心理学研究における統計改革の進展状況について ： p 値・効果量・信頼区間から

心理学研究における統計改革の進展状況について
－p 値・効果量・信頼区間から－

土　居　淳　子

Ⅰ　はじめに

「統計的仮説検定」あるいは「有意性検定」として広く用いられている統計的手法は，フィッシャーが展開した有意性検定論とネイマンとピアソンが展開した統計的仮説決定理論とを無理やり合体させ，統計ユーザー向けに折衷案として作られたハイブリッド仮説検定である。このハイブリッド仮説検定は、その誕生直後から多くの批判にさらされてきた（これについての詳細は，土居（2010）で述べた）。直近では，2016 年にアメリカ統計学会（American Statistical Association, ASA）が「統計的有意性と p 値に関する声明」を発表し（Wasserstein&Lazar, 2016），有意性検定の誤用と誤解に警鐘を鳴らしている。また，2019 年 3 月には，統計的に有意かどうかだけで結果を二者択一的に選別することに対する反対意見が，800 人以上の科学者の署名入りで Nature 誌に投稿された（Amrhein ら，2019）。このように，さまざまな分野で有意性検定偏重からの脱却を目指す「統計改革」が進行中であるが，以下では心理学における統計改革に限定して話を進めたい。

Cohen（1994）が “The earth is round（p<.05）” と題する論文において有意性検定を辛辣に批判したことがきっかけとなり，アメリカ心理学会（American Psychological Association, 以下 APA と略す）は,「有意」か「有意でない」かに注目する有意性検定一辺倒の姿勢から, 信頼区間を用いた推定重視に舵を切った。

1996 年には，APA の推測統計に関する専門委員会が心理学研究において推定・検定を使う場合のガイドラインを委員会提案として発表し，（1）有意性検定を行う際には検出力を考慮して標本の大きさを事前決定すること，（2）有意確率（p 値）だけでなく効果量を明記すること，（3）信頼区間を図示することなどをより適切な推測統計の方法として推奨した（Wilkinson ら，1999）。

この報告を受けて 2001 年に改訂された APA Publication Manual 第 5 版では，効果量，エラーバーを含む図，信頼区間による報告を推奨し，特に信頼区間の利用を「最適な報告戦略（the best reporting strategy）」であると述べた。さらに，2009 年に発表された第 6 版では，検出力の重要性を明記するとともに，効果量と信頼区間による結果報告を強く推奨し，可能な限り推定値と区間推定に基づいて結果を考察することを指示した。

このような APA の方針に呼応して，国内においても「心理学における統計改革」が始まった。第 6 版の日本語訳が 2011 年に出版され，発達心理学会の「論文原稿 作成のための手引き」（2013 年 7 月改訂版），日本心理学会の「執筆・投稿の手びき」（2015 年改訂版）などが APA に準拠する方向で改訂され，論文誌に掲載される論文や学会でのポスター発表などで，有意性検定の結果とあわせて効果量や信頼区間が報告されるようになった。

大久保（2009）によれば, 1982 年から 2008 年の『基礎心理学研究』に掲載され実証的検討を含んだ 199 編の論文のうち, 信頼区間を報告した論文はわずか 1 編, 効果量を報告した論文も 10％程度であった。しかも，効果量を報告した論文のほとんどは相関係数や回帰係数など従来から効果量自体が興味の対象になってきたものであり，それ以外のケースは分散分析の効果量 ω^2 を報告した論文 1 編のみだったという。

第 6 版（英語版）が出版されてから 10 年が経ち，本年 10 月には第 7 版が出版される。国内の統計改革はどの程度進展したのだろうか。本稿では，第 6 版で示されている推測統計の結果報告の方法を紹介した上で，日本心理学会が発行する 2017 年（88 巻 1 号）から 2019 年（90 巻 3 号）までの『心理学研究』に掲載された研究論文（原著論文，研究資料，研究報告）を対象に，有意確率（p 値），効果量，信頼区間がどのように扱われているかを集計し，その結果をもとに，

