研究委員会企画チュートリアルセミナー 1

心理学研究における効果量の活用と報告

— APA の指針をふまえて —

Using and Reporting Effect Sizes in Psychological Research :
On the Basis of the APA Guidelines

Hidetoki ISHII, Toshio YOSHIDA, Kensuke OKADA AND Tomokazu HAEBARA

企画・司会者 石井 秀宗 (名古屋大学)
企画者 吉田 寿夫 (関西学院大学)
話題提供者 岡田 謙介 (専修大学)
指定討論者 南風原 朝和 (東京大学)

話題提供

岡田 謙介

企画の趣旨

石井 秀宗・吉田 寿夫

研究における統計分析法の普及と並行して、心理学をはじめ多くの分野で、検定への過度の依存から脱却し、効果量や信頼区間を重視するという方向での改善の提案や努力がなされてきた。日本教育心理学会においても、既に1990年代から、南風原(1991; 1995)、孫(2010)などの講演や報告がなされ、統計的検定の問題点が指摘されるとともに、検定力を用いたサンプルサイズの設計、効果量や信頼区間を用いた実際の効果の大きさの推定の重要性が説かれてきた。こうした動きは世界的な規模で進行し、国際誌では効果量や信頼区間を報告することが必須となっているものも多くなっている。にもかかわらず、わが国の心理学領域の研究論文を見渡すと、統計的検定の問題点を知らないかのような、もしくは敢えて無視したかのような、いまだに検定に過度に依存して論を展開している論文が少なくない。

そのような中、『伝えるための心理統計—効果量・信頼区間・検定力—』(大久保・岡田, 2012) という、まさにこの問題点を捉えた書籍が発刊された。本セミナーでは、その内容が広く理解され、研究論文の改善に活かされることを期待し、著者に講演をお願いした。

統計改革がはじまった

心理学における統計分析と言えば、 t 検定や F 検定といった仮説検定がまず想起されるかもしれない。仮説検定は、実のところ有用な方法論である。仮説の真偽の検討という困難な問題を、仮説検定の手続きによって定型的な作業と2値的な判断に落とし込むことができる。結果の解釈上の利便性と相まって、検定によって $p < .05$ や ** などの表記で記載できる「有意な」結果を得ることが、長きにわたり研究者・分析者の念願とされてきた。そして仮説検定は心理学における統計分析の筆頭的な地位を占めてきた。

しかし、現在は大きな転換点にある。仮説検定に過度に依存することの弊害が多く、研究者から指摘され、これを受け、効果量や信頼区間を重視する立場からの改革が進んでいる。こうした動きは、心理学における統計改革 (statistical reform in psychology) と総称される。

実際、アメリカ心理学会 (American Psychological Association, APA) の最新版の論文作成マニュアルは、「仮説検定は分析のはじまりにすぎません。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ことを、アメリカ心理学会 (APA) は強調します」と明示的にうたっている (American Psychological Association, 2009, p.33)。また同様に、「すべての APA の論文誌において、著者には最低限、仮説検定の結果および適切な効果量の推定値と信頼区間をすべて報告することが期待されます」とも記述されている。実際に APA マニュアル中の投稿原稿例にも、効果量の指標が明示的に記載されている。こうした議論は心理学に限らず、教育学や医学・看護

学、生態学などさまざまな学問分野・領域で活発となっている。

企画の趣旨にもあるとおり、わが国の教育心理学会は、心理学における統計改革に早くから取り組んできた歴史がある。本チュートリアルセミナーではとくに効果量に焦点をあて、APAの指針に立脚しながら、教育心理学研究のための効果量に基づいた統計分析を概観した。また、その解釈や活用法について議論を行った。

検定と効果量

標本サイズ (sample size) が標本の大きさを表す量であると同様に、効果量 (effect size) は研究者が関心のある効果の大きさを表す量である。一般に、検定統計量はこの両者にそれぞれ依存する項の積として

$$\text{検定統計量} = \text{効果の大きさ} \times \text{標本の大きさ} \quad (1)$$

と表すことができる (南風原, 2002, p.163)。検定統計量が大きければ大きいほど¹、検定における p 値は小さくなる。 p 値が分析者が定めた有意水準 α よりも小さければ帰無仮説 H_0 が棄却され、有意という結果が得られる。したがって、効果の大きさが小さくても標本の大きさが十分大きければ、検定統計量は大きくなり、有意という結果が得られやすくなることがわかる。

