福岡教育大学紀要，第57号，第４分冊，37―47（2008）

心理統計における有意確率 p にかわる新指標：Sanabria & Killeen (2007) によって提唱された，再現確率 p_{rep} の有用性と心理教育研究への応用可能性

An Alternative to Null-hypotheses Significance Testing Based on p Values: The Usefulness of p_{rep} Statistic Proposed by Sanabria & Killeen (2007)

市　原　　学　　杉　村　智　子　　大　坪　靖　直

Manabu ICHIHARA　Tomoko SUGIMURA　Yasunao OTSUBO

（福岡教育大学　学校教育講座　心理学系）

（平成19年10月１日受理）

In a recent article, Killeen (2005) proposed the statistic p_{rep}, the probability of replicating an effect, as an alternative to traditional null-hypothesis significant tests (NHST). In the first part of this article, we describe the historical issue related to NHST in psychology and the currentinternational development of statistical practices including that of the effect size and p_{rep}. We then summarize the article of Sanabria & Killeen (2007) in which the idea of prep and its usefulness in practical decision making are addressed. Finally we highlight applied implications regarding decision making by means of p_{rep}: a new approach to educational research and improvement in the methods of statistic instruction.

Key words: statistics, NHST, prep

はじめに

Killeen (2005) は，従来の統計的仮説検定（NHST）における有意確率 p にかわる新しい指標として，再現確率 p_{rep} を提唱した。再現確率 p_{rep} とは，どのような統計指標であり，従来の有意確率pと比較するとどのような点において異なるのであろうか？　また，海外のジャーナルにおける p_{rep} の採用状況はどの程度なのであろうか？　本稿では，p_{rep} の詳細な算出方法等の数学的観点からの専門的内容にはふれないが，教育心理学的研究に携わっている人向けに書かれたと推測される Sanabria & Killeen (2007)の概略を中心に，新指標である p_{rep} がどのようなものかについて，その有用性も含めて解説していきたい。

I　統計的仮説検定（NHST）の考え方とジャーナル論文における統計記述の状況

ここではまず，統計的仮説検定（NHST）についての簡単な説明と，海外のジャーナル論文において，NHST や本稿でとりあげる再現確率p_{rep} 等の統計的記述がどのような状況であるかについてまとめておきたい。

（1）統計的仮説検定（NHST）における有意確率 p と再現確率p_{rep}

統計的仮説検定（以下，NHST）はFisher（1959）によって提案された意思決定のためのツールで，観察された事象が偶然に生じたものと考え，その事象が偶然に生じる確率を算出する。その確率が十分に低ければ，事象は偶然にではなく，何らかの作用によって引き起こされたのかもしれないとみなすことにしている。例えば，同一母集団から無作為抽出された2つのグループがあり，それぞれ異なる方法で指導を受けるという実験を考えてみる。一方は講義形式の指導（講義群）を，他方はテーマ学習形式の指導（テーマ群）を受け，指導後にテストを受ける。講義群とテーマ群のテスト成績の平均値を比べ，母平均の差異が0か（帰無仮説）否か（対立仮説）を確率論的に推定するのである。心理学では通常有意確率pを5％，もしくは1％に設定して，検定統計量がそれらの値以下の確率でしか見られないのであれば，「有意である」，「有意差あり」と記述する。このようにNHSTは，意思決定のためのツールとして心理学者の間では広く用いられてきた。

しかし，他方では有意水準の大きさが効果量（effect size）と同義であると混同されたり，母集団のあり様が点推定的にわかるといった誤解も見受けられる。このように，NHSTは意思決定のツールとしては有用でありながら，p値の解釈に多少の混乱や誤解が伴うといった問題も散見される。このようななかで，Killeen（2005）は，有意確率pに変わる，再現確率p_{rep}（probability of replication）という新しい指標を提案した。詳細は後述するが，p_{rep}とは，例えば2群比較実験を行いA群よりB群の得点が高かったとすると，追試で同様の結果が再現される確率を指している。

かなり誇張した例えで適切かどうかはわからないが，あえて例えるとすると，p_{rep}とは，“降水確率は教えるので，傘をもっていくかどうかの判断はそちらにおまかせします”といったときの，降水確率のようなものかもしれない。従来の有意確率pによる判断は，有意差があるかないかということが，直接なんらかの意志決定につながる傾向にあった。つまり，有意差があると判断されたら傘をもっていく，という具合である。これに対してp_{rep}が提供する情報は，例えば“降水確率は70％”だけであり，傘をもっていくかどうかの意志決定については，その他の諸条件（例えば，“今日はほとんど地下街しか歩かない”等）を考慮して判断を下す余地がのこされているといえるだろう。

（2）海外のジャーナル論文における統計的記述の動向

それでは，NHSTやp_{rep}等に関する海外の研究論文における状況はどのようなものであろうか。Cumming, et al,(2007)は，APA(アメリカ心理学会)発行のジャーナルを中心とした主要な心理学系ジャーナル10誌で，1998年，2003-4年，2005-6年の3つの時期に発行された論文についての統計的な記述内容についての調査を行っている。これによれば，それぞれの時期のおける，97.8%，97.7%，96.9%の論文が，NHSTを採用しており，現在においては，ほとんどの論文がNHSTによって結果を判断・解釈しているといえるだろう。また，Cumming, et al,(2007)では，American Psychological Association（2001)のパブリケーションマニュアル第5版では効果量（effect size)と信頼区間の記述を推奨するようになったが，個々のジャーナルの投稿規程に，統計的記述に関する具体的な奨励事項の記載があることは少ないことも報告している。

