第 2 期 医薬安全性研究会
第 25 回定例会 2019 年 11 月 09 日

最近の P 値をめぐる話題のまとめ

山田 雅之

キッセイ薬品工業株会社

要約: 2019 年 6 月の朝日新聞「統計的有意は誤解の温床」という記事は，Nature 誌に掲載された 800 名を超える科学者が署名を行っている Comment "Retire statistical significance"を基にした解説であった．そのきっかけとなった 2016 年のアメリカ統計学会（American Statistical Association: ASA）による「統計的有意性と P 値に関する声明（ASA 声明）」には，P 値はデータと特定の統計モデルが矛盾する程度を示す指標の 1 つであり，調べている仮説が正しい確率やデータが偶然のみで得られる確率を示すものではないこと，統計的有意性が効果の大きさや結果の重要性を示さないこと，科学的な結論やビジネスや政策の決定は統計的有意性のみに基づくべきではないこと，といった原則がまとめられている．また，検定に代わる方法として推定やベイズ流の方法といった P 値以外のアプローチの例も示されている．このような動向を受けて，New England Journal of Medicine 誌は 2019 年 9 月に，多くの副次的評価項目やサブグループ解析での P 値の過剰使用を避ける点から，全般的および臨床試験報告について統計に関する投稿規定を改定している．

キーワード: P 値，統計的有意性，ASA 声明，New England Journal of Medicine，投稿規定

目 次

1. 2019 年 6 月の朝日新聞の記事---------- 3
2. 検定のロジック---------- 5
3. 2016 年アメリカ統計学会声明（ASA 声明）の概説---------- 8
4. P 値論争の歴史的経緯と P 値に関する総括 ----------11
5. New England Journal of Medicine の統計に関する投稿規定の改定----------13
6. まとめ----------15
7. 参考文献 ----------16

最近のP値をめぐる話題のまとめ

2019.11.9

キッセイ薬品工業　山田 雅之

2

本日の内容

- 2019年6月の朝日新聞の記事
- 2016年 アメリカ統計学会声明（ASA声明）の概要
- P値論争の歴史的経緯
- New England journal of Medicineの統計に関する投稿規定の改訂

3

2019年6月の朝日新聞

- 2019年6月20日の朝日新聞の朝刊に以下の記事が掲載された

「統計的に有意」は誤解の温床
「やめるべき考え方」論文に研究者800人賛同

Natureに掲載された以下のCommentを紹介した記事

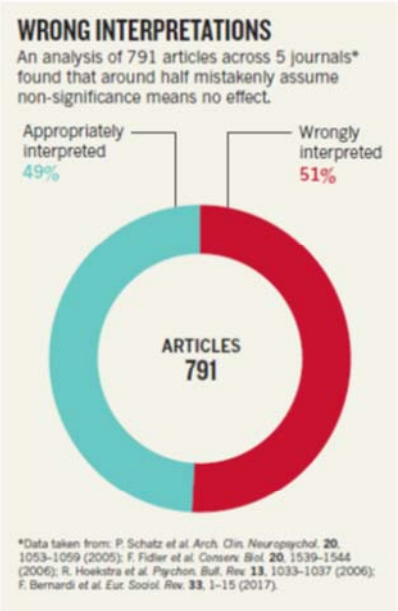
“Scientists rise up against statistical significance”
Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories
21MARCH2019, VOL.567,P305-307

2019 年 6 月 20 日の朝日新聞に，＜「統計的に有意」は誤解の温床，「やめるべき考え方」論文に研究者 800 人賛同＞という見出しで記事 [1)]が掲載された．記事の元となった論文は，2019 年 3 月の Nature 誌に掲載された"Scientists rise up against statistical significance"という Comment[2)]であり，このコメントには Supplementary Information として 800 人を超える研究者の署名が掲載されている．

4

Nature論文の主張

- 有意差なしを効果なしと解釈している例が多いことを指摘している
- 参加者の少ない試験では有意な差が出ない結果が出やすいが，そうした試験の立て付けの問題なのか，実際に薬の効果に差がないのかなどはわからない．にもかかわらず，有意でないことだけを根拠に「白か黒か」を言い切る例が多いことを警告している．
- ただし，P値の使用の禁止や，意思決定基準として使用を否定しているわけではない．



