales Zoologici Fennici, 46, 311–349.

- Kuffner, T. A., & Walker, S. G. (2017), “Why Are p-values Controversial?” The American Statistician, in Press, 1.
- Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Arg-amon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., Cross, E. S., Daniels, S., Danielsson, H., DeBruine, L., Dunleavy, D. J., Earp, B. D., Feist, M. I., Ferrell, J. D., Field, J. G., Fox, N. W., Friesen, A., Gomes, C., Gonzalez-Marquez, M., Grange, J. A., Grieve, A. P., Guggen-berger, R., Grist, J., van Harmelen, A.-L., Hasselman, F., Hochard, K. D., Hoffarth, M. R., Holmes, N. P., Ingre, M., Isager, P. M., Isotalus, H. K., Johansson, C., Juszczyk, K., Kenny, D. A., Khalil, A. A., Konat, B., Lao, J., Larsen, E. G., Lodder, G. M. A., Lukavský, J., Madan, C.
- R., Manheim, D., Martin, S. R., Martin, A. E., Mayo, D. G., McCarthy, R. J., McConway, K., McFarland, C., Nio, A. Q. X., Nilsonne, G., de Oliveira, C. L., de Xivry, J.-J. O., Parsons, S., Pfuhl, G., Quinn, K. A., Sakon, J. J., Saribay, S. A., Schneider, I. K., Selvaraju, M., Sjoerds, Z., Smith, S. G., Smits, T., Spies, J. R., Sreekumar, V., Steltenpohl, C. N., Stenhouse, N., Świątkowski, W., Vadillo, M. A., Van Assen, M. A. L. M., Williams, M. N., Williams, S. E., Williams, D. R., Yarkoni, T., Ziano, I., & Zwaan, R. A.) (2018), “Justify Your Alpha: A Response to ‘Redefine Statistical Significance,” Nature Human Behaviour, 2, 168–171.
- Lane, D. (1988), “Discussion of Berger and Wolpert,” IMS Lecture Notes-Monograph, 6, 175–181.
- Lang, J. M., Rothman, K. J., and Cann, C. I. (1998), “That Confounded P-value,” Epidemiology (Cambridge, Mass.), 9, 7—8.
- LeCam, L. (1988), “Discussion of Berger and Wolpert,” IMS Lecture Notes-Monograph, 6, 182–185.
- Lehmann, E. L. (1986), Testing Statistical Hypotheses, New York: Wiley.
- Lindeman, M., & Stark, P. B. (2012), “A Gentle Introduction to Risk-limiting Audits,” IEEE Security & Privacy, 10, 42–49.
- MacKay, D. J. C. (2003), Information Theory, Inference, and Learning Algo-rithms, Cambridge, Cambridge University Press, sec. 2.4, available at http://www.inference.org.uk/mackay/itila/book.html (http://www.inference.org.uk/mackay/itila/book.html)
- McShane, B. B., and Gal, D. (2017), “Statistical Significance and the Dichotomization of Evidence” (with discussion), Journal of the American Statistical Association, 112, 885–908.
- McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (2018), “Abandon Statistical Significance,” The American Statistician, this issue.
- Merriam-Webster Dictionary (2017), “Null,” available at https://www (https://www). merriam-webster.com/dictionary/null.
- Murdoch, D. J., Tsai, Y.-L., and Adcock, J. (2008), “P-Values are Random Variables,” The American Statistician, 62, 242–245.
- Neyman, J. (1977), “Frequentist Probability and Frequentist Statistics,” Synthese, 36, 97–131.
- Oxford Living Dictionary (2017), “Null,” available at https://en (https://en). oxforddictionaries.com/definition/null.*
- Perezgonzalez, J. D. (2015), “P-values as Percentiles. Commentary on: ‘Null Hypothesis Significance Tests. A Mix-up of two Different Theories: the Basis for Widespread Confusion and Numerous Misinterpretations’,” Frontiers in Psychology, 6, 341.
- Poole, C. (1987a), “Beyond the Confidence Interval,” American Journal of Public Health, 77, 195–199.
- (1987b), “Confidence Intervals Exclude Nothing,” American Journal of Public Health, 77, 492–493.
- Ritov, Y., Bickel, P. J., Gamst, A. C., and Kleijn, B. J. K. (2014), “The Bayesian Analysis of Complex, High-Dimensional Models: Can It Be CODA?” Statistical Science, 29, 619–639.
- Robins, J. M., and Wasserman, L. (2000), “Conditioning, Likelihood, and Coherence: A Review of Some Foundational Concepts,” Journal of the American Statistical Association, 95, 1340–1346.
- Royall, R. R. (1986), “The Effect of Sample Size on the Meaning of Significance Tests,” The American Statistician, 40, 313–315.
- (1997), Statistical Inference: A Likelihood Paradigm, New York: Chapman and Hall.
- Schervish, M. J. (1996), “P-values: What They Are and What They Are Not,” The American Statistician, 50, 203–206.

- Sellke, T. M., Bayarri, M. J., and Berger, J. O. (2001), “Calibration of p Values for Testing Precise Null Hypotheses,” The American Statistician, 55, 62–71.
- Senn, S. J. (2001), “Two Cheers for P-Values,” Journal of Epidemiology and Biostatistics, 6, 193–204.
- (2002), “Letter to the Editor re: Goodman 1992,” Statistics in Medicine, 21, 2437–2444.
- (2008), Statistical Issues in Drug Development (2nd ed.), New York: Wiley.
- Shannon, C.E. (1948), “A Mathematical Theory of Communication,” Bell System Technical Journal, 27, 379–423, 623–656.
- Spanos, A. (2013), “Who Should Be afraid of the Jeffreys–Lindley Paradox?” Philosophy of Science, 80, 73–93.
- Walsh, P., Rothenberg, S. J., and Bang, H. (2018), “Safety of Ibuprofen in Infants Younger than Six Months: A Retrospective Cohort Study,” PLoS One, 13, e0199493, available at https://doi.org/10.1371/journal (https://doi.org/10.1371/journal). pone.0199493
- Wasserstein, R. L., and Lazar, N. A. (2016), “The ASA’s Statement on p-values: Context, Process and Purpose,” The American Statistician, 70, 129–133.
- Wellek, S. (2010), Testing Statistical Hypotheses of Equivalence and Noninfe-riority (2nd ed.), New York: Chapman & Hall.
- Ziliak, S. T., and McCloskey, D. N. (2008), The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives, Ann Arbor, MI: University of Michigan Press.