心理学研究法の新指標“再現確率p_{rep}”の意志決定場面での有用性
——実験データ解析における有意確率pと再現確率p_{rep}の比較——

市原 学（imanabu@fukuoka-edu.ac.jp）
杉村 智子・大坪 靖直・黒川 雅幸・笹山 郁生・永江 誠司
〔福岡教育大学〕

Usefulness of the p_{rep} statistic as an alternative to null-hypotheses significance testing using p values: Comparisons of analytical results based on p_{rep} with p

Manabu Ichihara, Tomoko Sugimura, Yasunao Otsubo, Masayuki Kurokawa, Ikuo Sasayama, Seiji Nagae
Department of Psychology, Fukuoka University of Education, Japan

Abstract

In a recent article, Killeen (2005) proposed the statistic p_{rep}, the probability of replicating an effect, as an alternative to traditional null-hypotheses significant tests (NHST). In this article, two experiments were conducted and their analytical results based on p_{rep} and traditional p values were compared: non-significant results based on p values were reinterpreted as meaningful results in light of power analysis, calculating the effect size and p_{rep} value. The tendency of p value analyses not to reveal non-significant results (i.e. the file drawer problem) and the improvement of decision-making methods are discussed.

Key words

probability of replication, null-hypotheses significant tests, type II error

1. はじめに

本稿では、従来の統計的仮説検定（null hypotheses significance testing: NHST）に準拠する意思決定の問題点を挙げる。そして、NHSTの問題点を克服する可能性を持つかもしれない方法として、再現確率（probability of replication, p_{rep}; Killeen, 2005）を紹介し、p_{rep}に準拠した意思決定の方法について具体例を挙げながら論じる。

従来の実験研究では、あらかじめ有意水準（α、通常.05または.01に設定される）を定めて統計的仮説検定を行い、検定統計量の生起確率（p）が有意水準を下回ったとき、その結果は処遇の効果によるものと結論づけられる。そして、その結果を受けて、母集団においても処遇の効果がみられるであろう（$\delta \neq 0$）というように、結果の一般化を図ってきた。

ところで、有意な結果は、強い処遇効果を得るばかりではなく、サンプルの数を増やすことによっても達成される。これは、サンプルサイズが大きくなることで、母数推定値の信頼区間が狭まり、わずかな処遇効果でも検出されることを示している。裏を返せば、ある程度の処遇効果があるにもかかわらず、サンプルサイズが小さく母数推定値が安定しないため有意な結果が得られず、本来存在するはずの処遇効果を見逃してしまうというおそれもある（第2種の誤り）。

第2種の誤りを回避するために、検出力（$1-\beta$）分析や、効果量（Cohen, 1992）といった指標があるのだが、やはり前者ではサンプルサイズの問題を克服することはできないし、後者ではその結果の確からしさを確認することが難しい。

検出力とは、“本当は対立仮説が正しい場合に、正しく対立仮説を採択する確率”のことで第2種の誤り（β）とは背反の関係にある。有意水準、サンプルサイズ、グループ数、それから検定統計量（または効果量）により求めることができる。したがって、検出力を高め、正しく対立仮説を採択するには、たとえば“有意水準を5%から10%に引き上げる”、“サンプルサイズを大きくする”、それから“処遇の効果を強くする”といった方法がある。三つ目の“処遇効果”に関していえば、“誤差（グループ内の個人差）を小さくする”という方法もある。また，あらかじめ、検出力、有意水準、グループ数、それから先行研究の結果などから想定される検定統計量などを指定しておくことで、必要なサンプルサイズを求めることもできる。あまり強い処遇効果（大きな検定統計量、または効果量）を期待できない研究であるならば、大量のサンプルを確保することで第2種の誤りを回避できる。

