特別寄稿

神経心理学と統計
統計的仮説検定における効果量と検出力の問題

板口　典弘*

要旨：統計学的手法は多くの学問分野で，研究の主張における信頼性の一部を担保するために用いられる。しかしながら，統計学的手法の誤用や濫用の問題も少なからず指摘されている。神経心理学において，臨床検査を誤った方法で運用してはいけないように，統計手法もルールに則った適切な運用が求められる。そこで本稿では，統計的仮説検定の基礎およびその使用において考慮すべき検定の多重性，効果量，検出力について解説した上で，神経心理学研究における統計手法の運用状況を概観する。さらに，それらにかかわる問題の実用的な対策についていくつか提案する。

Key Words：検出力，効果量，検定の多重性，サンプル数，臨床研究

はじめに

統計学的手法とは，研究の主張における信頼性の一部を担保するひとつの道具である。ただし，使用する以上は，“たかが統計”であったとしても適切な運用が求められる。神経心理学領域の国際誌においては，様々な統計手法の問題が昔から繰り返し指摘，議論されている（Bezeau ら , 2001；Blakesley ら , 2009；Eichstaedt ら , 2013；Millis, 2003；Schatz ら , 2005）。これは，ただ使用者が勉強不足であるというだけでなく，神経心理学研究は小サンプルに基づいた考察や多数の検査結果の比較検討を含む点で，そもそも“統計的仮説検定”の適用が難しい性質を持っていることも一因であると考えられている（Beeson ら , 2006；Lezak ら , 1984；Loring ら , 2014；Silverstein, 1986）。さらに近年では，統計学的手法，特に統計的仮説検定の誤用・濫用が様々な学問分野で問題となっており（たとえば *p*-hacking, HARKing など[1]），わが国でも，心理学や医学･薬学･健康科学領域における学術雑誌において特集が組まれ，統計の専門家のみならず，ユーザーの立場からの議論も活発に行われている（松井, 2018；柳川, 2018）。一方で, わが国の神経心理学領域においては，統計学的手法の運用状況を直視し改善を促すような機運は高まってはいない。

そこで本稿ではまず，統計的仮説検定の運用において考慮するべき事柄である, ①検定の多重性問題, ②効果量, ③検出力の3点について解説する。次に,「神経心理学」誌に掲載された原著論文における統計手法の運用状況を簡単に報告する。加えて，これからの研究のためにいくつかの理論値を紹介しつつ, 上記問題への対策を紹介する。ただし本稿では，統計的仮説検定において考慮すべきすべての問題を扱っているわけではなく，解説や対応策の提案も概略的である点に留意していただきたい[2]。

1. 統計的仮説検定の基礎

様々な基準を用いて統計手法を分類することができるが，私たちが今まで使用してきた“古典的”統

1）*p*-hacking は統計的に有意な結果を“ひねり出す”方法の総称である。藤島ら（2016）では *p*-hacking を“実践”した結果を報告しており，具体的にどのような行為が *p*-hacking に該当するのかを理解するために有用である。HARKing（Hypothesizing After the Results are Known）とは，データに合わせて仮説を生成して（変えて）しまう方法である。

2）わかりやすさのために，用語も多少厳密でないものを用いた。

【受稿日　2019年5月7日】
Neuropsychology and statistical hypothesis testing：effect size and statistical power
＊ 静岡大学情報学部　Yoshihiro Itaguchi：Department of Computer Science, Shizuoka University

計手法においてわかりやすい分類は，「記述統計か推測統計か」というものであろう。記述統計とは手元にあるデータの性質を要約するものである。推測統計とはその背後に仮定される「母集団」（手元のデータが抽出された全体）の特徴を推測するものである。p値や統計的仮説検定は，推測統計の一部である[3)]。

a. 統計手法における自由と不自由

数多くある統計手法は，研究の主張をサポートするための道具である。ここで重要な点は，取得したデータは中立な“根拠”であること，および，統計手法は様々な理論（“論拠”）のうちのひとつであるということである（**図1**）。論拠とは，根拠に解釈を与え，主張を引き出すためのルールであると捉えるとわかりやすい。当然ながら，ひとつの論拠ですべての事象をカバーすることはできない。そのため，たとえ手元にデータが揃ったとしても，使用するべき統計手法は自動的には決定されず，データの性質や主張したい内容に応じて適切に選択する必要がある。言い換えれば，統計手法を選ぶ自由（と責任）は常に使用者自身にある。