このような状況の中で，Psychological Science誌は，2005年から投稿規定（Blackwell Publishing (2007)によるPsychological Science誌ホームページ）の中で，効果量を記述することと，従来のNHSTにおけるpではなくp_{rep}を採用することを推奨している。このことは，Cumming, et al,(2007)も指摘しているように，かなり異例のことであるといってよいであろう。少なくとも欧米では，効果量の記述については，上述したAPAマニュアルへの記載や，論文審査の段階で，効果量の記載を要求されることが多いことからも，一般的な統計指標となっているといえるが，p_{rep}についてはどのような状況にあるのだろうか。Killeen (2005)がp_{rep}を提唱してから約2年が経過しているが，ジャーナル論文において，p_{rep}はどの程度採用使用されているのだろうか。

Scopus（Elsevier社の学術情報検索ツール）によれば，2005年にKilleen（2005)が Psychological Science誌に，p_{rep}に関する論文を発表してから2007年9月下旬現在，Killeen（2005)を，引用している論文は69件であった。このうち，約53件が実際のデータの統計処理においてp_{rep}を適用していると推測される論文であり，約15件がp_{rep}も含めた統計技術そのものに関する内容であると推測される論文であった。前者においてはやはり，その大部分（44件）はPsychological Science誌に掲載された論文であり，それ以

外のジャーナルでは，Journal of Memory and Language 誌（Elsevier社）等の 9 件であった。Scopus は心理学系ジャーナルのすべてをカバーしているわけではないが，このような状況から推測すると，現在ではまだ，Psychological Science誌を中心としたごく一部の論文においてしかp_{rep} は採用されていないようである。

今後，p_{rep} の採用がどの程度の広がりを見せるのかについては現段階では予想できないが，Sanabria & Killeen（2007）が主張するように，少なくとも教育現場における介入研究等においては p_{rep} が有効な判断指標となりうる可能性が十分に見込まれる。従って本稿では次に，統計学的観点からの専門的な内容ではなく，おそらく教育心理学的研究や教育実践研究に携わっている人向けに書かれたと推測される，Psychology in the Schools誌に 2007 年に掲載されたSanabria & Killeen（2007）の概略を紹介する。この論文では，従来，結果判断の際に用いられてきた，p 値，効果量，信頼区間といった指標の不十分な点と，新指標であるp_{rep} 値がどのようなものかについて，統計の専門家でなくてもある程度理解できるように，具体的な研究例を交えた形で解説している。本稿で紹介するのはあくまでも概略であるので，論文の詳細については，必ずSanabria & Killeen（2007）を参照されたい。さらに，p_{rep} 等に関する統計的な専門的事項については，Killeen（2005）等の関連論文を参照されたい。

Ⅱ　Sanabria & Killeen（2007）論文の概略

2 つの方法のうちどちらかを選択しなければならないとしよう。他の条件がすべて同じであるなら，コストが大きく成果が少ない方法ではなく，成果が大きくコストが小さい方法を選択すればよい。しかし，判断に迷うのは選択肢が甲乙つけがたい場合である。大きな成果は得られるがそれだけコストがかかるといったようなジレンマ的な状況で選択をする際には，専門家や統計家の助けをかりることになる。実際，コストのかかる行動は，少しでもなんらかの改善が示されないと正当化できない。従って，少しでも改善がみられたら，そのわずかの差が統計的に有意な差であるかそうでないのかを知りたがるのである。

従来の帰無仮説棄却型の有意差検定（以下，NHST）は，このような判断をする時に半ば自動的に用いられる方法である。このNHSTの確固たる地位は，心理学や教育心理学の研究論文での統計的判断はほぼすべてそれでなされていることや，教育関係者や心理学学習者対象の統計の授業でも当然のようにそれが教えられていることをみても明白だろう。この論文では，NHST が，よくある 2 者択一型の選択を行う場合にどのように用いられているのかを説明し，NHST は，そのような状況で判断を下すためのツールとしては不十分であることを説明したい。また，最小のコストで最大の効果をあげることを念頭においた確率予測というものは，統計的に有意差があるかどうかだけを判断することよりも有用性があることと，その合理的な理由も同時に主張したい。

（1）[注1)] 従来の帰無仮説棄却型の有意差検定（NHST）の一般的な方法

英語の教授法について，ある新教授法の効果を検証するにはどのようにするだろうか？まず，できるだけ他の条件は同じにして，新教授法群と旧教授法群をもうけてある期間授業を実施し，授業終了時に，TOEFL などの客観テストを使用して，両群の点数の伸びを比較するためのデータを収集する。そして，片側もしくは両側t 検定などの帰無仮説棄却型の仮説検証手続きによって，観察された両群の点数の平均値の差が生じる確率が計算される。また，その差は次の式で標準化され，効果量（effect size）として表される。

d'＝（新教授群の平均値－旧教授群の平均値）/ 2 群をプールした標準偏差　(1)

2 群が同じ母集団からとりだされたと仮定した場合の帰無仮説は，「H_0：新教授群の平均値＝旧教授群の平均値」である。p 値は，帰無仮説が正しい場合にd'が得られる確率である。我々は，単純に 2 択の選択肢を設定し，伝統的な帰無仮説棄却型の仮説検証手続きによって結論を出していた。すなわち，p 値が有意水準（例えば.01）より小さければ，2 群の成績の差は‘有意’であるとし，他の条件が全部等しいと仮定したうえで，おそらく新教授法を採用してきたはずである。しかし，この結論の出し方には，2 つの問題点があるのだ。