Nature誌のCommentでは，有意差なしを効果なしと解釈している事例が多いことを指摘している．

スライド右のグラフは論文中に示されていた図で，有意差検定の結果と解析結果の解釈を調査した5論文をまとめたものである．

791論文のうち，有意差なしの意味を効果なしと示している論文は51%であると報告されている．有意差が出ない原因は，効果が十分でない場合だけではなく，症例数が足りない場合もあげられる．そのような試験のデザインに起因する要因を検討せず，有意でないという結果のみを根拠として効果の有無を判定している事例が多いことを警告している．

ただし，一律にP値の使用の禁止を求めたり，有意差検定の結果に基づく意思決定を否定しているわけではない．

5

検定のロジック

- 背理法による証明
  - 示したいこと:「AとBは差がある」
  - 仮説:「AとBは差がない」
  - 推論:「データから起こりうる可能性を推定」
  - 結果:「仮説が起こる確率は**である」
  - 判断:「起こりうる確率は，設定値よりも低い」

    仮説は否定⇒差がある

    「起こりうる確率は，設定値よりも高い」

    仮説は否定できない⇒判定保留

ここで，有意差検定の考え方を振り返る．有意差検定はまず「差がない」という帰無仮説を置き，得られたデータから帰無仮説が起こる確率を算出する．次に事前に定めた偶然と考えられる最小の確率（有意水準）とその確率を比較して，有意水準を下回る場合には帰無仮説を否定し，差があると判定する．有意水準を下回らない場合は仮説を否定できないため判定保留とする．

6

論理学（命題）

ある命題	「Aならば，Bである」
命題の逆	「Bならば，Aである」
命題の裏	「Aでないなら，Bでない」
命題の対偶	「Bでないなら，Aでない」

命題が正しいとすると，どれが必ず正しくなるか？

Nature 誌の Comment で指摘している，「有意差がなければ差がない」と判断することが誤りであることを論理学により説明する．

ある命題があり，その命題に対して，逆，裏および対偶を設定する．命題が正しい場合，対偶のみが常に正しくなる．逆および裏は，正しい場合もあるが正しくない場合もある．

7

検定での命題

命題　「P＜0.05ならば，差がある」
逆　「差があれば，P＜0.05」
裏　「P≧0.05ならば，差がない」
対偶　「差がなければ，P≧0.05 」

P≧0.05の場合は，裏に当たるので常に正しくなるとはいえない

正しくならない例：例数が少ない，試験の精度が悪い（バイアス，ばらつき）

前のスライドの内容を有意差検定に当てはめて考える.

ここで命題を「P<0.05 ならば，差がある」とすると，常に正しい対偶は「差がなければ，P≧0.05」であり，Nature 誌の Comment で指摘されている「P≧0.05 ならば，差がない」は裏にあたる.

命題の裏は正しい場合もあるが正しくない場合もある．例えば例数が少ない，バイアスやばらつきにより試験の精度が悪いなどがもとで，本来は差があるにもかかわらず，P 値に基づく判断では差がないとしてしまうことがある.

8

ASA声明の概説

- The ASA's statement on p-values: Context, process, and purpose.（The American Statistical Association, 2016）
- 日本計量生物学会が2017年に訳を公開している.

背景

- 公表された多くの科学的結論の土台となっているのは「統計的有意性」という概念であり，通常P値と呼ばれる指標で評価される．P値は有用な統計指標ではあるが，誤用と誤解がまかり通っている.
- ASAの公式な声明により，P値の適正な使用と解釈の基礎にある広く合意された原則を明らかにすることで，科学界が利益をえると考えている.

P値とは

- 特定の統計モデルのもとで，データの統計的要約（たとえば，2グループ比較での標本平均の差）が観察された値と等しいか，それよりも極端な値をとる確率である.

