Japanese Psychological Review
2016, Vol. 59, No. 1, 57–67

帰無仮説検定と再現可能性

大久保　街　亜
専修大学

Null hypothesis significance testing and reproducibility

Matia OKUBO
Senshu University

Although null hypothesis significance testing has been strongly criticized for decades, it has been the dominant statistical method in the field of psychology. Non-reproducibility of findings in psychology can be attributed, at least partially, to an arbitrary threshold (i.e., .05) in the null hypothesis significance testing and overrepresentation of *p*-values. The present study surveyed papers from the *Japanese Journal of Social Psychology* and examined whether or not such overrepresentation also existed among psychology researchers in Japan. Effect size measures and *p*-values did not correspond well when *p*-values were set at around .05. Moreover, the frequency of *p*-values just below .05 was greater than expected. These results imply that the overrepresentation of *p*-values can produce unreliable and irreproducible results. Two types of remedies are discussed to alleviate the problems of overrepresentation of the *p*-values.

Key words: null hypothesis significance testing, *p*-value, effect size, Bayesian statistics
キーワード：帰無仮説検定，*p* 値，効果量，ベイズ統計

1. はじめに

再現可能性は科学研究の根幹をなす特徴のひとつである。心理学を含めさまざまな実証科学では，諸条件を人為的に統制・操作し従属変数を測定する。その際，およそ同一の統制や操作がなされたとき一致した測定結果が得られる程度が再現可能性である。古典的な科学観に基づくと，再現可能性は科学に必要不可欠な条件となる。例えば，Popper（1959）は，「再現できない 1 回限りの出来事は科学にとって重要ではない（p. 66，拙訳）」と述べた。

ただし，現代の科学において再現可能性の重要度は分野により違いがある。上に紹介した Popper の主張はかなり極端なものである。実際，再現可能性が要求されないデータもある。例えば，惑星の衝突や火山の噴火など自然現象をデータとする場合，再現可能性を要求することは不可能である。また，現実の研究にはさまざまな要因が関わるため，およそ同一の統制や操作を再現することが困難なことがある。特殊な機器や手続きを採用した実験は簡単に再現できるものではない。実際，Rao（1985）が指摘するように，厳密な意味での再現可能性は，ほとんどの科学分野で明示的に検証されることはない。そのためもあり，しばしば再現可能性の低い（あるいはない）研究結果が報告されることになる。

近年，科学研究における再現可能性の低さはさまざまな分野で問題となっている。心理学も例外ではない。例えば，2012 年に *Perspectives on Psychological Science* は，心理学における再現可能性について特集を組んだ。2015 年に Brian Nosek らは，心理学の実験や調査を追試する大規模な再現可能性検討プロジェクトを行い，*Science* に報告した（Open Science Collaboration, 2015）。この結果は衝撃であった。著名な心理学の論文誌に掲載された研究結果の再現可能性は 39% 足らずであった。再現可能性が実証科学における証拠の信頼性（ここでは用語として credibility を想定している）を保証する条件なら，多くの心理学の研究結果はその条件を満たさないことになる。Nosek らはこのような低い再現可能性の原因として，出

版バイアス，検定力の低い研究計画，新しい技術に関する知識不足，出版への過度な圧力を指摘した（Open Science Collaboration, 2015）。他にもQRPs（Questionable Research Practicesの略，問題のある研究実践。本特集における池田・平石（2016）ならびに藤島・樋口（2016）を参照）や，データの捏造や改ざんも再現可能性を低める要因となるであろう。

再現可能性に関わる要因のうち出版バイアス，検定力，知識不足，QRPsのそれぞれには，直接的，間接的に心理統計が関わっている。これらの要因は全てデータの解釈に関わる。現代の実証的心理学においてデータから解釈を導く過程に統計的手法は不可欠である。そのため，これらの要因は必然的に心理統計と関わることになる。本論文では，心理統計の観点から心理学における再現可能性について検討を行う。まず，再現可能性を低める要因として心理学において過度な帰無仮説検定への依存があることを指摘する。そして，その問題を検討し，対策を提案する。

1.1　帰無仮説検定の原理とその問題点

帰無仮説検定は，現在の心理学において支配的な統計手法となっている。Lakatos（1978）が述べたように，帰無仮説検定はPopper（1956）の反証主義（方法論的反証主義）のアイデアと一致する論証形式を有する。つまり，方法論的反証主義の論理に確率的な表現を加えたものが，帰無仮説検定である（ただし，年代としては帰無仮説検定の成立は反証主義より先である）。Popperの方法論的反証主義では，仮説が真という前提のもとで生じないことが起これば（反証が得られれば），その仮説は棄却される。棄却されないあいだ，仮説は暫定的に支持される。一方，帰無仮説検定では，帰無仮説が真という前提のもと，観察される差（あるいは効果）が得られる確率が計算される（正確に述べると，求められた検定統計量より極端な値が得られる確率が計算される）。そして，その確率がある値より低ければ，帰無仮説は棄却され暫定的に差（あるいは効果）があると判断する。これは，「帰無仮説に対する反証が得られれば，帰無仮説は棄却される」ことと実質的に同じである。すなわち，反証主義の論理と一致する。