信頼性の問題：

もし他の条件が本当に等しいのなら，結論をだすためのツールとしての有意差検定は必要ないであろう。

ある方法と他の方法にかかるコストに差がないのであれば，TOEFLの点がより高くなる方法を採用すればいいだけのことである。しかし，他の条件が等しくないという状況では，どちらを選択するかの判断はたちまち難しいものとなる。英語の教授法の例で言うと，新教授法を採用するからには，実行するコスト（教員の再教育や新しい教材の準備等）を帳消しにするくらいの十分な効果が求められるに違いない。しかし，有意確率pと効果量d'から，効果が十分かどうかをどのように判断するのだろうか？　従来の方法は，効果量（d'）をみることである。もし，効果量が最低基準（例えば$d'=0.3$）よりも大きければ，コストの大きさにかかわらず，平均値が高い方法が採用される。しかしd'は，2群間の差のある一つの推定値にすぎない。いったいそのd'はどの程度確からしいのだろうか？　つまり，1回比較を行った時のd'が0.3だとして，もっと何度も同じような比較を繰り返しても，同じような効果が得られること（つまり，$d'>0$）がどの程度期待できるのだろうか？

よく用いられている方法とは，教授法による得点差（効果量d'）の信頼区間を求めることである。ここでいう信頼区間とは，例えば，教授法の違いによるTOEFLのスコアの比較を1000回行ったとし，さらに1000回ぞれぞれのスコアの差について95%信頼区間を求めたとすると，1000個の信頼区間のうち950個は，その信頼区間の中に，母集団における得点差の値（真の効果量d'）が含まれているということを意味する。しかし残念なことに，たった一度の比較における信頼区間について，1000回比較すればそのうち950回はd'の値がその範囲におさまることであると，誤って解釈されていることが多いのである。信頼区間を使用するにあたっては，この種の解釈の誤りをしないよう，細心の注意が必要であろう。

検出力の問題：

もし，d'は0より大きいけれど，pが.05もしくは.07さらには.25であった場合はどのように判断すればよいのか？　p値というものは，d'よりも大きな差が偶然にでる確率であって，差の程度を表すものではないということを思い出していただきたい。p値が大きいからといって，それは，新教授法が旧教授法より優れていないということではない。本当は存在する差を検出するのに十分な検出力がないだけなのかもしれないのである。

この問題を解決するためによく用いられている方法は，事前に検出力分析を行い，適切な被験者数を設定することである。検出力分析に必要なのは，あらかじめ設定された母集団効果量（δ^*）と，実際の効果量（d'）がδ^*と等しいという帰無仮説が棄却される確率（α）である。意志決定の場面では，δ^*には新しい方法を採用してもいいと判断する最小効果量を設定する。もし差が検出できなければよりコストの少ない方法が選択されるが，その判断には，第2種の誤りを犯す確率（β）があることがよく知られている。従って，新しい教授法を採用する十分な理由となりうる最低限の効果量と，第2種の誤りをおかす確率（慣例的にはβが20%であればよいとされている）という，2つの基準を設ける必要がある。

この検出力分析によれば，上述の教授法の比較研究を行うためには，それぞれの群で少なくとも1000人の被験者が必要になる。この数の被験者を確保することが物理的にも財政的にも無理なのであれば，第2種の誤りをおかす確率は高くなるだろう。しかし，たとえこの数の被験者を確保できたとしても，“有意でない”.07のようなp値が検出されると，問題はもとにもどってしまうのである。従って，検出力分析を行ったとしても，NHSTは，意志決定をするためのツールとしては不十分なのである。

（2）[注1)] NHSTに変わる方法：再現確率

上述したような問題を解決するためには，信頼区間ではわからなかった情報がえられ，かつ，微妙な有意差については1か0かの判断を下さなくてもよいような統計量p_{rep}を使うことをおすすめする。

NHSTにおけるp値から得られる情報について考えてみよう。帰無仮説（H_0）は効果量が0を中心に分布しているという図1-A[注2)]で表現される。陰影をつけられた部分は帰無仮説が真であるという仮定の下，基準となる効果量（d'_1）よりも大きな効果量（d'_2）が得られる確率$p(d'_2>d'_1 \mid H_0)$を表している。これがいわゆる「有意水準」である。有意水準とはある事象が生じたときに帰無仮説が正しいという確率$p(H_0 \mid d'_1)$のことをいう。しかし，有意水準pとH_0が正しい確率は同じではないということに注意してもらいたい。たとえ同じだとしても，追試におけるH_0から抽出される効果の確率は意思決定の根拠となる効果量の期待値とはならない。意思決定のために重要な情報は，追試をした場合に同方向の結果が再現される確率，最小の効果量d'_sを上回る確率（$PR=p(d'>d'_s \mid \delta)$）だろう。ここで，$d'$は積極的な意思決定（例えば，新しい教え方を採用する）を支持するための最小効果量（d'_s）を上回る値で，δは母集団における効果量

平均値である。ESL の教え方の例では，*PR*は，新しい教え方を採用するのに必要なテスト得点の最低限の伸びが得られる確率である。もし，この確率が十分なものならば，新しい教え方を採用してもいいだろう。そうでなければ，この新しい教え方を採用することはできないだろう。