しかしながら、あまり強い処遇効果が見込めない上に、大量のサンプルを確保することが難しい研究の場合には検出力分析を行っても、第2種の誤りを回避することは難しい。たとえば、有病率が1%を下回り、かつ患者ごとに多彩な症状を示す精神疾患（例えば、統合失調症）の治療研究においては、誤差の影響もあり強い治療効果は見込めず、また、大規模な無作為配置計画による治療効果の評価も難しい。

また、処遇効果の大きさを示す効果量は、当該研究における結果およびそこから推測される母集団効果量（δ）であり、後続の追試研究における結果の再現性に指針を示すものではない。そもそも、治療・教育実践者にとって知りたい情報は、“ある処遇を自分が行った場合に、多少なりともうまくいくかどうか”であり、処遇の真の効果量ではない場合が多い。そういった意味において、検出力分析や効果量の算出といった方法も不十分なのである。

こうした問題点をふまえて、Killeen（2005）は、実験研究や介入研究においてp_{rep}を使用するよう推奨している。p_{rep}とは、たとえば2群比較の実験研究を行い、グループ間に多少なりとも差異（$d'_1 > 0$）がみられたら、その後の追試において、それと同方向の結果（$d' > 0$）が得られる確率の推定値である。ところで、著者らの知る限りでは、p_{rep}を紹介した論文（市原・杉村・大坪, 2008）はあるものの、わが国ではp_{rep}を用いた心理学研究は見当たらず、日本ではまだ馴染みのないものであると思われる。そこで、本稿では、2群比較研究や1要因3水準の実験計画を用いて、p_{rep}の使用例を紹介する。具体的には、有意な結果が得られなかった2つの実験研究について、検出力や効果量とともにp_{rep}を報告し、その結果の確からしさについて考察を行う。p_{rep}は、従来の統計的仮説検定で有意な結果が得られなかったときに、検出力分析や、効果量といった指標とあわせて用いることで、大きな力を発揮すると思われる。

つまり、NHSTによるデータ分析の結果が有意水準を下回らないとき、その研究は公表されないという“お蔵入り問題”（Mullen, 1989）がある。Rosenthal & Hall（1981）によれば、有意な結果が出たことにより、出版された研究論文がk個あった場合、その背後には$5k + 10$個の有意でなかったために、公表されない（“お蔵入り”をしている）論文が存在すると指摘している。つまり、NHSTに頼る限り、有意な結果の背後には5倍以上の有意でない結果が存在すると推測され、結果の再現性という観点から見れば、NHSTによる意思決定は危険をはらんでいると考えられる。検出力分析や効果量といった指標とp_{rep}を併用することで、こうしたお蔵入り問題や、第2種の誤りといった問題を回避できるようになる可能性がある。

2. 研究1

2.1 概要

本研究ではプライミング効果を検討した。プライミングとは先行呈示された刺激材料が後続の課題遂行に影響をおよぼす現象である。プライミング効果の生起メカニズムについては、意味ネットワーク理論（Collins & Quillian, 1969）による説明が有力である。意味ネットワーク理論によれば、“りんご”、“みかん”などの果物に関する概念は互いに近接し、リンクによって結び付けられている。そのため、“りんご”の概念が活性化すれば、近接する“みかん”にも活性化が拡散し、利用されやすい状態となる。他方、たとえば“教師”という概念は“りんご”に近接していないため、“りんご”の活性化は“教師”までは拡散しない。プライミング効果の研究は、人間の知識構造を解明するうえで有益な知見を提供してくれ、現象の頑健性、再現可能性を確認することは有意義であると思われる。

2.2 方法

2.2.1 実験計画

対応のない2群の無作為配置計画。

2.2.2 実験参加者

大学生20名。プライミング群11名。統制群9名。

2.2.3 実験材料

(a) プライミング課題。プライミング群には「○○大学は“か□ご□”を養成する大学です」という問題を与え、統制群には「××大学は“き□う□ん”養成大学です」という問題を与えた（答えは順に“かんごし”、“きょういん”）。正答率はどちらのグループでも100%であった。(b) アナグラム課題（3題）。“ほ□ん”、“こ□せ□”、“し□ぞ□”（答えは順に“ほけん”、“こっせつ”、“しんぞう”の場合に正答とした）。