帰無仮説検定は反証主義と同様に後件否定の論証形式を用いた強力な推論である。科学研究の標準的方法である仮説演繹法を単純に適用した場合，後件肯定の誤謬を防げない。帰無仮説検定と反証主義は，後件否定の論証形式を導入することにより，論理的な破綻を逃れている。さらに，帰無仮説検定では，確率的な評価を行うことで，真偽の判定が可能な命題でなく，ある種の傾向や頻度を表す確率的な事象を扱うことができる。この点は方法論的反証主義よりも優れている。科学研究におけるデータは，多くの場合，仮説から演繹され論理的に導かれる命題の形となることは稀である。帰無仮説検定はそのような点で現実にも則した方法論でもある。

帰無仮説検定は1970年以降，ほぼすべての実証的な心理学の研究で用いられてきた。この結果は国際的な文献に関する調査（Hubbard & Ryan, 2000）でも，日本国内の文献に関する調査（Omi & Komata, 2005；大久保，2009）でも確認された。

このような支配的な状況にも関わらず，帰無仮説検定にさまざまな問題が指摘されてきた。帰無仮説検定の問題を指摘する論文は，そのアイデアが発表された直後の1940年代から存在し，1990年代には170編を超える批判が論文として出版された（Anderson, Burnham, & Thompson, 2000）。さまざまな研究者が長らくしかも激しく帰無仮説検定を批判してきたことを考えると（レビューとして，Cohen, 1994；Harlow, Mulaik, & Steiger, 1997；Kline, 2005；Krueger, 2001；Loftus, 1996；大久保・岡田，2012），この支配的な状況は驚くべきものだ。帰無仮説検定が抱える問題は多岐にわたる。例えば，帰無仮説検定における二値的な判断は，かのRonald A. Fisherが強く問題視したことで知られる（Fisher, 1959）。有意水準を分水嶺として，結果の有意か否かにデジタルに二分することには多くの批判がある。また，標本サイズによりp値が変化するため，注意して検定結果を解釈すべきだと多くの研究者が指摘してきた（e.g., 芝・南風原，1990）。

1.2　帰無仮説検定と再現可能性

帰無仮説検定が有するさまざまな問題点が再現可能性を低めることを本研究では強く主張する。帰無仮説検定における二値的な判断も再現可能性を低める重大な要因である。帰無仮説検定では，

p値が慣習的な有意水準より下回れば有意な差（あるいは効果），それより大きければ有意でない，とるに足らない差（あるいは効果）という二値的な判断を行う。慣習的な有意水準として，心理学では.05が用いられる。つまり，p値が.05を下回れば，論文で報告できる結果となり，上回ればゴミ箱行きとなる。p値は確率なので，このような二値判断に馴染まない。.051と.049という2つのp値に本質的な違いがあると考える研究者はほとんどいないであろう。Rosnow and Rosenthal（1989）は，帰無仮説検定の極端な二値判断を皮肉り「神はp<.06をp<.05と等しく，そして同じくらい強く愛してくださる（p. 1277，拙訳）」と述べた。確率は連続的なものなので，明確な分水嶺を置く帰無仮説検定のアイデアは確率の評価にそぐわない。そして，この二値判断のせいで，p-hackingやHARKingといったQRPsが横行することになる（QRPsについては，この特集の池田・平石（2016），ならびに藤島・樋口（2016）を参照してほしい）。

標本サイズも再現可能性にとって重大な問題である。有意水準（p値の評価基準），標本サイズ，効果量，検定力は互いに影響する。特に標本サイズによってp値は大きく変化する。例えば，r=.10という弱い正の相関は，標本サイズが100ならば5%水準で有意にならない。しかし，標本サイズが400になると有意水準を下回ってしまう。つまり，帰無仮説検定の基準に従えば，意味のある効果と判断される。これは同じ効果量が得られたとしても，標本サイズによって結果の解釈が異なることを示している。

標本サイズがp値に与える効果をRosenthal and Rosnow（1984）は，検定統計量との関係から，以下の式（1）により端的に表した。

$$\text{検定統計量} = \text{効果量} \times \text{標本サイズ} \qquad (1)$$

効果量が同じなら，標本サイズによって検定統計量が決まる。検定統計量によって検定結果は決まるので，結局のところ標本サイズによって検定結果が決まることになる。このために標本サイズが大きな研究では有意と判断された結果が，標本サイズの小さな研究では有意と判断できないことが生ずる。同一の手続きを用い，それほど違いのない効果量が再現されたとしても，検定結果が食い違う場合がある。この場合，帰無仮説検定の基準で考えるなら，再現可能性はないことになる。

さらに検定の繰り返しの問題がある。検定を繰り返してはならないことを研究者なら誰でも知っている。しかし，実際の研究場面でそれを軽視することが報告されている。John, Lowenstein, and Prelec（2012）は，およそ6000名の心理学者を対象に調査を行い，2155名から有効回答を得た（この論文について藤島・樋口（2016）に詳細な議論がある）。そして，およそ70%の研究者が分析の結果を見ながら（中途解析における帰無仮説検定の結果を見ながら），データの取得を停止したことがあることを明らかにした。先に述べたように標本サイズは検定統計量に直接影響を与えるので，このような後付けの理由によるデータ取得の停止は望ましくない。データ取得の半ばで帰無仮説検定を行うことは検定を繰り返すことである。検定を繰り返せば，実質的な有意水準は上昇するので有意な差や効果を得やすくなる。これを意識していない研究者が多い。このような研究者のナイーブな感覚に警鐘を与えたのがSimmons, Nelson, and Simonsohn（2011）の研究である。彼らはコンピュータ・シミュレーションを行い，全く差がないデータセットを用い，データの取得に伴う検定を繰り返しの効果について検討した。その結果，本来差がないはずのデータセットでもデータの取得ごとに（例，1人の参加者のデータが得られたごとに）検定を行うと，22%のケースで有意な差が検出された（ただし，Murayama, Pekrun, & Fiedler, 2014も参照[1]）。また，Cramer et al.（in press）は，分散分析を用いた解析において，探索的な分析を行うことで実質的な有意水準が上昇することを警告した。例えば，2要因の分散分析には，2つの主効果と1つの交互作用がある。これらについて探索的に検定を行った場合，第1種の誤りは上昇し，14%にまでなる。つまり，有意水準5%で帰無仮説検定を行っているつもりでも，実質的には14%水準の検定となっているのである。Cramer et al.（in press）は*Journal of Experi-*