図 1-B[注2)] ではこの考え方を示したものである。母集団効果量の分布は正規分布で表現され，中央に（未知ではあるが）母集団効果量の平均値 δ が置かれる。測定値 d'_1 に対する標本誤差は，

$$\Delta_1 = d'_1 - \delta \qquad (2)$$

で表現される。

ここで，簡便のために最小効果量 d'_s を 0 としておく（すると，ほんのちょっとした改善が見られただけでも，新しい教え方を採用するという意思決定を支持することになる）。図 2B中の $d'_s = 0$ の右側の陰影をつけられた部分が*PR*を表す。ところで，(a) *PR*の算出に際しては，帰無仮説 H_0 は軽視されることもあるし，(b) δ がなければ*PR*を算出することもできない，という 2 点については注意されたい。ただし，δ は未知であるので，d'_1 に基づいて*PR*を推定するしかない。

効果量の分布は正規分布に近似する。そして標準誤差は

$$\sigma d'^2 \fallingdotseq \frac{n^2}{n_1 n_2 (n-4)'} \qquad (3)$$

で表現される。なお，n_1，n_2 は各群のサンプルサイズであり，$n = n_1 + n_2$ である。そして，$-1 < d'_1 < 1$ の範囲をとり，$n_1 = n_2$ ならば（3）式は，

$$\sigma d'^2 \fallingdotseq 4/(n-4) \qquad (4)$$

で簡略化される。

そして，効果量推定値（d'_1）とその分散（$\sigma d'^2$）を用いることで，*PR* 推定値が算出される。これを p_{rep} 推定値とする。

p_{rep} の算出方法

p_{rep} の算出方法は若干専門的で難しい。しかし，図 1 を見れば直観的に理解することはできよう。効果量の標本分散は $\sigma d'^2$ であり，その誤差は 2 倍となる。そして誤差の平方和は標本分散に対応する。したがって，図 2 のCやDで示されているように，p_{rep} の分散は

$$\sigma_R^2 = 2\sigma d'^2 \qquad (5)$$

である。そして，p_{rep} の分布は

$$p(d'_2 \mid d'_1) \sim N(d'_1, \sigma_R) \qquad (6)$$

となる。

p_{rep} の推定値 $p(d'_2 > 0 \mid d'_1)$ または p_{rep} は，図 1-C[注2)] で陰影をつけられた，$d'_1 > 0$ となる部分である。これは正規分布上の

$$z = d'_1 / \sigma_R \qquad (7)$$

における下方領域の面積の割合（確率）と同値である。ここで得られた z 値をExcelの関数NORMS-DIST（）に入力すれば，p_{rep} を算出することができる。

ここで，仮説的な例を用いて p_{rep} の算出方法を紹介しよう。2 つのグループがあり，それぞれ 127 人の学生が割り当てられ（$n_{OLD} = n_{NEW} = 127$；$n = 254$），古い方法と新しい方法というそれぞれ異なる方法でESLの訓練を受けたとする。実験結果に影響を与えうる剰余変数（e.g., 事前のTOEFL得点，年齢，性別，学歴など）については等質性が保証されている。1 学期間訓練を受けた後，各グループそれぞれCBT-TOEFL（得点範囲 40－300 点）を受けた。古い方法で訓練を受けたグループ（Group OLD）の平均値は 206.1，新しい方法で訓練を受けたグループ（Group NEW）の平均値は 240 で，プールされた標準偏差 S_{POOLED} が 169.3 だとする。すると，