2.3 結果と考察

本研究におけるデータはすべて、“R 2.5.1”（R Development Core Team, 2007）によって分析された。特に断りのない限り、研究2についても同様である。

正答数の平均値はプライミング群2.55（$SD = 0.50$）、統制群1.89（$SD = 0.99$）であり、その差は有意ではなかった（$t(11.12) = 1.71, p = .12$）。次に、検出力、効果量、および、p_{rep}を算出した（$1 - \beta = .37, d' = .77, p_{rep} = .93$）。効果量、$p_{rep}$の具体的な算出方法については、Appendix 1および2を参照されたい。Cohen（1992）の基準によると、本研究で得た効果量は比較的大きなサイズである。それにもかかわらず、結果が有意でなかったのは、検出力が低いためだと考えられる。しかしながら、p_{rep}をみると、おそらく今後追試を行うとすれば、本研究と同方向の結果（$d' > 0$）が100回中93回程度得られると予想される。このことから、プライミングは頑健かつ再現可能な現象であろうと推測できる。

それではなぜ、“看護師”ということばによって、“骨折”、“心臓”などのことばがプライミングされるのだろうか。看護師とは医師の指示のもと診療の補助や、患者の療養の世話に従事する職業である。そのため、看護師という概念は身体器官やその損傷に関する概念と近接しており、一方の活性化が他方にまで拡散したのだろう。他方、“教師”という概念はそういった身体器官およびその損傷に関する概念とは近接していないため、プライミング効果が波及しなかったのだと考えられる。

このように、検定結果が有意でないときには、検出力や効果量とともにp_{rep}を参照することで、間違った結論（第2種の誤り）を引き出してしまう恐れを回避できるかもしれない。

3. 研究2

3.1 概要

感情が記憶の再認課題における遂行成績におよぼす影響を検討した。Forgas（1998）によれば、ネガティブな感情を喚起された場合には精緻な情報処理がなされ、記憶課題の遂行成績が向上することが報告されている。本研究ではネガティブ感情、ポジティブ感情を実験的に喚起し、その喚起された感情が記憶の再認課題におよぼす影響を、統制群と比較しながら検討する。

3.2 方法

3.2.1 実験計画と手続き

1要因3水準の実験参加者間無作為配置計画であった。実験は集団形式で実施された。実験は、“エッセイ・物語問題”、“説明文問題”2セッションから構成されていた。“エッセイ・物語”問題で感情操作を行い、“説明文問題”で記憶の再認課題を行った。

3.2.2 実験参加者

大学生28名。ネガティブ感情（以下N感情）群9名、ポジティブ感情（以下P感情）群9名、統制群10名であった。

3.2.3 実験材料

(a) エッセイ・物語問題。N感情群には虐待に関する文章を、統制群には製品開発に関する文章を、および、P感情群にはコメディ文章を読ませた。読了後、12項目の感情語リスト（“笑いたい”、“悲しい”など）に回答させ、感情操作が妥当だったことを確認した。(b)説明文問題。ウィルスや免疫に関する文章を読ませた。説明文問題では、問題文と解答用紙を配り、まず問題文を読ませてから、読了後、解答用紙へ答えを記入させた。内容は解答用紙の中で、問題文から変更のあった箇所を指摘させるものであった。変更箇所は全部で20箇所あった。

3.3 結果と考察

正答数の平均値はN感情群9.11個（$SD = 2.98$）、P感情群7.67個（$SD = 1.20$）、統制群6.90個（$SD = 1.23$）であった。分散分析を行ったところ、有意ではないものの、強い効果量が得られた（$F(2, 25) = 3.09, p = .06, 1 - \beta = .62, \eta^2 = .25, f = .50$）（$\eta^2$や$f$についてはAppendix 3参照）。そこで、N感情群－統制群、P感情群－統制群、N感情群－P感情群についてt検定を行い、検出力や効果量、およびp_{rep}を算出した（Table 1）。