1) Murayama et al.（2014）は，有意となる見込みがありそうなとき（つまり，.05<p<.10）のみデータ取得を継続するという条件をつけると，有意な差が検出される確率（偽陽性の確率）が7.1%にまで下がることを示した。

mental Psychology: General, Psychological Science, Journal of Abnormal Psychology などを含む 6 つの論文誌を対象に調査を行った。そして，このような他要因計画の分散分析が分野を問わず多用されていることを報告した。6 つの論文誌全体で 47.62%，*Journal of Experimental Psychology: General* では実に 84.61% の論文が多要因の分散分析を用いていた。しかし，有意水準の修正を行っている研究はほとんどなく，全体では 1.03%，*Journal of Experimental Psychology: General* では 0% であった。無自覚に検定を繰り返すことで実質的な有意水準は上昇する。これが第 1 種の過誤を高め，再現可能性を低めることになる。

p 値だけに依存した判断も再現可能性を低める。帰無仮説検定に依存し *p* 値を見るだけでは，適切に現象を捉えられない可能性がある。例えば，論文で報告された *p* 値と効果量がしばしば一致しないことが示されている。効果量は効果の大きさを表す指標であり，*p* 値と異なり帰無仮説が正しくない程度を量的に表すものである（Cohen, 1988）。一方，*p* 値は，繰り返し述べてきた通り，帰無仮説が正しいという前提のもと，観察される差（あるいは効果）が得られる確率である。つまり，*p* 値は差や効果の大きさについて一切の情報を含まない。しかも，標本サイズを大きくすれば，式 (1) に示した通り検定統計量が大きくなるため *p* 値は小さくなる。結果として，わずかな効果量にもかかわらず，*p* 値から判断する限り有意な差や効果と判断されることが起こる。このようなことが原因となり *p* 値と効果量の不一致が生じると考えられる。Wetzels et al.（2011）は，*Psychonomic Bulletin & Review* と *Journal of Experimental Psychology: Learning, Memory & Cognition* の 2 誌に掲載された研究について対応のない *t* 検定を対象に検討を行った。その結果，全体として，*p* 値と *t* 検定の効果量である Cohen の *d* には負の相関があり，*p* 値が小さいほど *d* が大きくなった。つまり，全体として *p* 値と効果量は一致した解釈を導くことができた。ただし，*p* 値と効果量の不一致も観察された。例えば，*d*=.05 のとき *p* 値は .001 から .05 までばらつくことが示された。また，*p* 値が .01 のとき *d* は .02 から 1.0 まで，すなわち Cohen の基準では小さな効果量から大きな効果量の範囲にまでばらついた。日本の研究においても，*p* 値と効果量が一致しない場合が多くあることを波多野・吉田・岡田（2015）が明らかにした。これらの研究結果は，*p* 値のみに依存した解釈をした場合，効果量を無視した解釈となることを示している。同一の現象から得られる解釈に齟齬があるなら，現象は適切に解釈されていないと考えるべきだろう。*p* 値のみに依存した解析は，現象の解釈を歪めてしまうかもしれない。歪んだ解釈は追試にあたっては修正されると考えられるので，*p* 値のみに依存した解釈は再現可能性を低めると考えられる。

帰無仮説検定，特に *p* 値に対する依存によって，研究論文における *p* 値の報告が歪んだものになることも指摘されている。Masicampo and Lalande（2012）は，論文で報告された *p* 値の分布を調べた。彼らは *Journal of Experimental Psychology: General, Journal of Personality and Social Psychology*, そして *Psychological Science* の 3 誌で 2007 年から 2008 年の間に報告された *p* 値を集計した。*p* 値を横軸に取ったとき，低い *p* 値の報告が多くなるという指数減数型の分布が得られた。ただし，有意水準（.05）よりわずかに小さい *p* 値について報告数がスパイク状に増加し，この分布曲線から外れることが示された。この増加は，*p* 値に過度に依存した心理学の統計解析の結果であり，出版への圧力や QRPs が生み出している可能性が指摘された。Leggett et al.（2013）は，*Journal of Personality and Social Psychology* と *Journal of Experimental Psychology: General* について 1965 年と 2005 年に報告された *p* 値を集計した。どちらの年度でも，有意水準のすぐ下の *p* 値について報告数がスパイク状に増加した。ただし，その増加は 2005 年の方が大きかった。また，2005 年の上昇は *Journal of Personality and Social Psychology* で顕著であった。Leggett et al.（2013）は，近年強くなってきた出版への圧力によって，スパイク状の増加が生じていると解釈した。おそらく 2005 年になり，出版への圧力のため，*p* 値に過度に依存した解析の負の側面がさらに強調され，並行して QRPs も多くなったと考えられる。これらの結果は，現状として，心理学の研究が帰無仮説検定における *p* 値に過度に依存していること，そしてその結果として，再現可能性が低い結果が数多く報告される可能性があることを示唆している。