その一方で，データと統計手法が決定されれば，主張は一意に決定される。すなわち，ある統計手法を適用した後の結果の解釈には自由がない。たとえば，研究のスタンスによって，統計的仮説検定におけるp値の解釈の仕方が変わることはない。どのような研究であっても，ある特定の統計手法を用いた限りは，その結果の解釈には特定のルール（仮定や限界）が適用される。このルールを無視することは，学術論文においては暴力に等しい行為である（濫用＝abuseとも呼ばれる）。特に統計的仮説検定においては，使用者の意図にかかわらず“仮説”が設定されている点に注意が必要である。

b. 仮説検証の手続き

統計的仮説検定における仮説の検証は以下のような手順を踏む。まず，「今回得られたデータが単一の母集団から生じる確率」を計算する。これがp値である。その確率が非常に低い（5%以下）場合に，「データが単一の母集団から生じた」という仮説が間違っていたと考え（帰無仮説の棄却），「データが複数の母集団から生じた」と結論する（対立仮説の採択）。このとき，最初から「仮説が間違っていた」という結論を求めている点で，統計的仮説検定は背理法である。t検定の例で説明すると，使用者は「2群の母集団の平均値が異なる≒2群のデータは複数の母集団から生じた」という結論を目指して「2群のデータが単一の母集団から生じた」確率を計算し，それが非常に低いことを示す。この論理はすべての統計的仮説検定で同一である。ちなみに，分散分析においてn群の比較を行った場合には，「n群のデータは複数（≠n）の母集団から生じた」という対立仮説を採択する点に注意が必要である。

c. p値とデータ数

統計的仮説検定の有意性判断ではp値が5%を下回るかどうかを基準とする。この5%という値そのものは恣意的な値である。しかし，恣意的な値であるから軽視しても良いという論理は通らない。5%という値よりも大事なことは，p値の正しい解釈を行うことである。このとき重要な要素になるのがデータ数である。なぜなら，p値の大小には，効果の大きさに加えて，データ数が大きくかかわってく

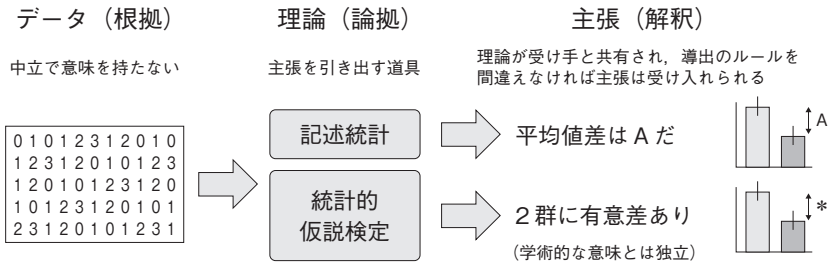

図1 データと統計手法の関係

3）私たちが一般的に使用している統計的仮説検定はNeyman–Pearson流の解釈を用いている。一方で，仮説検定（対立仮説の採択）を行わず，p値のみを解釈するという立場もある（Fisher流の検定）。この2つの立場の比較は柳川（2018）がわかりやすい。