方程式（1）より，

$$d'_1 = (240 - 206.1)/169.3 = 0.2$$

方程式（4）より，

$$\sigma d'^2 = 4/(254-4) = 0.016$$

方程式（5）より，

$$\sigma_R = 2 \cdot 0.016 = 0.032$$

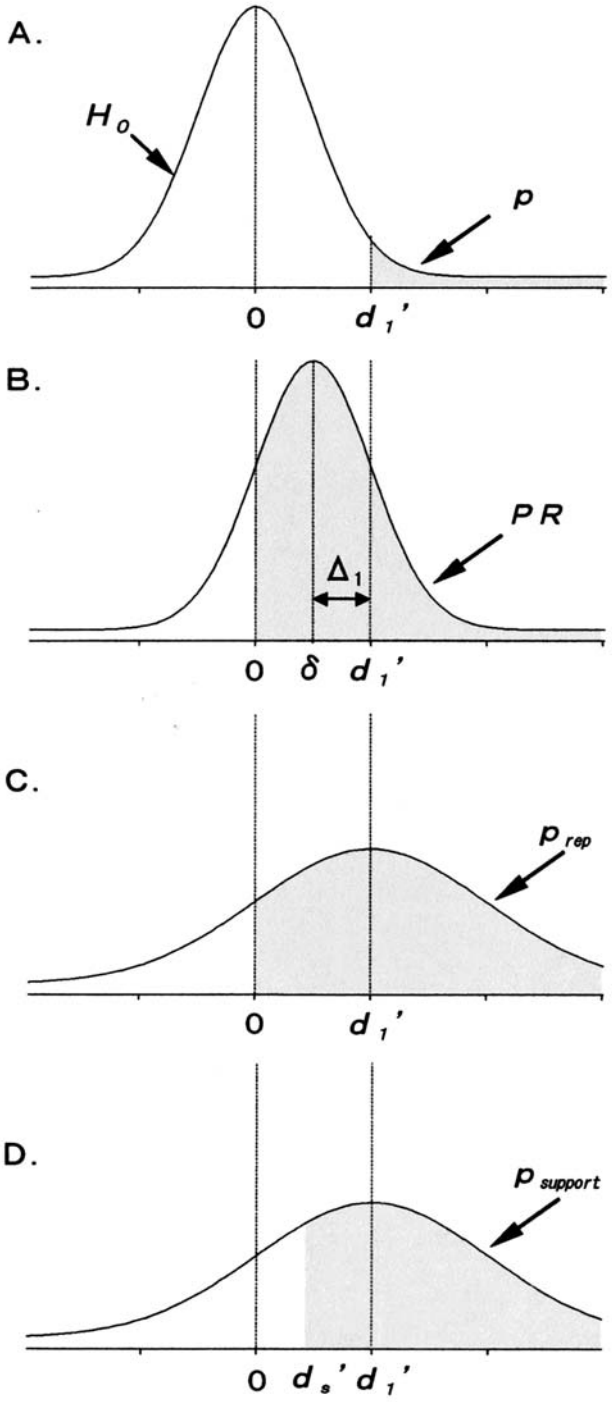

図1[注2)] 統計的検定の概念図
各分布の影の部分は以下のことを示す：(A) 帰無仮説の下でd_1'より大きな効果が得られる確率。(B) δの差がある正規分布の下で0より大きな効果が得られる確率（再現確率；PR)。(C) d_1'から推定された効果量分布の下での$PR(p_{rep})$の推定値。(D) 基準量d_s'より大きな効果量が得られる再現確率の推定値。

方程式（7）より，

$d'_1 / \sigma_R^2 = 0.2 / \sqrt{0.032} = 1.12$

そして最後に，

p_{rep} ＝ NORMSDIST（1.12）＝ .87

これらの結果は，新しい方法で訓練を受けた場合，サンプルサイズは 127 という条件下で，10 グループのうち 9 グループは古い方法で訓練を受けた場合よりもTOEFLでよい点を取ることができることを示唆している。

最小差異基準（d'_s）の導入

最小差異基準（d'_s）が 0 以上でさえあればよいとすると，どんな小さな差異でもよりコストのかかる新しい訓練方法を採用することになってしまう。しかし，ほんの小さな差異であるならば，それが確からしいものであっても新しい方法を採用することはためらわれるのではないか。そうした場合には，まず，二つの訓練方法それぞれのコストの差を計算し，意思決定の基準となる最小の効果量d'_sを決めておくとよい。例えば，外国語ティーチングアシスタントに必要とされるTOEFL の最低得点について考えてみよう。古い方法で訓練を受けた学生のTOEFL 平均点が 206.1 点だったとする。しかしティーチングアシスタントになるには最低 223 点が必要である。合格と不合格の間の最小差異は 16.9 点である。最小差異を効果量に変換し，一般化する。つまり，最小差異をプールされた標準偏差（S_{POOLED}）で除する。ここで各グループの平均値やプールされた標準偏差には上記の例の値（M_{NEW}＝240，M_{OLD}＝206.1，S_{POOLED}＝169.3）を用いる。もし最低 16.9 点の得点の向上が必要ならば，$M_{NEW} - M_{OLD}$ ＝33.9，d'_s＝16.9/169.3＝0.1 である。こうしてみると，新しい訓練方法を採用するには最低 0.1 の効果量が必要であることになる。

p_{rep} が追試において$d'>0$ となる確率の推定値であるのに対して，$p_{support}$ は追試において$d'>d'_s$となる確率の推定値である。$p_{support}$ は図 1-Dにおいて陰影をつけられた$d'>d'_s$となる部分の確率である。そして正規分布上の

$z = (d'_1 - d'_s) / \sigma_R$ （8）

における下方領域の面積の割合（確率）と同値である。

方程式（8）より，

$z = (d'_1 - d'_s) / \sigma_R = (0.2 - 0.1) / 0.18 = 0.56$

$p_{support}$ ＝ $N_{ORMSDIST}$（0.56）＝ .71

となる。この結果は，新しい方法で訓練を受けた場合，サンプルサイズは 127 という条件下で，10 グループのうち 7 グループは平均して 16.9 点以上のテスト得点の向上が見込めることを示唆している。

（3）[注1] p から p_{rep} と $p_{support}$ へ：論文 2 事例を通じて

Scime & Norvilitis（2006）の事例（表 1-1 [注3] 参照）：

Scime & Norvilitis（2006）は，算数の成績とフラストレーション反応に関する研究で，ADHD群と統制群の児童に，難しいパズルの完成課題と，問題を解くたびに難易度が上がるように作成された算数課題に取り組んでもらった。そして，2 群間の，17 の課題成績（パズル課題成績，フラストレーション反応，感情のコントロール能力，算数課題等）の差を比較した。さて，ここでt検定をたくさん繰り返すことになるが，この研究では第 1 種の誤りを回避するためBonferroniの修正を施した。しかし，結果的に有意水準の基準がα＝.003 まで厳しくなってしまい，17 の課題成績のうち，たった 3 つでしか有意差が検出されなかった。また，この研究では，算数課題の全問回答率ににおいては 2 群間に差が見られたが，正答率には差は見られなかった。彼らはこの結果から，“ADHD児は課題の全問回答率では劣るが，回答した問題についての正答率は統制群と差がない。”と結論づけた。