2. 日本における状況：社会心理学研究を対象として

これまで海外における研究結果を中心に検討を進めてきた。日本における心理学の研究においても帰無仮説検定への過度の依存があり再現可能性が低いデータが報告されているかもしれない。そこで，(1) *p* 値と効果量の不一致（波多野・吉田・岡田，2015；Wetzels et al., 2011）ならびに（2）有意水準付近における *p* 値報告数のスパイク状の増加（Masicampo & Lalande, 2012；Leggett et al., 2013）が，日本における心理学の研究で観察されるか検討を行った。*p* 値と効果量の不一致については，すでに波多野・吉田・岡田（2015）が報告をしており，その確認が主な目的となる。一方，有意水準付近における *p* 値報告数の増加はこれまで日本の研究を対象に検討されたことはない。さらに，これらが同じデータセットから報告されたことはなく，本研究が最初の報告となる。

本研究では日本社会心理学会が発行する論文誌である社会心理学研究の 22 巻 1 号（2006 年 8 月）から 29 巻 3 号（2014 年 3 月）に掲載された 170 篇の論文を対象とした。対象となった多くの論文において効果量の記載がなかった。そこで論文に記載された情報から簡便に効果量を計算できる対応のない *t* 検定を検討対象とした。対応のない *t* 検定を行った論文は 170 篇のうち 32 篇であった。標本サイズの決定にあたり，検定力に基づく例数設計を行った。検定力 .80 で中程度の効果量を想定すると，81 が適切な標本サイズである。論文中の記述不足や記述間違いによる除外が予想されたため，その数を超える 101 の検定結果を分析の対象とした。ただし，今回の分析にあたって，除外対象となった検定はなかった。

Figure 1 に対応のない *t* 検定における *p* 値とその効果量である Hedges の *g* の散布図を示した。*p* 値と *g* には弱い負の相関があり，*p* 値が小さくなるほど *g* が大きくなった，$r(101)=-.28$，$p=.005$。ただし，2 つの指標には不一致もあった。まず，$p=.05$ の周辺で *g* は 0.2 から 0.8 の値を取り，大きなばらつきがあった。さらに，中程度の効果量である $g=.05$ 周辺では，有意水準を下回ることも上回ることもあり，$p=.001$ から .2 のあいだに広がった。これらのパタンは先行研究と一致していた。すなわち，全体として負の相関があることから *p* 値と *g* は一致した解釈が可能なことが多いものの，有意水準周辺では解釈に齟齬が生ずることが示された。

さらに *p* 値の分布を検討したところ，全体として *p* 値が低いほど報告数が多く，$p<.01$ で急激に減少する指数減衰型の分布が得られた。先行研究の手続きに則り（Masicampo & Lalande, 2012；Leggett et al., 2013），$p>.01$ から $p<.10$ の範囲について，.01 刻みで *p* 値を集計し，指数減数関数の当てはめを行った。全体の傾向と同様に $p>.01$ から $p<.10$ の範囲についても *p* 値は指数減数型に分布しており，その関数を当てはめたところ $R^2=.61$ の説明率であった。ただし，Figure 2 に示した通り，.04 において *p* 値の報告数は増加しており，分布曲線から大きく外れていた。今回のデータセットでは標本サイズが先行研究よりも少ないため，スパイク状の増加と呼べるか判断が難しい。それでも，先行研究に類似したパタンが観察された。今後，大きな標本サイズでの再検討が求められる。

Figure 1 と 2 はそれぞれ（1）*p* 値と効果量の不一致（波多野・吉田・岡田，2015；Wetzels et al., 2011）と（2）有意水準付近における *p* 値報告数の増加（Masicampo & Lalande, 2012；Leggett et al., 2013）が日本における心理学の研究でも生じるこ

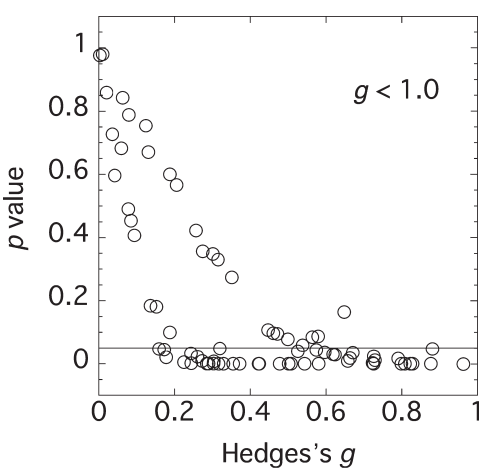

Figure 1 Scatterplot of *p* values and effect sizes (*g*) in independent *t* tests. Tests with $g<1.00$ were not presented. The horizontal line represents the conventional α level (.05).