るためである。具体的には，データ数が多いと小さな効果であってもp値は小さくなってしまう（結果，有意だと判断される）一方で，データ数が少ないと大きな効果があってもp値は小さくならないという構造がある。したがって，有意性判断の解釈には，必ずデータ数を考慮に入れなければならず[4]，データ数が適切でない場合の解釈には慎重を要する。p値にかかわる重要な要素（効果とデータ数）を独立させた指標が，次に紹介する効果量と検出力である。

d. 効果量と検出力

まず効果量とは，データ数に依存しない，「データのばらつき（ノイズ）に対する効果の大きさ」の指標である。たとえばt検定の場合はdという指標が提案されている[5]。Cohen（1988）は，効果量dの目安を0.2，0.5，0.8（小，中，大）としている。d＝1であるとき，2群間には標準偏差ひとつ分の平均値差があると考えてよい。ちなみに相関係数rも効果量の一種である。次に検出力とは，「母集団が単一でない場合に，それを正しく検出する力」を意味する指標をいう。Cohen（1988）が推奨する検出力は0.8以上である。検出力が0.8であるとは，ある統計デザインにおいて「帰無仮説が誤っている場合に，80％の確率で帰無仮説を棄却できるデータ数である」ことを意味する。データ数が多くなれば検出力も大きくなるため，実用上の概念としては，データ数の多さを統計デザイン間で比較できるように標準化したものと捉えて問題ない（**図2**）。

実験を計画する段階で，想定される効果量・検出力・有意水準の3つを定めることにより，その効果量を検出するために必要なデータ数（すなわち実験参加者数）を計算することができる（事前の分析，鈴川ら，2012）。検出力と有意水準は慣習的な値（0.8と0.05）が存在する。そのため，検出したい効果量と統計デザインが決まりさえすれば，必要な参加者数が決まる。検出すべき効果量は研究や状況によって異なるため，一概に決定することはできない。しかしながら，先行研究等やCohen（1988）の基準などを参考に，学術的に意味がある値を暫定的に設定することは可能である。このように効果量と検出力に対して実験計画段階から注意を払うことにより，p値と有意性判断における不毛な議論（5％という基準に意味があるのかどうか，など）を防ぐことができる。

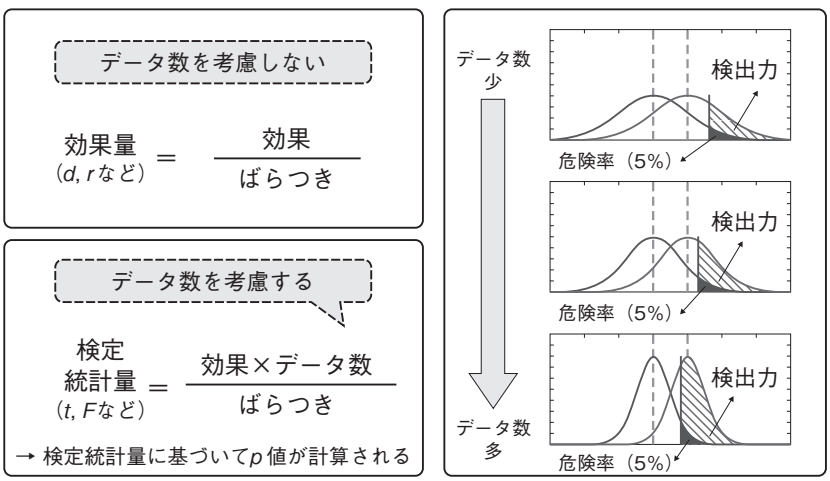

図2　効果量，検定統計量，検出力
データ数が多いほど母集団平均値への推定精度が高くなるため，分布の形が急峻になる（右図）。その結果，同じだけの平均値差を想定した場合に検出力が高くなる。