しかしながら，この結論は，表 1-1 に示したように，データからは支持されない。統計的には有意ではないが，統制群はADHD群よりも，算数課題の正答率が高いのである。実際，p_{rep} の値は，もし同様の実験が複数回行われれば，そのうちの 65%の実験で統制群の正答率がADHD群の正答率を上回ることを示している。回答率に関するp_{rep}（99%）に比べてこの割合（65%）ははるかに低いが，正答率において両群の差がないというように解釈すべきではない。65%という数値に意味があるか否かという問題は，理論的，実践

表1-1[注3] Scime & Norvilities (2006) の結果の一部についての統計解析

変数	n	ADHD群平均値(SD)	統制群平均値(SD)	d'	p	p_{rep}
算数の正答率	64	83.81 (15.44)	86.44 (18.71)	0.15	.285	.649
算数の回答率	64	26.88 (22.14)	44.52 (15.69)	0.98	.002	.994
TMMS-C 情動的注意	64	3.30 (0.43)	3.65 (0.56)	0.67	.028	.958
TMMS-C 情動表出	64	3.63 (0.69)	3.85 (0.64)	0.34	.085	.806

*算数の得点は正答もしくは回答した問題の数に基づく
**すべてのp値は，報告されているデータから算出したt検定（対応のない片側検定）の推測値

表1-2[注3] Linares et al. (2005) の結果の一部についての統計解析

変数	n	ベースライン期平均値(SD)	2年後の平均値(SD)	d'	p	p_{rep}
読み(UMSP実施群)	94	2.63 (0.91)	2.81 (0.79)	0.21	.163	.753
読み(統制群)	102	2.54 (1.02)	2.73 (0.89)	0.20	.169	.748
計算(UMSP実施群)	94	2.41 (0.98)	2.75 (0.85)	0.36	.044	.884
計算(統制群)	102	2.78 (1.02)	2.67 (0.95)	−0.11	n.a.	.353

*読みと計算の得点範囲は1（不十分）から4（優秀）
**すべてのp値は，報告されているデータから算出したt検定（対応のない片側検定）の推測値

的な観点から議論，判断されるべきであり，統計的に判断されるべきではないのである。

Scime & Norvilitis（2006）は有意水準を厳しく設定したことで，いくつかの検定結果は有意な傾向しかないと報告した。たとえば，児童用メタ気分尺度（TMMS-C）の情緒的注意を見てみると，追試では96%の確率で統制群がADHD群よりも高得点を取ることが予想される（p_{rep} =.96）。このような高い再現率にもかかわらず，この研究では，その差は有意ではないものとして報告されている。もう少しゆるい基準を設けたとしても有意にはならない結果もあるようだが（例えば，情動表出では，p =.085，p_{rep} = 81%,），このような矛盾は，誤差の割合を比較の集合に設定する（familywise error collection）という厳しい手法によるものだろう。p 値と同じように，p_{rep} もやはり誤差の影響をうける。比較する回数が増えると，標本誤差により，極端なp_{rep} 値が得られる確率も高くなる。多重比較による決定をするときには，このような問題点を考慮すべきであろう。

Linares（2005）の事例（表1-2[注3] 参照）：

Linares（2005）らの介入比較研究では，2年間にわたってUMSPというプログラムを実施した学校の4年生とそうでない学校（統制群）の4年生の，読み能力と算数能力等の14種類の評定値を比較した。Linares（2005）らは，読み能力については，両者に差はないが，算数能力については，測定時期と学校の交互作用がみられたと報告している。この交互作用は，実施校の児童の方が，算数能力の伸びが大きいことを意味している。p_{rep} を見ても，UMSPを採用するよう意思決定してもよいと思われる。

しかし，強い効果が得られているにも拘わらず，学校間のベースライン期の成績の違いがあることで，測定時期×学校の交互作用について明確な解釈をすることが難しくなっている。つまり，UMSPの効果が，ベースライン期の成績の差と関わる剰余変数と交絡してしまっているのである。ベースライン期における両群の成績の差がなくなるよう調整することで，この問題は解決できるであろう。他にはUMSPを受けたことによる読みや算数の成績の伸びの期待値を設定しておくことも1つの手である。最低基準（読みで.10，算数で.26）を設けて，それぞれS_{POOLED} で除して，最小差異基準d'_sを算出する（読みで0.12，算数で0.28）。そして，得られた$p_{support}$ の値（読みで.619，算数で.611）から，追試が5回行われれば，そのうちの3回はUMSPを2年間受けた者は受けなかったものに比べて読みや算数の成績が伸びるということが示唆される。

（4）[注1] 誤った確からしさの感覚をぬぐい去ることで得られるもの

有意でない結果が得られると意思決定をしにくくなるが，有意な結果が得られると真の効果がわかったよ

うな気になってしまう。有意水準が低ければ低いほどその傾向は顕著になる。たとえば，$p < .001$ で有意な結果が得られた場合，効果量の真値について誰が疑うだろうか。このような，誤った確からしさの感覚は，何らかの比較をするときに，その比較の結果見いだされた効果量は母平均や母分散のもとに分布しているものである，とは考えないで，唯一無二のものであると考えてしまうことによるところが大きい。

効果の有無だけを考える 2 者択一的な考え方をするのなら，従来のように 2 つの可能性だけを考えればよい。テストの効果が検出され（いわゆる，有意差がある），その効果が実際にあるとみなすか，もしくは，効果は検出されず（いわゆる，有意差がない），それについては何も言及しないかのどちらかである。特定の 2 つの集団（例えば，2004 年のルイジアナ州とテキサス州の高校生）のある属性（例えば，学力成績GPAの平均値）の差などについては，差があるかないかの単純な判断をすることが必要な場合もあるかもしれない。だが，これはまれなケースだろう。