国内の「心理学における統計改革」の進展状況を考える。

Ⅱ　APA 第 6 版でどのように述べられているか

APA Manual 第 6 版では統計的仮説検定について次のように述べている。すでに土居（2010）で紹介しているが，ここで改めて再掲したい。

> 歴史的に心理学研究者は，多くの（しかし，全てではない）分析的なアプローチの出発点として帰無仮説有意性検定（NHST）を強く信頼してきた。APA は，NHST が単に出発点でしかないこと，分析結果の意味するところを完全に伝えるためには，効果量，信頼区間，詳細な記述を付加して報告することが必要であることを強調する。各論文誌が NHST を重視する程度は，それぞれの編集者が決めることである。しかしながら，検定を行ったすべての仮説と適切な効果量および信頼区間の推定結果を報告することは，すべての APA 論文誌にとって最低限の期待である。科学研究者は，研究結果を正確かつ信頼できる形で報告することに対して常に責任を負っている。（APA , 2009, p.33，筆者訳）

具体的には，*t* 検定や *F* 検定などの推測統計の報告には，検定統計量，自由度，正確な *p* 値，効果量を示すこと，効果量の推定値の精度を示すために，可能な限り信頼区間を示すこと，また，標本平均や回帰係数の推定値を報告する場合は標準誤差などによって推定精度を示すことを求めている（APA, 2009, p.34）。以下では，正確な *p* 値の報告，効果量，信頼区間それぞれについて簡単に紹介する。

1. 正確な *p* 値

有意性検定にもとづく *p* 値を報告する場合は，正確な *p* 値を報告することが求められている（APA, 2009, p.113）。例えば，無相関検定の結果を文章中で報告する場合，

$$r(24) = -.43,\ p = .028$$

と書く。

従来は，*p* 値そのものを報告せずに，どの有意水準で有意になるかを示していた。上記の例の場合，

$$r(24) = -.43,\ p < .05$$

となる。この表記方法では，*p* 値の小ささに応じて，*p*<.05，*p*<.01 あるいは *p*<.001 と適用する有意水準が変わる。有意水準によって *p* 値がどれくらい小さかったか（＝帰無仮説を棄却することの確からしさ）を示す方法である。APA 第 6 版では，このような方法で *p* 値を報告する慣行は，限られた臨界値の分布表しか利用できなかった時代の産物である（p.114）とし，正確な *p* 値を報告することを求めている。

例外として，相関行列の表など *p* 値の正確な値を記載することで表が分りにくくなる場合は，旧来の「*p*<」形式を採用してもよいとしているが，その場合においても結果を本文で論じる場合は正確な *p* 値を報告することを求めている。

2. 効果量と信頼区間

p 値は「帰無仮説を棄却することの確からしさ」の指標であり，「効果の大きさ」を表す指標ではない。よく知られているように，実質的な差がほとんどない場合でも，サンプルサイズ（標本数）が大きくなればなるほど，*p* 値は小さくなり，統計的に有意であるという結果が得られやすくなる。

心理学では，統計的仮説検定の有意水準を慣習的に 5％に設定しており，有意確率が 5％より小さいかどうかで，有意かどうかを判定している。これは，効果量（＝実質的な差の大きさ）の 95％信頼区間に帰無仮説で設定する値（多くの場合は 0）が含まれているかどうかを見ることと同値であるので，95％信頼区間を示すことで，5％水準で有意かどうかを判定することができる。さらに，効果量の大きさとその推定精度（どのくらいの誤差を含んでいるか）も把握することができる。したがって，効果量とその 95％信頼区間を報告すれば，仮説検定の結果を報告する必要はないことになる。

APA 第 6 版では，適切な効果量を報告すること，可能な限り信頼区間も報告することを求め，1 標本の平均，2 群の平均値差，ピアソンの積率相関係数，回帰係数と決定係数，オッズ比などに対して，効果量とその信頼区間を報告する例が具体的に示されている。一方で，分散分析については，実験デザインによっては信頼区間の算出が必ずしも容易ではないためか，効

果量のみの報告例を示すに留まっている。

本文で報告する場合として，たとえば，次のような例が示されている。

> For immediate recognition, the omnibus test of the main effect of sentence format was statistically significant, $F(2, 177)=6.30$, $p=.002$, est $\omega^2=.07$. The one-degree-of freedom contrast of primary interest (the mean difference between Conditions 1 and 2) was also statistically significant at the specified .05 level, $t(177)=3.51$, $p<.001$, $d=0.65$, 95% CI [0.35, 0.95]. (APA, 2009, p.128)