4）「t（20）＝10.0」などと統計結果に自由度（データ数に比例）も記載しなければならないのはこのためである。
5）検定の種類に応じて様々な種類の効果量が提案されている。そのため，分散分析やカイ二乗検定についても効果量を算出することができる。効果量の概念や算出方法については水本ら（2008）が参考になる。

e. 統計的仮説検定運用における問題

統計的仮説検定を使用する上で，注意を払うべき問題を2つに分けて検討する。1点目は，検定の多重性による危険率の上昇である。統計的仮説検定は通常，ひとつの検定につき5％という危険率（有意水準）を設定している。“5％水準で有意差がある”という検定結果は，厳密な表現ではないが，5％の確率で“本当は差がないのに，差があると誤って判断する”危険性があることに相当する。当てずっぽうの答えを続けていけばいつか正解するのと同じ論理で，複数回の検定を行うと全体としての危険率は検定回数に応じて上昇してしまう。このような検定の多重性問題は，多重比較という手法によって回避することができるものの，一般的にその対策がなされているのは，分散分析後に実施される検定の繰り返しにおいてのみである。しかしながら，どのような検定の繰り返しにおいても検定の多重性問題は生じうるものであり，かつ対処しなければならない大きな問題である（Blakesleyら, 2009；Frane, 2015；橘, 1986）。

2点目は，統計的仮説検定の検出力が高すぎる場合と低すぎる場合に生じる問題である。有意性判断の基準となるp値にかかわる要素は，①興味のある要因の効果（t検定であれば2群の平均値差），②興味のない要因の効果（群内のばらつき＝ノイズ），③データ数，の3つである。このうちデータ数の要素は，研究者がどれだけの対象者をリクルートしたかという数字に過ぎない。それにもかかわらず，データ数が多すぎる（検出力が高すぎる）と，学術的に意味のない効果であっても有意だと判断されてしまう。この問題を精査しないまま議論がなされた場合，もっとも被害を受けるのは統計学に明るくない大多数の読者である。さらに，見せかけの効果に基づいた議論は，学術分野のみならず社会的な問題にもつながりうる。一方で，データ数が少なすぎる（検出力が低すぎる）ために，学術的に意味のある効果が検出できない場合にも問題が生じる。つまり，データ数が足りないという理由だけで，学術的に考察されるべき現象が議論されない，あるいは報告対象ともならないケースである（お蔵入り）。これらのケースは，当該研究に要した資源（労力や資金）が無駄になるだけでなく，後の研究の方向性をミスリードする危険性がある。たとえば，同じような研究が繰り返され，どの研究においても有意な効果が得られないと報告された場合には，人々はそれを“効果がなかった”証拠として捉えてしまう。しかしながら，どの研究でも安定した効果量が得られていた場合には，実際には議論・報告すべき重要な知見かもしれない。このように，統計的仮説検定の判断結果に依拠しつつ，学術的に誠実かつ妥当な結論を導くためには，効果量と検出力の双方を考慮する必要がある。

2. “神経心理学”誌における検定運用の実際

本節では，「神経心理学」誌に掲載された170編の原著論文における統計手法の運用状況を報告する。紙幅の関係上，データの詳細については紹介しない[6]。また，検出力は実験を実施する以前に“検出すべきである”と想定された効果量との関係で検討しなければ，値そのものを吟味する意味は薄いため，本調査では算出していない。

多重性と効果量の問題の前にまず，検定の運用状況を簡潔にまとめる。それぞれの研究が対象とした参加者数は1名（一例報告）がもっとも多く（37％，63編），参加者10名以下までの研究が49％（84編）であった。一方で，100名を超える参加者を持つ研究も16.5％（30編）存在した。また，統計的仮説検定を1回以上実施している論文[7]は全体の51％（87編）であった。これ以降の検討では，それら87編の論文を検討対象とした。

a. 検定の多重性問題

分散分析後の多重比較を除き，1論文あたりの検定の繰り返し数を算出した。1論文あたりの総検定回数をみると，最大値は312回，中央値は13回であった。検定回数が2回以上の論文は81編であった。相関分析が実施される際には特に，質問項目や検査項目すべてに対して総当りの相関係数が計算されることが多かった。このような中，解析対象となった論文のうち，2編（2％）のみで全体の危険率が多重比較法によって調整されていた（それぞれ相関係数の