ところで帰無仮説 H_0 とは「差が 0 である」「関連が 0 である」といった極端な仮説である。帰無仮説が心理学の実際の問題において厳密に正しいことはまず考えられない。つまり「帰無仮説は常に間違っている」(Loftus, 1996)。したがって、現実的には、標本サイズさえ十分に大きくとれば、仮説検定においては有意な結果を得ることができる。極端に言えば、仮説検定は大きな標本サイズを集めた研究者に対する「努力賞」ともみなしうることになってしまう。これは仮説検定について指摘されるいくつかの欠点の中でも、重大な問題点と言える。

一方、(1) 式の関係より、効果量は検定統計量から標本の大きさに依存する部分を除外した量と考えることができ、上記の問題に陥らずに研究者の考えている効果の大きさについての定量的な議論を可能にする。これが現在、仮説検定だけに頼った分析をいましめ、効果量を使った議論を行うことが望まれている大きな理由の 1 つである。

d 族と r 族の効果量

効果量というのは広い概念であるが、2 群の平均値

差を表す d 族の効果量と、2 変数間の関係の大きさを表す r 族の効果量とに大別して考えられることが多い。もっとも、オッズ比やある種のノンパラメトリックな効果量など、必ずしもこれらの 2 分法では論じられない効果量も少なからず存在する。しかし、現状、心理学研究で利用され議論されるのはこの 2 種類が支配的であるため、今回もこの 2 種類を扱う。

推測統計における多くの問題と同様に、効果量を用いた議論にあたっても、母集団の効果量と標本の効果量とを区別することが重要である。

代表的な d 族の母集団における効果量 (母効果量) に、2 群の母平均の差を標準化した

$$\delta = \frac{\mu_1 - \mu_2}{\sigma} \quad (2)$$

がある。ここで μ_1 , μ_2 はそれぞれ第 1・第 2 群の母平均、 σ は 2 群に共通な母標準偏差である。つまり、 δ は 2 群の平均値差が標準偏差いくつ分であることを表す量である。 δ は母集団における効果量であり、実際には標本から推定する必要がある。この目的から、Cohen の d や Hedges の g など、いくつかの標本効果量が知られている。ただし、これらの標本効果量の指標には名称の混乱も見られる。ここでは McGrath & Meyer (2006) など現在の心理学研究における多数派の記述に従い、標本平均の差をプールした標本分散 (分母が $n_1 + n_2$; ただし n_1 と n_2 はそれぞれ第 1 群と第 2 群の標本サイズ) の正の平方根で割り算した標本効果量を Cohen の d 、同じくプールした不偏分散 (分母が $n_1 + n_2 - 2$) の正の平方根で割り算した標本効果量を Hedges の g と呼ぶことにした。なお、これらは δ の不偏推定量ではなく、厳密な補正項を与えた Hedges & Olkin (1985) による δ の不偏推定量や、一方が実験群であり他方が統制群であるときに利用される Glass の d など知られている。

一方、心理学研究では分散分析が用いられる場面も非常に多い。分散分析では、従属変数の分散を要因に由来する分散や誤差の分散などに分解する。たとえば最も単純な 1 要因被験者間 (固定効果) の分散分析は、母集団における全分散 σ_A^2 が要因の分散 σ_A^2 と誤差の分散 σ_E^2 との和に分解されると考える統計モデルである。このとき、母効果量

$$\eta^2 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2} = \frac{\sigma_A^2}{\sigma^2} \quad (3)$$