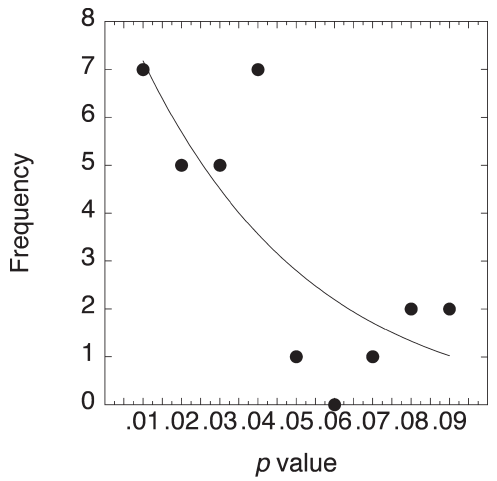

Figure 2　Frequency of *p* values for independent *t* tests. The line shows the best fitting exponential decay function. *p*s < .01 and > .10 were not presented.

とを示すものである。これらの結果はどちらも有意水準周辺で生じる現象であり，帰無仮説に過度に依存した現状が日本にもあることが明らかになった。しかも同一のデータセットにおいて，有意水準周辺において *p* 値と効果量の不一致があり，かつ，*p* 値の報告数が増加することが示された。これまでの研究ではこの 2 つの現象の関連は示されていなかったが，今回の分析においてそれらが同時に生じることが明らかになった。

3. 再現可能性を確保するために

著名な心理学論文誌に載った 100 件の実証的研究のうち，39% 足らずしか結果が再現されず，再現性可能性の低さに警鐘がならされた（Open Science Collaboration, 2015）。本論文では再現可能性の低さをもたらす大きな要因として，帰無仮説検定への過度な依存があることを指摘した。そして，その結果として，日本における心理学の研究においても，*p* 値と効果量に不一致があり，かつ，慣習的な有意水準である .05 よりわずかに低い *p* 値において報告数が増加することが示された。本論文で検討の対象としたのはごく限られた期間に過ぎず，論文誌も 1 誌しか対象としてない。しかし，このような結果は国際的な論文誌でもすでに報告されている（Masicampo & Lalande, 2012；Leggett et al., 2013；Wetzels et al., 2011）。おそらく心理学全体の問題であり，それが本邦でも同様に存在すると考えても良いであろう。

今回観察された *p* 値と効果量の不一致と有意水準周辺における *p* 値報告の増加は，帰無仮説検定への過度な依存が一因と考えられる。この過度の依存を廃し，心理学研究における再現可能性を確保しなくてはならない。本論文では，そのためのアプローチを大きく分けて 2 つ紹介する。一つは現在でも広く使用されている帰無仮説検定を継続的に用い，その難点を修正するものである。現在の圧倒的な使用率を考えると帰無仮説検定を直ちに廃止することはあまり現実的でないかもしれない。このアプローチはその現実的な要請に応えたものと言えるだろう。

なお，帰無仮説検定を修正するアプローチや信頼区間を重視するアプローチは母数が定数であると想定した解析である。この立場は，データは母集団からランダムに抽出された確率変数であると考えるため頻度主義と呼ばれる。このような考えと依って立つ前提が異なる立場もある。それがベイジアン・アプローチである。これはベイズ主義と呼ばれる立場に基づくもので，頻度主義とは異なり，母数は未知で不確実なため確率変数と考え，手元にあるデータを定数であると考える。ベイズ主義に基づくと，帰無仮説検定の難点であるその論理的な構造，さらに標本サイズや検定の繰り返しの問題を回避することができる。ただし，あまり広く使用されておらず，難易度が高い。本論文では頻度主義とベイズ主義に基づくアプローチを紹介する。

3.1　頻度主義に基づくアプローチ

帰無仮説検定の大きな枠組みを維持しつつ，過度な依存から脱する最も簡便な方法は，*p* 値のみを統計的判断の材料とせず，複数の指標から統合的に評価し判断することである。これは American Psychological Association（以下，APA）が発行する最新の Publication Manual（第 6 版）でも明記されている（American Psychological Association, 2009）。APA Publication Manual によれば，“仮説検定は分析のはじまりにすぎない。結果の意味を十分に伝えるためには，効果量，信頼区間，そして詳細な説明が不可欠であると APA はここに強く主張する（APA, 2009, p. 33，拙訳）”と述べられ，帰無仮

説検定の p 値と共に効果量と信頼区間を記述することは最低限求められること（minimum expectations, APA, 2009, p. 33）となった。2015 年に改定された日本心理学会投稿・執筆の手引きでも，p 値を記載する場合，効果量を付記することが求められるようになった（日本心理学会，2015）。複数の指標を統合的に評価することで p 値のみに依存した判断を避けることができる。効果量や信頼区間については大久保・岡田（2012）など日本語でも詳しく解説した書籍が出ているので参考にしてほしい。

上述した p 値と効果量の不一致からも明らかなように（波多野・吉田・岡田，2015；Wetzels et al., 2011），これまでの心理学の研究において複数の指標に基づく統合的な判断はあまりなされてこなかった。実際，p 値のみを重視し，効果量を無視した解釈がほとんどであった。本論文で報告した社会心理学研究を対象とした分析結果も，統合的な判断が十分にできていないことを裏付けるものであろう。もちろん論文投稿の規則が変化し，p 値だけでなく効果量や信頼区間の報告が義務付けられたことで，論文の結果の記載も変化し，複数の指標が記載されるようになってきた（Cumming et al., 2007；Fritz, Morris, & Richler, 2012）。ただし，日本ではまだまだ不十分である（大久保，2009）。複数の指標を掲載することが浸透すれば，それらを用いた統合的な判断を研究者が自然に行うようになることが期待される。しかし，効果量や信頼区間を報告しても，統合的な判断はなされず，p 値のみに依存した解釈が行われがちなことが医学や疫学の論文を対象とした研究から示されている（Fidler et al., 2004）。論文投稿の規則を変えるだけでなく，時間をかけた教育や啓蒙が併せて必要なのであろう。