また，効果量や信頼区間を示さない場合は，読者が必要に応じて効果量や信頼区間を算出できるように，セル平均，標準偏差，サンプルサイズ，相関などの基本的な統計量を報告することが求められている（APA, 2009, p.116）。独立測定のみの分散分析の場合，F 値や自由度とあわせて平均平方誤差 MSE が分れば，信頼区間を算出することができる（大久保・岡田, 2012）。

Ⅲ　日本心理学会の「執筆・投稿の手びき」でどのように述べられているか

国内の心理学系論文誌の多くが準拠する日本心理学会の「執筆・投稿の手びき」（2015 年改訂版）では，統計改革の流れがどのように反映されているだろうか。「1.2 論文の構成」において，APA 第 6 版に準拠する形で効果量と信頼区間の報告の重要性が示され，「3.4.3 統計記号，その他」において，正確な p 値，効果量および信頼区間による報告が，次のように指示されている。

> 検定結果については，t，F，χ^2 などの検定統計量の値，自由度，p 値，および効果量と効果の方向を記述する。点推定値（標本平均や回帰係数など）を示す場合には，推定精度に関する情報（標準誤差など）をあわせて示す。論文中では一貫した有意水準によって信頼区間を表示することが望ましい。（日本心理学会，2015，p.29）

しかし，具体的な記述例は，下記の通り，旧来の方法が示されている。正確な p 値，効果量，信頼区間のいずれも報告されていない。

> $F(1, 10)=6.18$, $p<.05$; $F(4,40)=22.71$, $p<.01$, $MSe=.005$
> $t(22)=6.16$, $p<.01$
> $\chi^2(4, N=90)=10.51$, $p<.05$

正確な p 値を報告するべきかどうかについては，意見が分かれているように見える。発達心理学会の「論文原稿作成のための手引き（2013 年 7 月改訂版）」の「4.1 検定結果の表記」においては，p 値ではなく有意水準による報告を指示している（別の箇所で、正確な p 値を報告することも認めている）。

> t，F，χ^2 検定の場合は，検定の手法，t，F，χ^2 の各値，自由度（χ^2 検定の場合はサンプルサイズも），有意水準。また，できる限り効果量も記載する。（発達心理学会，2013，p.6）

なお、発達心理学会の手引きには、分散分析の結果の報告例として効果量を含めた例が、日本教育心理学会（2017）の「日本教育心理学会 論文作成の手引き」には信頼区間の表記方法が具体的に示されている。

Ⅳ　『心理学研究』vol.88（1）～ vol.90（3）における報告の状況

2017 年度（88 巻 1 号）から 2019 年度半ば（90 巻 3 号）までの『心理学研究』に掲載された論文 148 編のうち，推測統計にもとづく分析結果を報告している 138 編について，p 値，効果量，信頼区間（CI）が報告されているかどうかを集計した。

表 1 は年度別に，表 2 は論文種類別に集計した結果である。正確な p 値の報告については，p 値が 0.001 より小さい場合を除き p 値そのものが報告されている場合を「あり」，有意でない場合のみ p 値が報告されている場合を「一部あり」，従来の方式（有意水準によって p 値の小ささの程度を示す）を用いている場合を「なし」とした。効果量と信頼区間については，論文中に少なくとも 1 つ報告されている場合を「あ

表 1 『心理学研究』における p 値、効果量、信頼区間の報告状況（年度別）

	対象論文数	正確な p 値の報告			効果量		信頼区間（CI）	
		あり	一部あり	なし	あり	なし	あり	なし
2017 年	55	10（18.2%）	5（9.1%）	40（72.7%）	29（52.7%）	26（47.3%）	11（20.0%）	44（80.0%）
2018 年	55	15（27.3%）	5（9.1%）	35（63.6%）	39（70.9%）	16（29.1%）	23（41.8%）	32（58.2%）
2019 年	28	8（28.6%）	1（3.6%）	19（67.9%）	20（71.4%）	8（28.6%）	9（32.1%）	19（67.9%）
計	138	33（23.9%）	11（8.0%）	94（68.1%）	88（63.8%）	50（36.2%）	43（31.2%）	91（68.4%）