2016 年に American Statistical Association (ASA)が P 値に関する声明（ASA 声明）[3)]を示している．2017 年に日本計量生物学会が ASA 声明の日本語訳 [4)]を作成し公開している．以後，日本計量生物学会の日本語訳から，ASA 声明の概要を説明する.

ASA 声明の背景で，多くの科学的結論の土台となっている P 値で表される「統計的有意性」について，誤用と誤解がまかり通っていることを指摘している．ASA 声明により適切な使用と解釈の基礎にある広く合意された原則を明らかにすることで，科学界が利益を得ることを目的としている.

ASA 声明で示す P 値は，「特定の統計モデルのもとで，データの統計的要約（たとえば，2 グループ比較での標本平均の差）が観察された値と等しいか，それよりも極端な値をとる確率である．」と定義している.

9

ASA声明の概説

原則

1. *P値はデータと特定の統計モデル（訳注: 仮説も統計モデルの要素のひとつ）が矛盾する程度をしめす指標のひとつである.*
   - P値が小さいほどデータの帰無仮説の統計的矛盾の程度が大きくなる. ただし, P値の計算の背後にある仮定がすべて正しければ, であるが.
2. *P値は, 調べている仮説が正しい確率や, データが偶然のみでえられた確率を測るものではない.*
   - P値は仮説やその計算の背後にある仮定に基づいたデータについての記述であり, 仮説や背後にある仮定自身についての記述ではない.
3. *科学的な結論や, ビジネス, 政策における決定は, P値がある値（訳注: 有意水準）を超えたかどうかにのみ基づくべきではない.*

10

ASA声明の概説

原則（続き）

4. *適正な推測のためには, すべてを報告する透明性が必要である.*
   - 見込みのありそうな結果をいいとこ取り—データのどぶさらい, 有意症, 有意クエスト, 選択的推論, P値ハッキングとも呼ばれる—すると, 出版された論文に統計的に有意な結果が誤って過剰に報告されるため, 厳に避けなければならない.
5. *P値や統計的有意性は, 効果の大きさや結果の重要性を意味しない.*
6. *P値は, それだけでは統計モデルや仮説に関するエビデンスの, よい指標とはならない.*
   - 背景情報やほかのエビデンスがなければ, P値は限られた情報しか提供しないことを研究者は認識すべきである.

11

ASA声明の概説

P値以外のアプローチ

- 信頼区間，信用区間，予測区間などの，検定よりも推定を強調した方法
- ベイズ流の方法
- 尤度比やベイズファクターなどのことなったエビデンスの指標
- 決定理論やFalse Discovery Rateといったアプローチ

結語

すぐれた研究デザインとその実施という原則，多様な数値およびグラフによるデータの要約，研究対象である事象の理解，背景情報に基づく結果の解釈，すべてを報告すること，そしてデータの要約の意味の適正な論理的かつ定量的理解．ひとつの指標が科学的推論の代わりとはなりえない．

ASA 声明では，P 値に関する 6 つの原則を通じて P 値のみで判定することの危険性を指摘している．ASA 声明では P 値以外のアプローチとして，検定より推定を強調した方法，ベイズ流の方法等を紹介している．これらの方法はすぐに P 値に代わるものではないが，多角的な面で科学的判断を行うための有用な方法であると考える．

ASA 声明は，「すぐれた研究デザインとその実施という原則，多様な数値およびグラフによるデータの要約，研究対象である事象の理解，背景情報に基づく結果の解釈，すべてを報告すること，そしてデータの要約の意味の適正な論理的かつ定量的理解．ひとつの指標が科学的推論の代わりとはなりえない．」という結語で締められている．

12

P値に関する議論

P値論争の歴史的経緯

時代	出来事
1986年頃	American Journal of Public HealthにおけるP値論争
2010年以降	P値をめぐるいろいろな非難
2014/2/13	Natureの記事で衝撃的なP値への非難
2015/2/12	Basic an Applied Social PsychologyでP値禁止の論説
2015/3/5	NatureでP値禁止がニュースに
2016/3/8	ASA声明
2016/3/10	NatureでASA声明がニュースに