いずれも中程度以上の効果量であったにもかかわらず、有意な結果にならなかったのは、検出力が低かったためと考えられる。p_{rep}をみると、いずれについても、100回中80回もしくはそれ以上の確率で、同方向の結果が得られることを示唆していた。特にN感情群－統制群では、N感情の正答数が上回る確率が100回中91回程度と見積もられており、ネガティブ感情は記憶課題における遂行成績を向上させるという、Forgas（1998）の結果を支持していた。

また、P感情群と統制群の間にも中程度以上の効果量が得られ、100回中81回程度は同方向の結果が得られる可能性が示された。従来の研究知見（Storbeck & Clore, 2005）では、ポジティブ感情は誤記憶を増大させるということが報告されていたが、本研究ではそれとは異なる結果が得られた。本研究で得られたデータからは、なぜこのような結果が得られたのかを明らかにすることはできないので、これ以上の考察は控える。しかしながら、p_{rep}の結果をふまえれば、再現可能性の高い結果であり、今後詳細な検討が望まれる。

NHSTのみに依拠して意思決定していた場合、上述のP感情群の遂行成績は見過ごされてしまいがちである。しかしながら、p_{rep}を用いて総合的に結果を吟味することにより、より深く現象を捉えることができるようになり、“お蔵入り問題”を回避できるようになるだろう。

Table 1：一対比較におけるd'、$1-\beta$、およびp_{rep}

	d'	p	$1-\beta$	p_{rep}
N感情群-統制群	1.00	.05	.54	.91
P感情群-統制群	.63	.68	.26	.81
N感情群-P感情群	.64	.28	.24	.80

4. 総合考察

以上、本稿ではp_{rep}を用いた研究例を紹介してきた。従来の研究では、得られた検定統計量が有意水準を下回るかどうかに注目が集まり、検定統計量が有意水準を上回ったときには、$d' = 0$、および$\delta = 0$と結論づける傾向があった。しかしながら、本稿で紹介してきたように、検出力や効果量とともに、p_{rep}を用いることで、誤った結論を導くおそれを回避することができるようになるかもしれない。たとえば、研究1のような2群配置の実験計画で.80以上の検出力をもって、仮説検定を行いたければ、1群あたり28人のサンプルを集めなければならない(1)。冒頭でも述べたように、有病率が極めて低い精神疾患、たとえば解離性人格障害の治療研究を行うとなれば、これだけのサンプルを集めるのには大変なコストがかかってしまう。このように、p_{rep}は小さなサンプルサイズで実験を行わなければならないようなときに、検出力や効果量とともに用いることで、有効性を発揮するものと思われる。

治療や教育などの実践に携わる者にとっては、技法の真の効果よりも、直面する患者や学習者に対して、当該の技法が効くのか否かが問題となる。このように考えると、所与の統計量から、同方向の結果が再現される確率を推定するp_{rep}は実践家向けの統計指標であるとも考えられる。近年、わが国でも“証拠に基づく臨床心理学”が普及しつつあるが（丹野, 2002）、p_{rep}は、実践家が当該の技法を採用するか否かにおいて、有力なツールになりうると考えられる。

しかしながら、p_{rep}にはNHSTの有意水準のように、どの程度の値を示せば、その結果が信頼できるのかという明確な基準はない。つまり、どの程度のp_{rep}値をもって十分と

するかは、研究者本人や、論文の読者の裁量にゆだねられている部分が大きい。たとえば精神疾患の治療場面において新しい治療方法を実施する場合、その治療方法の副作用に注目し、軽微なもの（例：吐き気、微熱）ならば、.80程度のp_{rep}値を示していれば十分だろうが、重篤なもの（例：意識障害、呼吸困難）ならば、.95以上のp_{rep}値はほしいであろう。これは極端な例ではあるが、新しい処遇を行うにあたっては、その処遇の費用対効果をふまえて、p_{rep}の基準を設定すればよいと思われる。こうしてみるとp_{rep}はNHSTと比べて、明確な基準はないものの、処遇の費用対効果といった実情を柔軟に反映させることができ、より実用的な意思決定のツールと考えることもできる。

引用文献

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155-159.