表 2 『心理学研究』における p 値、効果量、信頼区間の報告状況（論文種類別）

	対象論文数	正確な p 値の報告			効果量		信頼区間（CI）	
		あり	一部あり	なし	あり	なし	あり	なし
原著論文	52	16（30.8%）	5（9.6%）	31（59.6%）	39（75.0%）	13（25.0%）	19（36.5%）	33（63.5%）
研究資料	40	6（15.0%）	0（0.0%）	34（85.0%）	22（55.0%）	18（45.0%）	14（35.0%）	26（65.0%）
研究報告	46	11（23.9%）	6（13.0%）	29（63.0%）	27（58.7%）	19（41.3%）	10（21.7%）	36（78.3%）
計	138	33（24.8%）	11（8.0%）	11（66.9%）	88（63.8%）	50（36.2%）	43（31.2%）	95（68.8%）

り」，全く報告されていない場合を「なし」とした。ただし，無相関検定における相関係数や回帰分析における回帰係数など，従来から報告することが慣例化している効果量のみが報告され，かつ，信頼区間を伴っていない場合は，効果量が報告されているとはカウントしなかった。

2017 年度から 2018 年度にかけて，p 値，効果量，信頼区間の報告が増えているが，2019 年度は足踏みしており，2019 年度においても効果量の報告が 7 割程度，正確な p 値や信頼区間の報告は 3 割程度にとどまっている（表 1）。論文種類別では，原著論文とそれ以外で差が生じており，研究資料や研究報告の半数近くが効果量の報告を行なっていない（表 2）。

表 3 は，p 値，効果量，信頼区間の報告の有無によって，報告タイプ別に集計したものである。この表では，有意でない場合のみ p 値を報告している論文は p 値報告なしとして集計している。「p 値あり」「効果量あり」「信頼区間あり」は全体の 9.4%にとどまっているが，「p 値なし」「効果量あり」「信頼区間あり」も 21.7%あり，効果量と信頼区間を報告している論文が全体の 30%程度あることがわかる。一方，「p 値なし」「効果量なし」「信頼区間なし」の従来のスタイルは 33.3%であった。

以上の結果は，論文中で効果量や信頼区間が少なくとも 1 つ報告されていれば「あり」として集計した結果であるが，信頼区間を報告している 43 編の論文のうちの 11 編は，級内相関，間接効果の推定値，モデルの適合度指標などの信頼区間のみを報告し，t 検定や分散分析の効果量、あるいは直接効果の推定値など、報告が期待されるその他の効果量の信頼区間は報告されておらず，適切に報告されているとは言い難い状況であった。

表 3 『心理学研究』における推測統計の報告タイプ

p 値	効果量	CI	論文数	割合
○	○	○	13	9.4%
○	○	×	16	11.6%
○	×	×	4	2.9%
×	○	○	30	21.7%
×	○	×	29	21.0%
×	×	×	46	33.3%
	計		138	100.0%

○は報告あり、×は報告なし

表 4 は，平均値差の検定や回帰分析などの心理学研究でよく用いられる分析手法について，p 値，効果量，信頼区間の報告状況を集計したものである。2 群の平均値差の分析における効果量の報告率は 72.2%であり，その内訳は，Cohen の d（23 件），Hedges の g（1 件），r（2 件）であった。一要因被験者間分散分析における報告率は 55.6%，η^2（6 件），偏 η^2（4 件）であった（一要因被験者間の場合，この 2 つはおなじものである）。その他の分散分析における報告率は 68.7%で，η^2（10 件），偏 η^2（22 件），一般化 ω^2（1 件）であった。η^2 を報告した 10 件のうち 6 件は参加者内を含む混合計画であった。平均値差の分析や分散分析におい

表 4 『心理学研究』における効果量、信頼区間の報告状況（分析方法別）

分析方法	効果量		効果量の CI の報告
	あり	なし	
2 群の平均値差（対応なし、あり）	26（72.2%）	10（27.8%）	5（13.9%）
分散分析			
一要因被験者間	10（55.6%）	8（44.4%）	1（5.6%）
その他	33（68.7%）	15（31.3%）	3（6.3%）
無相関検定	72（100%）	0（0%）	8（11.1%）
重回帰分析	29（100%）	0（0%）	1（3.4%）
独立性の検定（クロス表）	3（27.3%）	8（72.7%）	0（0%）

ては，ほとんどの場合に記述的な Cohen の d，η^2 および偏 η^2 が用いられていることがわかる。

相関分析および重回帰分析では，効果量（相関係数 r や偏回帰係数 β）は 100%報告されているが，独立性の検定（クロス表）で効果量が報告されることは少なく（27.3%），報告された 3 件はすべてクラメールの V であった。