次に再現可能性を高めるための具体的な手段として例数設計について紹介しておこう。帰無仮説検定への過度の依存により再現可能性が低下する要因として，（1）標本サイズが p 値に与える影響と（2）検定の繰り返しについて指摘した。また，（3）p 値と効果量の不一致は再現可能性が低いことを示す実証例である。適切な例数設計を行うと，これら 3 つを避けることができる。例数設計とは，信頼できるデータを得るために必要な標本サイズを推定することである。心理学研究では，検定力に基づく例数設計を行うことが多い。検定力とは 1−β，つまり，全体の確率から第 2 種の過誤を引いたものである。全体の確率から，第 2 種の過誤（差があるにもかかわらず，差がないと判断する確率）が除かれるので，差があるときに差があると判断できる確率が検定力となる。この定義から考えて，例数設計のため検定力を用いることが妥当性であることがわかるだろう。例数設計にあたり，Cohen（1988）が提案した .80 を検定力の基準として採用することが多い。もっともこれはあくまでも目安に過ぎず，固定的なものではない。第 2 種の過誤が甚大な影響をもたらす場合（例，環境問題，人命にかかわる問題），検定力を .95 上げることが推奨されている。

例数設計を行うことで上にあげた 3 つの問題を一気に片付けることができる。（1）例数設計を適切に行えば，標本サイズが大きすぎたり逆に小さすぎたりすることがない。適切な標本サイズで研究が行われているなら，p 値に対する標本サイズの影響をさほど気にしないで良い（他の研究との比較は必要となる）。また，（2）例数設計により標本サイズが決まっているのでデータの取り足しが生じず，そのために検定を繰り返すこともない。そして，（3）例数設計において，効果量と有意水準の関連がすでに考慮されているので，解釈の不一致が生ずる可能性は小さい。このように例数設計によりさまざまな問題を一気に解決できる。検定力に基づく例数設計は G*Power などフリーソフトを使って簡単に行えるので試してみると良い（Faul et al., 2007）。

例数設計には幾つか種類がある。検定力を用いる方法だけでなく，信頼区間を使用した正確度分析や適応的な停止規則を採用するものもある。正確度分析による例数設計とは，事前に設定した信頼区間におさまるよう標本サイズを決定する手続きである。一方，適応的な停止規則とは事前に決定した規則に基づいて，その基準に達した段階でデータ取得を停止する方法である。この場合，明確な標本サイズを事前に決定しない。これは動物研究など標本サイズを大きくすることが難しい分野で採用されることが多い（e.g., Frick, 1998；Fitts, 2010）。

APA Publication Manual は，例数設計について記述することを明示的に求めている（APA, 2009,

p. 30）。Psychonomic Societyの統計ガイドラインでも検定力を考慮し，どのように標本サイズを決定したか記述することが必須とされた。このように心理学の様々な論文誌で，例数設計に関する記述が明示的に求められており，例数設計を行わず研究を行うことは不適切だと認識されるようになってきた。これまでの心理学の研究では中程度の効果量を想定した場合，.39–.66程度の検定力しかない場合が多く，Cohen（1988）が提唱した.80という水準にはるかに及ばないことが指摘されてきた（Aberson, 2010）。日本国内の研究を対象として論文でもおおよそこの程度だと報告されている（杉澤，1999；鈴川・豊田，2012）。心理学では，p値に対する過度の依存があり，検定力や例数設計について十分に考慮をした研究が少なかった。今後，例数設計が定着することで，心理学研究における検定力が全体的に増加し，延いては再現可能性を高めることにつながることが期待される。

なお，頻度主義のアプローチを取り，かつ，帰無仮説検定を廃することも可能である。例えば，医学や疫学では帰無仮説検定を禁止し，信頼区間を重視するアプローチを採用した論文誌もある（Fidler et al., 2004）。心理学でも *Basic and Applied Social Psychology* が帰無仮説検定を廃する編集方針を発表した（Trafimow & Marks, 2015）。信頼区間を重視するアプローチについては紙幅の制限もありここでは詳しく触れない。Altman et al.（2013）など丁寧な解説があるので参考にしてほしい。

3.2　ベイジアン・アプローチ

ベイジアン・アプローチを採用することで，帰無仮説検定の論理的な構造に起因する問題を解決できる。頻度主義の立場では，帰無仮説検定において帰無仮説が正しいという前提のもと，データが得られる確率を計算する。つまり，$P(D|H)$を求めることになる。ここにまず大きな問題がある。帰無仮説検定は，先に述べたようにPopperの方法論的反証主義のアイデアと一致するものである。そのため，特殊な前提をおき確率の計算を行う。特殊な前提とは，帰無仮説が真という前提である。この前提をおくと後件否定の論証形式を用いることができ，強力な推論が可能になる。しかし，この前提には問題がある。現実の測定において，比較条件間で全く差がないことは，ほぼありえないからである。現実的にはほぼありえない前提に基づく確率$P(D|H)$が，どの程度有益な情報を与えるか疑問が残る。帰無仮説検定では論理的な整合性を重視するあまり，研究者が本来求める情報が得られなくなっている。