大抵の場合，研究者が知りたいのは，比較するたびにその集団構成員が異なるような集団間の差なのである。先に示したELS教授法の例では，もう一度比較をすれば，それぞれの教授法に対して 1 回目と異なった反応をする別の生徒が参加しているはずである。2 回の比較のうち，どちらが，真の教授法の効果をよく反映しているのだろうか？　あるものの効果について，それを 1 回効果があったという事実としてではなく，それは真の効果を中心に分布しているもののうちのひとつであると認識していれば，つまり，ある教授法は大抵の場合は有効であるがいつも有効であるとは限らないということをふまえてさえいれば，そのような問を考える必要はないのである。

確率分布の考え方に基づくあいまいさによって，実際の行動をおこす決断がにぶってしまうのかもしれない。しかし，統計手法を用いた洗練された意志決定の方法が誤った仮説のもとに行われている場合には，結局誤った結論をだしてしまうのである。本論文では，d_1 や σ_r を用いて，追試を行った場合の確率分布を推定するという簡単な方法で，このようなミスを減らせることを主張してきた。p 値によって判断すると誤った結論を出してしまうが，p_{rep} を用いると,結果をよく吟味したうえでの意志決定ができるようになるのである。また，$p_{support}$ のように，最低基準を設定するなど効果量の補正を行うこともできる。しかし一番重要なことは，　p_{rep} を用いることによって，今日教育関係者たちが直面している複雑な意志決定場面において，教育関係者同士の意思疎通が容易になることであろう。

Ⅲ　心理教育研究におけるp_{rep}の有用性

以上，Sanabria & Killeen（2007）の概略をみてきたが，この中で言及されている，NHSTに付随する効果量や検出力分析等については，本邦の心理学系ジャーナルにおける統計記述においては，一般的に普及していない状況にある。従って，まず，これらの事項について簡単に言及したうえで，p_{rep} の応用可能性について述べることにする。

（1）NHSTにおける有意水準pと効果量d'

心理学の論文では，時々「1％水準で大きな有意差が見られた」,「5％水準で小さな有意差が見られた」などのように，有意水準の値と効果量を同義のものとして扱っている研究が見られる。確かに，サンプルサイズが一定であるならば，効果量が大きいほど有意水準は低くなる傾向にある。しかし反対に，効果量が一定でも，サンプルサイズが大きくなるほど有意水準は低くなる傾向にある。これはサンプルサイズが大きくなるほど，母数推定値の信頼区間が狭まり，帰無仮説を棄却しやすくなるためである。有意水準とは，帰無仮説が真であると仮定したときに観察された効果量d'_1 以上の効果量d'_2 が得られる確率$p(d'_2 > d'_1|H_0)$を示すものである。つまり，有意水準とは帰無仮説が正しい確率を示しているわけではないし，$1-\ p(d'_2 > d'_1|H_0)$が対立仮説（H_1：$\delta \neq 0$，もしくはH_1:$\delta > 0$）の正しさの確率を示しているわけではない（鋤柄，2001）。こういった有意水準pに対する解釈の難しさは心理学初学者には理解しにくいものだと思われる。

さらにはd'についてでさえも，解釈の混乱が見受けられる。2 群比較実験を行い，d'_1 という効果量が得られ，$p<.05$ で帰無仮説H_1：$\delta \neq 0$ が棄却されたとしても,それは$\delta = d'_1$ を意味しているのではない。d'_1 から推定される区間内に$\delta = 0$ となる領域がないことを示しているに過ぎない。また，心理学実験における無作為割り当て実験で得られた結果は，その被験者にしか適用できない（豊田，2002）ということもあり，追試においては必ずしも同等の効果量を期待できるわけではない。このような効果量に関する問題は実践に

おいては非常に大きな意味を持つことになる。厳密な実験によって有効であると提案された介入技法も，いざ実践で追試したら，結果が再現されない，時には被援助者に危害を加えてしまうという可能性もある。

これらの問題点を踏まえると，有意水準 p や効果量d'だけを頼りにして介入実践を行うのは難しい。実践においては，ある介入技法の有効性，効果量に関する情報もそうだが，どれくらい安心してその技法を使えるかという情報も必要であろう。

（2） NHSTと検出力分析

それでは次に，NHSTを行い，有意な結果が得られなかった（つまり，$p>.05$）場合はどうだろうか。その結果について，多くの人が以下の 2 つのように考えるだろう。1 つ目は，仮説は正しいのだが，実験計画に剰余変数が混在しており従属変数に誤差の影響が出たという考え方，2 つ目には，仮説そのものが間違っていたという考え方である。しかし，第 2 種の過誤について言及することはほとんどないのではないだろうか。第 2 種の過誤とは，「本当は有意な結果であるにもかかわらず，その有意性を拾い出せなかった」という間違いである。そして有意な結果を正確に拾い出すことを検出力という。検出力は独立変数が従属変数に及ぼす効果の大きさ（効果量）とサンプルサイズの 2 つによって規定され，ある程度の効果量があってもサンプルサイズが小さければ，有意性を拾い出せないことになる。従って研究者は実験を行う以前に，先行研究などを手がかりに実験で得られるであろう効果量をあらかじめ予測し，十分な検出力（例えば.80）をもって検定できるようサンプルサイズを決めておくことが望ましい。これを検出力分析という。