6）詳細は，板口・福澤（2019）を参照してほしい。
7）前提としてあるべき差や，差がないことを確認する検定はカウントから除外した。

有意性検定を8回と23回の繰り返し）[8]。国際誌における神経心理学研究を対象とした調査でも，本調査と同様に1論文内で多くの検定が繰り返されていることが明らかとなっている（最大値で139回，中央値で24回）。ただしその一方で，全体の18.2％の論文で危険率の調整が行われていたことはわが国の現状と対照されるべき点である（Bezeauら, 2001）。

検定を複数回実施するすべての論文において危険率調整が必要なわけではないものの，5％水準で検定を13回（＝今回の中央値）繰り返したときの全体の危険率は約50％になることを考えると，検定の多重性問題はもっと強く意識され，対処されるべきである。これらの対策に加えて，もっとも避けるべきは，実際は多くの検定を実施したにもかかわらず，都合のいい（たとえば，有意になった）ものだけを報告するという行為である（p-hacking）。これは，読者に正しい知識があったとしても認識・対処することができない。このような行為は意識的にせよ無意識的にせよ，研究不正につながってしまう重大な問題として認識しなければならない。

b. 効果量と検出力の問題

本調査では，t検定のみを効果量検討の対象とした。その結果の前にまず，t検定におけるデータ数，効果量，検出力の理論的関係を確認する。**表1**をみると，たとえば効果量d＝0.5を片側検定・検出力0.8で検出するためには，対応ありの場合には27名，対応なしの場合には102名の参加者が必要となることがわかる。同時に，対応あり・片側検定であれば，データ数が156を超えると，小さな効果量（0.2以下）を検出する確率が高くなることもわかる。なお，対象論文中17編が，実験参加者数150を超えていた。

表1 検出力0.8で任意の効果量dを検出するために必要な1群あたりのデータ数（括弧内は参加者数）

p＝0.05，$power$＝0.8*				
	片側検定		両側検定	
	対応あり	対応なし**	対応あり	対応なし
d	n	n	n	n
0.2	156	310（620）	199	394（787）
0.5	27	51（102）	34	64（128）
0.8	12	21（ 42）	15	26（ 52）
1.0	8	14（ 28）	10	17（ 34）
1.2	6	10（ 20）	8	12（ 24）
1.5	5	7（ 14）	6	9（ 18）
2.0	4	4（ 8）	5	6（ 12）

*統計ソフトR「pwr」パッケージを用いて計算した。小数点第一位を繰り上げて整数にしている。

**対応なしt検定の括弧内の数字は，2群合わせた場合の数値，すなわち参加者数である。対応ありの場合は，1群あたりのデータ数と参加者数は一致する。フリーソフトG*Powerを用いても同様の計算が可能である。

次に，調査対象となったt検定の効果量を，有意差の有無で分けて報告する。有意差があった検定の効果量dの最小値は0.4であった（中央値は0.8）。Cohen（1988）の基準にもとづくと，この数値は小～中程度の効果量である。そのため，結果論ではあるが，今回調査対象となったt検定に関しては，検出力が高すぎるために統計学的に小さ“すぎる”効果を有意と判断したケースはなかったようだ。一方で，有意差なしと判断された検定の効果量dの最大値は0.7であった（中央値は0.29）。もし中程度の効果量（d＞0.5）を仮定して検定を実施したのであれば，検出力が足りなかったために有意差を“得ることができなかった”可能性がある。逆に，もし効果量大（d＞0.8）が学術的に報告するべき差であり，それ以下はそうではないと仮定して当該検定における参加者数を設定したのであれば，その目論見は成功したといえる。ここで重要なことは，ある効果量が学術的に大きいか小さいかという議論は，Cohen（1988）の基準のような統計学的観点ではなく，最終的にはその分野の理論的背景に依拠する点である。たとえば，健常者と患者という要因の効果量は“大きくて当然”であるし，患者に対するリハビリテーションの効果量も“小さな効果では意味がない”ことが多いため，そのような背景を十分に考慮するべきである。

3. 対策とまとめ

調査の結果，神経心理学研究では少なくない論文において，検定の多重性問題が生じていることが判明した（中央値13回，全体としての危険率は約50％）。多数の検定を実施し，かつその有意性に基づいた議論を行いたい場合には，適切な危険率の調整を行う必要がある。これは，仮説探索型の研究でも変わら