Collins, A. M. & Quillian, M . R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8, 240-247.

Forgas, J. P. (1998). On being happy and mistaken: Mood effects on the fundamental attribution error. *Journal of Personality and Social Psychology*, 75, 318-331.

市原学・杉村智子・大坪靖直（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　福岡教育大学紀要，57(4)，37-47. (Ichihara, M., Sugimura, T., & Otsubo, Y. (2008). An alternative to null-hypotheses significance testing based on p values: The usefulness of p_{rep} statistic proposed by Sanabria & Killeen (2007). *Bulletin of Fukuoka University of Education*, 57(4), 37-47.)

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

Mullen, B. (1989). Advanced basic meta-analysis. Lawrence Erlbaum Association.（小野寺孝義（訳）(2000)．基礎から学ぶメタ分析　ナカニシヤ出版）

R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rosenthal, R. & Hall, J. (1981). Critical values of Z for combining independent probabilities. *Replications in Social Psychology*, 1(2), 1-6.

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.

Storbeck, J. S. & Clore, G. L. (2005). With sadness comes accuracy: With happiness, false memory. *Psychological Science*, 16, 785-791.

丹野義彦（2002）．認知行動療法の臨床ワークショップ―サルコフスキスとバーチウッドの面接技法　金子書房

Appendix 1: 効果量（d'）の算出方法

効果量（d'）を求める式は以下の通りである（Cohen, 1992）。

$$d' = \frac{M_1 - M_2}{S_{POOLED}} \tag{1}$$

M_1、M_2は各グループの平均値、S_{POOLED}は2群を込みにした標準偏差であり、

$$S_{POOLED} = \sqrt{\frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{n_1 + n_2 - 2}} \tag{2}$$

によって求められる。n_1、n_2は各グループの人数、s_1^2、s_2^2は各グループの分散を表す。

Appenndix 2: 再現確率（p_{rep}）の算出方法

p_{rep}を求める式は以下の通りである（Sanabria & Killeen, 2007）。

$$z = d' / \sigma_R \tag{3}$$

p_{rep}は正規分布中、zの下方分布の面積の占める割合であり、Microsoft Excelの関数NORMSDIST()を用いれば簡単に算出できる。

なお、σ_Rは、効果量（d'）の標準誤差であり、以下の式で求めることができる。

$$\sigma_R = (n_1 + n_2)\sqrt{\frac{2}{n_1 n_2 (n_1 + n_2 - 4)}} \tag{4}$$

n_1、n_2は各グループの人数を表す。

Appendix 3: 効果量η^2、fの算出方法

η^2やfは分散分析における効果量を表し、全体としての処遇の効果の大きさを表すものである。それぞれ、

$$\eta^2 = \frac{SS_{between}}{SS_{total}} \tag{5}$$

$$f = \sqrt{\frac{SS_{between}}{SS_{within}}} \tag{6}$$

によって求めることができる。なお、$SS_{between}$、SS_{total}、SS_{within}はそれぞれ、グループ間の平方和、全体の平方和、グループ内の平方和を表す。

注

(1) Rでは、あらかじめ効果量、有意水準、および検出力を指定しておくと必要なサンプルサイズを出力してくれる関数（power.t.test、power.anova.testなど）が用意されている。

（受稿：2008年7月18日　受理：2008年9月26日）