折笠秀樹，薬理と治療，2018 一部改変

P値の議論は，ASA声明でいきなり行われたわけではなく，それ以前から行われていた．富山大学の折笠先生がその経緯をまとめられている[6]．

1986年ごろの論争では，最終的にP値を否定するまでには至らなかったが，P値より信頼区間のほうに軍配が上がり，British Journal of Medicine誌で，信頼区間の記載が義務付けられるきっかけとなった．その後論争は落ち着いたが，2010年にScience Newsに「統計解析での仮説検定は薄っぺらな土台に基づくもので，そのようなもので科学的方法といってよいのか」という内容の記事が掲載され，再度議論となった．極めつけは2014年2月13日のNature誌にP値を批判する論説が掲載され，この論説の中で「ベイズ因子」が提唱された．

2015年にはBasic and Applied Social PsychologyでP値禁止の論説が掲載され，これが2015年3月5日のNature誌に取り上げられことで，さらに話題となった．このような議論の中で，2016年3月8日にASAが緊急に統計的有意とP値に関する声明を発表し，これもすぐNature誌のコラムとして取り上げられ，広く周知されることとなった．以後，現在まで，統計的有意とP値に関する論争が続いている．

13

P値に関する総括

① カット点(P<0.05)はとくに意味はないので，目安として利用するのがよい．
P値は基本的に正確な値を示し,その値を見て,異常性(帰無仮説を棄却する可能性)については複数カット点で見てもよいだろう．

② P値についての誤解は，教育で払拭すべきである．
P値はFisherが初めて用いたものであり,「前提が正しい時に現データ(またはそれより極端データ)の出る確率」である．

③ P値の選択的報告は避けるべきである．
都合のよいP値を拾い,都合の悪いP値を捨てると,偽陽性へと結論を曲げることにつながる．

④ 研究計画，実施上のバイアスも勘案のうえ結論すべきである．
P値は偶然誤差を考慮するが,それ以外の系統誤差の可能性も判断して結論すべきである．

⑤ 統計的有意と臨床的有意は異なることを認識すべきである．
効果サイズ(推定値)にも注意を払うべきである．

折笠秀樹，薬理と治療，2018より

P値は誤解や誤用があるため批判もあるが，その信ぴょう性を見分けるための目安としては有用かもしれない．安易にP値を禁止するなどというのは無用な議論と言わざるを得ない．もっとフレキシブルに，必要に応じて信頼区間やP値を明記すればよいと思う．

P値については，2016年のSander Greenlandの総説やStuart Pocockが示した留意点などが示されているが，折笠先生はこれらをまとめて総括している．

なお，Pocockが示した留意点は参考資料として添付した．

14

NEJMの統計に関する投稿規定の改訂

- 2019年7月にThe New England journal of Medicineの統計に関する投稿規定が改訂され，その背景のEditorialsが示されている
- 近年議論となっているP値に関する議論に基づき，多くの副次評価項目やサブグループ解析でのP値の過剰使用を避けるため，報告方法を見直した
- 適切に設計されたランダム化研究または観察研究には，主要仮説と事前に指定された解析方法があり，その解析で得られたP値は，観察されたデータが介入と関連のない帰無仮説と矛盾する程度の信頼できる指標であるとしている

近年のP値に関する様々な議論を踏まえ，2019年にThe New England Journal of Medicine誌（NEJM）の統計に関する投稿規定[8]が改定された．また，改定の背景がEditorial[7]に示されている．

Editorialでは，ASA声明の原則4に示されている，多くの副次評価項目やサブグループ解析でのP値の過剰使用を避けるため報告方法を見直したこと，適切に設計されたランダム化研究または観察研究には，主要仮説と事前に指定された解析方法があり，その解析で得られたP値は，観察されたデータが介入と関連のない帰無仮説と矛盾する程度の信頼できる指標であることが述べられている．