8）Bonferroni法は変数間が独立であることを仮定している。そのため，変数が相関している可能性がある場合には，過度に保守的となる点には注意が必要である。

ない（Blakesleyら, 2009；Eichstaedtら, 2013）。それどころか，探索的な性質の研究こそ検定間の論理的関係を仮定できないため，多重比較法による危険率の調整が必要となる（Bezeauら, 2001；Eichstaedtら, 2013；Frane, 2015；Gelmanら, 2017）。ただし，本稿や先行研究での議論は，すべての検定に対する盲目的な危険率調整を推奨するものではない。たとえば，研究の目的から外れた確認的分析のような副次的検定までもカウントして危険率補正をすることは研究の生産性を低めてしまう（Bezeauら, 2001；Gelmanら, 2017；Lezakら, 1984）。

危険率を調整するためにはBonferroni法が簡便かつ有用である。しかし一方で，検出力の過度な低下も招いてしまう。これを避ける現実的な方法は2通りある。1つ目は，検定回数に応じてデータ数を多く設定する方法である（Bezeauら, 2001；Silverstein, 1986）。そもそもの検出力を高くすれば補正後の検出力の低下は相殺される[9]。t検定におけるBonferroni補正とデータ数の関係を**図3**に示す。たとえば対応ありのt検定の場合, $d=0.8$の効果量を検出力0.8で検出するためには12名の参加者が必要になる（**表1**参照）。これに対し，検定を20回繰り返したときにBonferroni補正を用いて危険率を5%に保とうとすると，25名の参加者が必要になることがわかる。このように，データ数（参加者数）を用意することが可能ならば，Bonferroni補正に対しても統計学的には十分に対応できることがわかる。

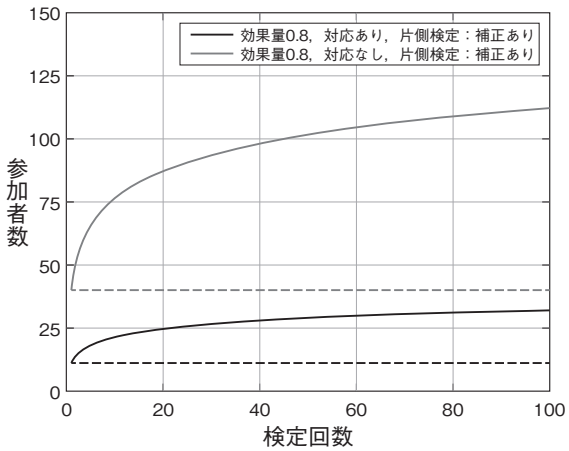

図3 Bonferroni補正による必要参加者数の変化
片側t検定において, 効果量$d=0.8$を検出力0.8で検出する場合を想定している。図中の点線は補正をしない場合, つまり検定回数が1回の場合の必要な参加者数である。

2つ目は，Bonferroni法とは異なる補正方法や統計手法を用いる方法である。Bonferroni法の改良版としてHolmの方法やShafferの方法が提案されており, Holmの方法は手計算でも導入可能である。また，場合にもよるものの，多変量分散分析（MANOVA），またはFalse Discovery Rate（FDR）とよばれる考え方に基づいて危険率を補正する方法も推奨される。しかしながら，これらの方法を用いても検出力の低下は避けられないため，できるだけ1つ目の方法と組み合わせ，さらに検定の回数を減らす努力が必要である。また，FDRは総当りの検定にのみ適用が限定される点にも注意が必要である[10]（松田, 2008）。