ベイジアン・アプローチでは，研究者が本来知りたい情報を確率として得ることができる。ベイズ主義では母数をデータと考えるので，母数に関する確率を求めることができる。帰無仮説検定を含む頻度主義では，逆に母数を定数と考えるため確率的に扱うことができない。母数に関する確率を求められることは頻度主義にはないベイズ主義の利点である。科学研究における仮説は多くの場合母数に関する仮説である。従って，ベイズ主義に立てば，そのデータが得られたことを前提に，本来知りたい仮説そのものが正しい確率確率$P(H|D)$を得ることができる。すなわち，研究者が知りたい仮説の正しさを確率として直接推定できるのである。

かつてベイジアン・アプローチに基づく確率の推定は，主に技術的な理由で困難であった。計算に非常に手間がかかり，現実的に意味のある事態について確率を計算することができなかった。また，主観的に事前確率を設定することへの激しい抵抗もあった。しかし，コンピュータの爆発的な進歩により技術的な問題が解決し，事前確率の設定についても客観的な基準が少しずつ整備されるようになり，状況は大きく変わってきた。

帰無仮説検定で問題となる標本サイズが有意性に与える効果も，ベイジアン・アプローチを取るとあまり問題にならない。ベイジアン・アプローチでは，得られたデータを分析に足し込み，確率を更新することができる。一般に標本サイズが増えるほど推定は正確になるので，標本サイズが増えるほど，$P(H|D)$，すなわち，仮説の正しさを正確に評価できる。帰無仮説検定で，標本サイズが問題になるのはp値に基づき帰無仮説の棄却・採択を行うためである。比較条件間で全く差がないことはほぼありえないため，標本サイズが増えていけばいずれ帰無仮説は棄却されてしまう。結果として，現実的には，ほぼ無視して良いような差や効果が意味のある差と判断されることになる。

ただし，これはあくまでも帰無仮説を棄却し，暫定的に対立仮説を採択するという複雑な論理構造が帰無仮説検定にあるために生ずる。研究者が知りたい仮説の正しさを直接計算できるベイジアン・アプローチでこの問題は生じない。

Johnson（2013）は，帰無仮説検定における p 値に依存した判断ではなく，ベイズファクターを用いた判断に変更すべきだと主張した。ベイズファクターとは，データによって得られた2つの仮説の妥当性をオッズ比として表したものである。Johnsonは，帰無仮説検定が正当化できないほど高い有意水準を用いているために第1種の誤りが生じ，結果として再現可能性が低くなることを指摘した。彼の分析によると，帰無仮説検定の枠組みならば，有意水準を .005 あるいは .001 に設定するくらいの厳しさがなければ，十分な再現可能性が担保されないという。

もっともベイジアン・アプローチに関する批判も多い。確率を推定するためには，事前確率を設定しなくてはならず，その設定に十分な客観性を与えることは難しい。また，比較的新しく開発された手法が多いため，評価が定まっていないことも多い。例えば，ベイズファクターについてさまざまな議論がある。ベイズファクターは概念的には理解が簡単なものの実際の計算は複雑で簡単に利用ができない（詳しくはPericchi, 2005）。また，ベイズファクターを用いる場合，帰無仮説検定に比べ基準が圧倒的に厳しくなる（Johnson, 2013）。効果量にもよるが，一般に，基準が厳しくなると安定した推定を得るために大きな標本サイズが必要である。臨床的テストや侵襲性がある実験の場合，標本サイズを大きくすることには倫理的な問題が伴う。単純に増やせばよいというものではない。この点から，基準が厳しくなるベイズファクターによる評価を批判する研究者もいる（Gaudart et al., 2014）。

ベイジアン・アプローチは，帰無仮説検定が主流となっている心理学において現時点ではあまり用いられていない。帰無仮説検定とは依って立つ立場が大きく異なるため知識のない研究者は手を出しづらい。解析に使用できるソフトウエアも現時点では限られており，使いこなすのにも知識が必要である。また，ベイジアン・アプローチを紹介した書籍で，心理学の研究にそのまま用いることができる詳細かつ具体的なものはまだまだ少ない。ただし，近年になり少しずつ出版されるようになった。日本語で読めるものとして，豊田（2015）やLesaffre and Lawson（2012）の翻訳がある。豊田（2015）は，歴史的背景や，ベイズの定理，確率分布に関する解説も丁寧に行っており，統計の基礎知識があれば無理なく理解できる。また，岡田（2014）は，分散分析に変わるものとしてベイジアン・アプローチによる情報仮説の評価について実例に基づく紹介を行った。英語の書籍なら，Lee and Wagenmakers（2014）が大変参考になる。基礎的な解析と併せて，著者らが行った研究に対応させた分析を紹介しており具体的でわかりやすい。歴史的な背景や哲学的な議論にも関心がある場合は，Dienes（2008）に読みやすくかつ詳細な情報がある。興味がある方はこれらの書籍や論文を手に取ってみると良いであろう。

3.3 むすび

帰無仮説検定に関する批判は長らく続いており，その批判を受け改革が心理学では進んできた（Cumming et al., 2007；Kline, 2005；大久保・岡田，2012）。しかし，帰無仮説検定への過度な依存は現時点でも強い。本研究における分析でもその過度な依存がデータをもって示された。これが再現可能性を低める一因となっていると考えられる。上述のアプローチを採用するなどして，心理学における統計改革をさらに進め再現可能性を高めていくことが求められる。

謝　辞

本研究は，平成23–27年度文部科学省私立大学戦略的研究基盤形成支援事業「融合的心理科学の創成：心の連続性を探る」（S1101013）の助成を受けた。

文　　献

Aberson, C. L. (2010). *Applied power analysis for the behavioral sciences*. New York: Routledge Academic.