15

NEJMの統計に関する投稿規定の改訂

"Statistical Reporting Guidelines"の全般的事項及び臨床試験報告の主な事項

- すべての原稿の”Method”には，研究のサンプルサイズと検出力に関する簡単な説明と，主要及び副次解析の簡単な説明を含める必要がある．
- すべての原稿の”Method”には，**欠測データの処理方法の説明を含める必要がある．**欠測がまれでない限り，一般にComplete Case解析は主要解析として受け入れられないため，欠測メカニズムを考慮して，適切な方法に置き換える必要がある．
- 有意差検定には，推定効果の大きさ，関連性の尺度，またはその他の関心のあるパラメータの信頼区間を添付する必要がある．**信頼区間は，対応する検定で有意水準に対して行われた調整と一致するように調整する必要がある．**

NEJM の統計に関する投稿規定は，すべての試験に関する内容，臨床試験に関する内容および観察研究に関する内容に分かれて記載されている．

本資料では，すべての試験に関する内容と臨床試験に関する内容について主な改定個所を示した．

16

NEJMの統計に関する投稿規定の改訂

"Statistical Reporting Guidelines"の全般的事項及び臨床試験報告の主な事項

- オリジナルおよび最終のプロトコルと解析計画書（SAP）を原稿とともに提出し，プロトコルとSAPに加えられた変更の表と変更の日付と内容を示す．
- 臨床試験の結果を報告する原稿の主要な解析結果は，特別な状況を除き，元のプロトコルで事前に指定された解析と一致する必要がある．**プロトコルと一致しない解析は，原稿の”Method”で正当化される必要がある．**
- 臨床試験のプロトコルまたはSAPで，多数の推論を調整する方法または誤検出率を制御する方法が指定されていない場合，**すべての副次エンドポイントおよび探索エンドポイントの報告は，治療効果の点推定値と95%信頼区間に限定する必要がある．**
  そのような場合，”Method”で信頼区間が多重性に合わせて調整されておらず，示された推論が再現できない可能性があることに注意を促す．**これらの分析ではP値は報告されない．**

17

まとめ

- 有意差検定を用いたP値による結果の解釈は誤用があることが広く認知されている
- そのため，P値ではなく推定値を用いた結果の表記が主要な雑誌の投稿規定でも求められるようになってきている．
- しかし，有意差検定を一律に否定するものではなく，規制当局のDecision Making等では，適切にデザインされた試験の主要な結果を有意差検定によるP値で評価することは，従来通り受け入れられている．

18

参考資料

1. 朝日新聞記事（2019年6月20日）
https://www.asahi.com/articles/DA3S14062710.html
2. V Amrhein, et al., Scientists rise up against statistical significance, Nature 2019; 567: 305-307.
3. RL Wasserstein, and NA Lazar, “The ASA's statement on p-values: context, process, and purpose”, The American Statistician 2016; 70: 129-133.
4. 統計的有意性とP値に関するASA声明, 日本計量生物学会
http://www.biometrics.gr.jp/news/all/ASA.pdf
6. 折笠秀樹, “P値論争の歴史”, 薬理と治療 2018; 46: 1273-1279.
7. D Harrington et al., “New Guidelines for Statistical Reporting in the Journal”, New England Journal of Medicine 2019; 381: 285-286.
8. Statistical Reporting Guidelines, New England Journal of Medicine
https://www.nejm.org/author-center/new-manuscripts

19

参考資料

統計的有意とP値に関する指針（Pocock）

① P値は実値(actual P value)を記載する.

② P値と信頼区間のあいだの関係を理解する.

③ いつも両側検定を使う.
ただし,片側検定を使うこともある. そのときは,1−sidcd P=0.04などと書く.

④ P値が小さいということは,(帰無仮説の下で)偶然には起こりえないことが起きたことを意味する.
しかし, もしかすると計画. 実施上に生じたバイアスかもしれない.

⑤ 統計的有意と臨床的有意を区別する.

⑥ 少数例のときは非有意でも臨床的意義のことはある(過小検出)が,多数例のときは臨床的意義はなくても有意になることはある(過剰検出).

⑦ 非劣性試験で統計的有意というのは,優越性の統計的有意とは異なるので注意が必要である.

Pocock SJ et al., Making sense of statistics in clinical reports. Part I. J Am Coll Cardiol 2015; 66:2536-49.

折笠秀樹, 薬理と治療, 2018より