仮説探索型研究の場合には，統計的仮説検定を用いないという選択肢もある。なぜなら，仮説探索・生成は記述統計のみを用いても十分可能であるためである。その場合には，効果量が議論の根拠として有効である（Bezeauら, 2001；Gelmanら, 2017）。ただし，p値も依然として有用な情報であるため，p値を報告することが“許されない”わけではないことには留意されたい。問題となるのは，繰り返しになるが，検定の多重性・効果量・検出力を無視し，有意性判断のみに基づいて結果を解釈する行為である。効果量のほかにも，マルチレベル分析（Gelmanら, 2017）や，ベイズによる確率推定も探索型研究において有効な方法であると考えられる。特にベイズ統計は，参加者数の調整が難しい研究向きかもしれない。しかし，ベイズ統計に関しては現段階ではノウハウが蓄積・確立されていない。さらに今回の調査からも推測できるように，雑誌の査読者は古典的な統計手法に関してすら適切な指摘ができない可能性がある。いずれにせよ，どのような統計手法であれ使用者は適切な知識と誠実な態度を身につけなければならない[11]。

9）参加者数(データ数)は実験計画段階で設定したものに従う必要がある。実験開始後にデータをみながら増減させることは許されない。
10）統計的仮説検定を含まない相関行列の報告は検定の多重性とは関係がない。ただし，「相関行列を確認した後に任意のペアを選んで統計的仮説検定（p値の計算）を実施する」のは詐欺行為に当たる。
11）仮説探索によって仮説が生成された暁には，ターゲットを絞った仮説検証型研究を実施することが可能となる。またそれこそが仮説探索型研究を行う意義でもある。神経心理学検査よりもはるかに比較数の多い研究における多重性問題についても同様の議論がなされているので参照されたい（松井, 2018）。

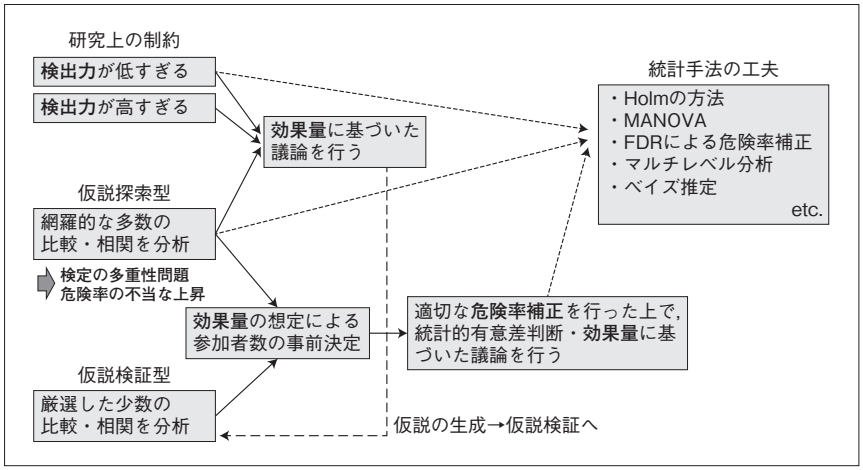

図4　研究計画に関する簡易フローチャート

最後に，今回のt検定のみを対象とした調査においては特別な大きな問題はみられなかったが，効果量と検出力に対する具体的対策について簡潔に紹介する。まず，効果量はだいたいの統計ソフトで簡単に算出可能である。国際的な学術雑誌では報告が強く推奨あるいは義務付けられており，できるだけ記載するべきである。また，先行研究に効果量が記載されていなくても，適切な統計値が記載されていれば事後に算出することも可能である。検出力は，統計ソフトRの“pwr”パッケージや，G*Powerなどのソフトウェアで計算できる。効果量が小さくて有意であった場合には，その効果にどのような意味があるのかをよく考察する必要がある。効果量に基づいた理論的な考察がなされていれば，検出力が大きすぎたとしても問題はない。逆に，効果量が大きく有意でない場合には，検出力が不足していた可能性を検討し，それに言及する必要がある。これまで述べた対策のまとめを，**図4**に簡単なフローチャートとして示す。これも繰り返しになるが，学術的にもっとも大切なのは，統計的有意差判断の結果ではなく，理論に照らし合わせて効果量を考慮することである。