Altman, D., Machin, D., Bryant, T., & Gardner, M. (2013). *Statistics with confidence: confidence intervals and statistical guidelines*. New York: John Wiley & Sons.

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: problems, prevalence, and an alternative. *Jourrnal of Wildlife Management*, *64*, 912–923.

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

Cohen, J. (1988). *Statistical power analysis for the behavior sciences (2nd ed.)*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist*, *49*, 997–1003.

Cramer, A. O. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P. P. P., Waldorp, L. J., & Wagenmakers, E. J. (in press). Hidden multiplicity in multiway ANOVA: Prevalence, consequences, and remedies. *Psychonomic Bulletin & Review*.

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., Lo, J., McMenamin, N., & Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science*, *18*, 230–232.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. London: Palgrave Macmillan.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science*, *15*, 119–126.

Fisher, R. A. (1959). *Statistical Methods and Scientific Inference (2nd ed.)*. Edinburgh: Oliver and Boyd.

Fitts, D. A. (2010). The variable-criteria sequential stopping rule: generality to unequal sample sizes, unequal variances, or to large ANOVAs. *Behavior Research Methods*, *42*, 918–929.

Frick, R. W. (1998). A better stopping rule for conventional statistical tests. *Behavior Research Methods, Instruments, & Computers*, *30*, 690–697.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*, 2–18.

藤島喜嗣・樋口匡貴（2016）社会心理学における "*p*-hacking" の実践例　心理学評論，*59*, 84–97.

Gaudart, J., Huiart, L., Milligan, P. J., Thiebaut, R., & Giorgi, R. (2014). Reproducibility issues in science, is P value really the only answer? *Proceedings of National Academy of Science USA*, *111*, E1934.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum.

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.

Hubbard, R., & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology — And its future prospects. *Educational and Psychological Measurement*, *60*, 661–681.

池田功毅・平石 界（2016）心理学における再現可能危機：問題の構造と解決策　心理学評論，*59*, 3–14.

John, L. K., Lowenstein, G., & Prelec, D. (2012) Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of National Academy of Science USA*, *110*, 19313–19317.

Kline, R. B. (2005). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington DC; American Psychological Association.

Krueger, J. (2001). Null hypothesis significance testing. On the survival of a flawed method. *American Psychologist*, *56*, 16–26.

Lakatos, I. (1978). The methodology of scientific research programmes. In J. Worrall & G. Currie (Eds.) *Philosophical Papers Volume 1*. Cambridge: Cambridge University Press.

Leggett, N. C., Thomas, N. A., Loetscher, T., & Nicholls, M. E. (2013). The life of *p*: "Just significant" results are on the rise. *The Quarterly Journal of Experimental Psychology*, *66*, 2303–2309.

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

Lesaffre, E., & Lawson, A. B. (2012). *Bayesian Biostatistics*. New York: John Wiley & Sons. 宮岡悦良（監訳）(2016) 医薬データ解析のためのベイズ統計学　共立出版.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*, 161–171.

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below. 05. *The Quarterly Journal of Experimental Psychology*, *65*, 2271–2279.

日本心理学会（2015）日本心理学会　執筆・投稿の手びき（2015 年改訂版）　日本心理学会.

大久保街亜（2009）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.

岡田謙介（2014）ベイズ統計による情報仮説の評価は分散分析にとって代わるのか？　基礎心理学研究，*32*, 223–231.

Omi, Y., & Komata, S. (2005). The evolution of data analyses in Japanese psychology. *Japanese Psychological Research*, *47*, 137–143.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716.

Pericchi, L. R. (2005). Model selection and hypothesis testing based on objective probabilities and Bayes factors. In D. Dey, & C. R. Rao (Eds.) *Bayesian thinking: modeling and computation. Handbook of statistics vol. 25* (pp. 115–149). Amsterdam: Elsevier/North-Holland.

Popper, K. R. (1959). The logic of scientific discovery. Hutchinson, London, United Kingdom.

Rao, K. R. (1985). Replication in conventional and controversial sciences. In B. Shapin & L. Colby (Eds). *The Repeatability Problem in Parapsychology* (pp. 22–41). New York: Parapsychology Foundation.

Rosenthal, R., & Rosnow, R. L. (1984). Applying Hamlet's question to the ethical conduct of research: A conceptual addendum. *American Psychologist*, *39*, 561–563.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276–1284.

芝 祐順・南風原朝和（1990）行動科学における統計解析法　東京大学出版会.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

杉澤武俊（1999）教育心理学研究における統計的検定の検定力　教育心理学研究，*47*, 150–159.

鈴川由美・豊田秀樹（2012）“心理学研究”における効果量・検定力・必要標本数の展望的事例分析　心理学研究，*83*, 51–63.

豊田秀樹（2015）基礎からのベイズ統計学：ハミルトニアンモンテカルロ法による実践的入門　朝倉書店.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*, 1–2.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science*, *6*, 291–298.

― 2016. 3. 1　受稿，2016. 3. 20　受理 ―