信頼区間はすべての場合に 95%信頼区間が報告されていたが，どの分析方法においても報告件数は少ない。実証的研究の基盤となるこれらの分析において，信頼区間の導入が特に遅れていることがわかる。一般的な統計ユーザーの多くが用いている分析ソフトウェア SPSS では信頼区間がダイレクトに出力されないからかもしれない。信頼区間を報告している論文の多くは，R，SAS，Mplus などを用いていた。なお，信頼区間を報告していない論文の多くは，分散分析における誤差の平均平方（MSE）や回帰分析における回帰係数の標準誤差（SE）など，信頼区間の算出に必要な情報を報告していなかった。

V　おわりに

1982 年～ 2008 年に『基礎心理学研究』に掲載された論文のなかで，信頼区間を報告しているのは 1 編のみ，効果量の報告も 10%程度に留まっていたことと比較すると（大久保，2009），2017 年度以降の『心理学研究』での報告は格段に増加している。しかし，信頼区間や正確な p 値の報告はいまだ少なく，旧来の方法を使い続けている論文が 3 分の 2 以上を占めている。とくに信頼区間の報告は，実質的にはあまり普及していないことがわかった。また，効果量の報告は増えてきているが，結果の考察や解釈は従来通りの有意か有意でないかのみに依拠し，効果の大きさに言及していないケースも散見される。形式的な変革は浸透しつつあるが，本来の目的である「有意性検定一辺倒の姿勢から，信頼区間を用いた推定重視に」という本質的な変革はまだ始まったばかりという段階であろう。

本来，効果の大きさ（効果量）とその推定精度（信頼区間）で母集団の特性を推測する方法は，有意性検定よりも直感的で理解しやすいものである。また，波田野ら（2015）が示したように，p 値と効果量の大きさは必ずしも対応していない。効果が小さいにもかかわらずサンプルサイズが大きいために有意と判定されているケース，効果が小さくないにもかかわらずサンプルサイズが小さいために有意ではないと判定されているケースが少なからず存在する。今回の調査においても，同様の状況が散見された。

繰り返し指摘されているように，有意性検定には多くの問題がある。今後，検定力分析，効果量とその信頼区間を用いた優れた論文が多く発表され，標準的な分析ソフトウェアで平易に効果量や信頼区間が算出できるようになり，統計ユーザーを対象とした授業や解説書がリニューアルされることで，心理学における統計改革が本当の意味で浸透することを期待したい。

引用・参考文献

American Psychological Association（2009）*Publication Manual of the American Psychological Association*（6th ed.）, Washington, DC, American Psychological Association

American Psychological Association（2010）*Concise rules of APA style（6th ed.）*, Washington, DC, American Psychological Association

アメリカ心理学会（2011）APA 論文作成マニュアル 第 2 版（前田・江藤・田中訳） 医学書院

土居淳子（2010）帰納的推論ツールとしての統計的仮説検定―有意性検定論争と統計改革― 年報人間関係学 13, 5-36

Amrhein, V., Greenland, S. & McShane, B.（2019）Scientists rise up against statistical significance, Nature 567, 305-307

波田野結花・吉田弘道・岡田謙介（2015）『教育心理学研究』における p 値と効果量による解釈の違い 教育心理学研究 63, 151-161

日本発達心理学会（2013）「論文原稿 作成のための手引き」（2013 年 7 月改訂版）日本発達心理学会 HP http://www.jsdp.jp/contents/~cmhenshu/paper/kitei.html （閲覧日：2019 年 9 月 16 日）

日本教育心理学会（2017）「日本心理学会 論文作成の手引き」日本教育心理学会 HP https://www.edupsych.jp/toukou/ （閲覧日：2019 年 9 月 16 日）

日本心理学会（2015）「執筆・投稿の手びき」（2015 年改訂版）日本心理学会 HP https://psych.or.jp/manual/

大久保街亜（2019）日本における統計改革―基礎心理学研究を資料として― 基礎心理学研究 28, 88-93

大久保街亜・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

Wasserstein, R.L. & Lazar, N.A.（2016）The ASA Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133

Wilkinson, L. and the Task Force on Statistical Inference（1999）Statistical methods in psychology journals: Guideline and explanations, American Psychologist, 54, 594-604