そもそも，データ数やサンプリング方法に制限のある臨床研究においては，統計的仮説検定は第一の選択肢でなくてもよい[12]。今回取り上げた問題以外にも，統計的仮説検定には母集団の決定やランダムサンプリングなど様々な制約があり，神経心理学研究がすべてを満たすことは難しい。本稿は統計手法に関する問題のほんの一部をなぞっただけであるが，今後の神経心理学の発展に少しでも寄与すれば幸いである。

文　献

1）Beeson, P. M., Robey, R. R.：Evaluating single-subject treatment research：lessons learned from the aphasia literature. Neuropsychol Rev, 16：161-169, 2006.
2）Bezeau, S., Graves, R.：Statistical power and effect sizes of clinical neuropsychology research. J Exp Neuropsychol, 23：399-406, 2001.
3）Blakesley, R. E., Mazumdar, S., Dew, M. A., et al.：Comparisons of methods for multiple hypothesis testing in neuropsychological research. Neuropsychology, 23：255, 2009.
4）Cohen, J.：Statistical Power Analysis for the Behavioral Sciences. Lawrence Earlbaum Associates, New York, 1988.
5）Eichstaedt, K. E., Kovatch, K., Maroof, D. A.：A less conservative method to adjust for familywise error rate in neuropsychological research：the Holm's sequential Bonferroni procedure. NeuroRehabilitation, 32：693-696, 2013.
6）Frane, A. V.：Planned Hypothesis Tests Are Not Necessarily Exempt From Multiplicity Adjustment. Journal of Research Practice, 11：2, 2015.
7）藤島喜嗣，樋口匡貴：社会心理学における“p-hacking”の実践例．心理学評論，59：84-97, 2016.
8）Gelman, A., Geurts, H. M.：The statistical crisis in sci-

12）上下肢の運動機能に関するポピュラーな検査においては，MDC（Minimum Detectable Change）やMCID（Minimal Clinically Important Difference）という指標が提供されており，この指標を用いて検査得点の解釈・考察を行うことも可能である。

ence：how is it relevant to clinical neuropsychology? Clin Neuropsychol, 31：1000-1014, 2017.
9）板口典弘，福澤一吉：神経心理学誌における統計解析方法の実態と理論的検討：推測の意味を考える．PsyArXiv：https://doi.org/10.31234/osf.io/n4uc5, 2019
10）Lezak, M. D., Gray, D. K.：Sampling problems and nonparametric solutions in clinical neuropsychological research. J Clin Exp Neuropsychol, 6：101-109, 1984.
11）Loring, D. W., Bowden, S. C.：The STROBE statement and neuropsychology：lighting the way toward evidence-based practice. Clin Neuropsychol, 28：556-574, 2014.
12）松田眞一：FDR の概説とそれを制御する多重検定法の比較．計量生物学，29：125-139, 2008.
13）松井茂之：オミクス研究における検証的解析と探索的解析：多重検定と P 値を中心に．計量生物学，38：127-139, 2018.
14）Millis, S. R.：Statistical practices：The seven deadly sins. Child Neuropsychol, 9：221-233, 2003.
15）水本　篤，竹内　理：研究論文における効果量の報告のために：基本的概念と注意点．関西英語教育学会紀要「英語教育研究」，31：57-66, 2008.
16）Schatz, P., Jay, K. A., McComb, J., et al.：Misuse of statistical tests in Archives of Clinical Neuropsychology publications. Arch Clin Neuropsychol, 20：1053-1059, 2005.
17）Silverstein, A. B.：Statistical power lost and statistical power regained：The Bonferroni procedure in exploratory research. Educ Psychol Meas, 46：303-307, 1986.
18）鈴川由美，豊田秀樹：“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, 2012.
19）橘　敏明：医学・教育学・心理学にみられる統計的検定の誤用と弊害．医療図書出版社，東京，1986.
20）柳川　堯：p 値は臨床研究データ解析結果報告に有用な優れたモノサシである．計量生物学，38：153-161, 2018.