しかし，検出力分析にも問題がないわけではない。介入研究ではよくあることだが，実験によって十分な効果量（例えばd'＝1.00 以上）が得られることはまれである。独立した 2 群比較の実験研究で，中程度の効果量（d'＝.50），.50 の有意水準，.80 の検出力で分析を行いたければ，各群 64 名，合計 128 名のサンプルを集めなければならない。この 128 名というサンプルサイズは実験を行うには大きすぎる。特に臨床心理学で行われる治療研究では，多数の等質な患者を集めることが困難であり，例えば 100 名を越えるうつ病患者を集めることは，多くの場合不可能である。したがって，治療研究の中には，サンプルサイズが小さいために検出力が低く，有意な結果が得られず，公表されなかった研究が多数存在する可能性がある。他方で，サンプルサイズを一定に保ちながら検出力を高めるには，有意水準を例えば 5%から 10%へと引き上げればよい。しかし，そうすると今度は第 1 種の過誤を犯す確率が高まってしまう。このように，研究者は第 1 種の過誤と第 2 種の過誤という互いに相容れない問題を抱えながら，分析を行わなければならないのが現状であろう。

（3） 再現確率p_{rep}の応用可能性

心理学研究の目的の 1 つは，人間の行動を記述し，説明することである。この目的を達成するためには，ある程度の個人差は考慮するにしても，一般的な人間の行動を説明する理論やモデルが必要になる。このような法則定立的研究においては，提唱された理論やモデルから導かれる仮説の真偽を判定するために，帰無仮説型の推測統計が用いられてきた。

これらの研究領域では，研究者が主張しようとしている仮説と一致した平均値の様相が観察されたにもかかわらず，統計的に仮説が支持されなかった（$p>.05$）場合，その研究結果は報告されることが少なかった。もちろん，統計的に仮説を支持する結果はそのまま報告されることが多いので，報告された研究結果を総合して結論を導きだそうとしても，“引き出し効果”と呼ばれるバイアス（仮説検証に失敗した研究結果が“引き出しにしまい込まれるように”報告されない現象）のために，総合的な結論を下すことが困難であった。

本論文で紹介した“再現確率p_{rep}”という新しい統計指標は，これらの問題を完全に解決してくれるわけではないが，総合的な結論に到達することを促す可能性を持っている。なぜならば，従来の有意水準とこの指標を併用すると，平均値の方向は仮説に一致したが有意水準に達しなかった結果の報告数が増えると考えられるからである。さらに，このような研究結果の報告数が増えると，理論やモデルを洗練していくための個人変数や状況変数についての示唆が得られるという波及効果も期待できよう。

また，Sanabria & Killeen(2007)が主張するように，新しい教育技法の導入といった意思決定においては，“再現確率p_{rep}”は非常に有益な指標となることが期待される。通常，責任ある意思決定をしなければならない人達が，統計が得意な人達とは限らない（むしろ苦手な人が多いことが一般的である）。そのような人達にとっては，帰無仮説型の有意確率に比べて，“再現確率（p_{rep}）”ははるかにわかりやすい手がかりを

提供してくれるだろう。あえて主張したいことと背反する帰無仮説から始まって，観察された事実との確率的な不整合を根拠に対立仮説を採択する背理法の考え方よりも，“今回観察された平均値の様相と同じ方向の平均値の差が得られる確率は○%です”や“平均点を△点以上向上させる確率は□%です”という考え方の方が容易に理解できる。そういう意味では，“再現確率p_{rep}”は，確率的な現象についての意思決定により多くの人が参加する可能性を提供してくれるのではないだろうか。

引用文献

American Psychological Association (2001). *Publication manual of the American Psychological Association*, 5th ed, Washington, DC: Author.

Blackwell Publishing (2007). Psychological Science: Author guideline. Blackwell Publishing < http://www.blackwellpublishing.com/journal.asp?ref=0956-7976&site =1> (September, 28. 2007)

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., et al. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science*, 18, 230-232.

Elsevier, B. V. (2007). Scopus. Elsevier B. V. <www.scopus.com/scopus/home.url> (September, 28. 2007)

Fisher, R.A. (1959). *Statistical methods and scientific inference*. New York: Hafner

鋤柄増根 (2002). 測定・評価部門　研究法の理解とデータ分析における学生の誤解　教育心理学年報，41, 104-113.

Killeen, P.R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, 16, 345-353.

Linares, L.O., Rosbruch, N., Stern, M.B., Edwards, M.E., Walker, G., et al. (2005). Developing cognitive-social-emotional competencies to enhance academic learning. *Psychology in the Schools,* 42, 405-417.

Sanabria, F., & Killeen, P. R. (2007). Better statistics for better decisions: Rejecting null hypotheses statistical tests in favor of replication statistics. *Psychology in the Schools*, 44, 471-481.

Scime, M., & Norvilitis, J. M. (2006). Task performance and response to frustration in children with attention deficit hyperactivity disorder. *Psychology in the Schools*, 43, 377-386.

豊田秀樹 (2002). 項目反応理論［事例編］　―新しい心理テストの構成法―　朝倉書店

脚注

注 1) Sanabria & Killeen (2007)にはタイトル番号はないが，概略をわかりやすくするために筆者が番号を挿入した。

注 2) Sanabria & Killeen (2007)のFIGURE 2 を参考にして筆者が作成した。本稿ではSanabria & Killeen (2007)のFIGURE 1 については紹介していなため，この図が図 1 となっている。

注 3) Sanabria & Killeen (2007)のTable 1 を一部改編して筆者が作成した。