すなわち全分散に占める要因の分散を考えると、この η^2 を標本から推定する標本効果量にも、

¹ 正確には検定統計量の絶対値が大きいほど、であるが、煩雑になるため本稿ではこのように略記する。

複数の量が提案されている。たとえば、(3) 式における母分散をそれぞれ対応する標本の平方和で置き換えた

$$\hat{\eta}^2 = \frac{SS_A}{SS_T} \quad (4)$$

がまず考えられる。しかしこの $\hat{\eta}^2$ には正のバイアスがあり、一般に (3) 式の η^2 を過大推定する。そこで、これを補正する量として、回帰分析における自由度調整済み R^2 の考え方を使った

$$\hat{\varepsilon}^2 = \frac{SS_A - df_A MS_E}{SS_T} \quad (5)$$

や、(3) 式の各要素の分散を、その対応する不偏推定量で置き換えた

$$\hat{\omega}^2 = \frac{SS_A - df_A MS_E}{SS_T + MS_E} \quad (6)$$

といった量が知られている。ただし df_A は要因の自由度であり、 MS_E は誤差の平均平方である。これら 3 者の間には

$$\hat{\omega}^2 \leq \hat{\varepsilon}^2 \leq \hat{\eta}^2 \quad (7)$$

の関係がある。

要因が 2 つ以上が増えた場合、とくに被験者要因が加わった場合には、母効果量が (3) 式のような

$$\eta^2 = \frac{(\text{当該要因の分散})}{(\text{全分散})} \quad (8)$$

という形ではなく、

$$\eta_p^2 = \frac{(\text{当該要因の分散})}{(\text{当該要因の分散}) + (\text{誤差分散})} \quad (9)$$

という量になることがある。この種の効果量は偏効果量 (partial effect size) と呼ばれる。さらに、これらになるべく多くの研究デザイン間で比較するための量として、一般化効果量

$$\eta_c^2 = \frac{(\text{当該要因の分散})}{(\text{当該要因の分散}) + \left(\text{それ以外の操作} \right. \\ \left. \text{されていない要因} \right) \text{の分散の総和}} \quad (10)$$

も開発されている。これらの量はいずれも母効果量であり、それを推定する $\hat{\omega}^2$ などの標本効果量の具体的な形は研究デザインに依存して変化する。

これら d 族や r 族の効果量の解釈にあたっては、Cohen の基準がよく知られている。これは、 $\delta=0.2$ 程度であれば小さな効果、 $\delta=0.5$ 程度であれば中程度の効果…といった、多くの社会科学分野の結果に基づいて提唱された一種の目安である。しかし、これは

あくまでも大まかな目安にすぎず、分野や検討する問題による差異が非常に大きいことに注意すべきである。Cohen 自身この点について幾度も警鐘を鳴らしていたが、それにもかかわらず同基準を機械的に適用して考察に代える論文は後を絶たない。Thompson (2002) はこのような現状を「 $\alpha=.05$ が使われるのと同じ硬直さで効果量の基準を使うならば、新しい基準を使う愚か者のままだ」と批判している。

効果量は、同じ分野や同じ変数を扱う先行研究と比較されるときに、その力を発揮できると考えられ、先行研究と比較してその意味や大小を論じることが有用であろう。これを系統的・網羅的に行うのがメタ分析である (山田・井上, 2012 など参照)。

標本効果量の実現値は、標本抽出に由来して確率的に変動する。したがって、母効果量を所与の確率で含む信頼区間 (confidence interval, CI) を構成し、これに基づいて母効果量の推定における不確実性を評価することもまた重要である。効果量の信頼区間を用いた区間推定による母集団についての推論は、仮説検定よりも結果の再現性の点で優れた性質を持つ。Cumming (2012) はこれを「 p 値のダンス」のデモンストレーションによって親しみやすく表現しており、本セミナーでもこれを紹介した²。具体的には、

$$\begin{aligned} x_1 &\sim N(50, 20^2) \\ x_2 &\sim N(60, 20^2) \end{aligned} \quad (11)$$

という、母分散が等しく母平均が 10 異なる 2 群からそれぞれ $n_1=n_2=32$ ずつの標本を取り出し、独立な 2 群の t 検定を繰り返してその p 値を「 $p>.10$ 」「 $.05<p\leq.10$ 」「 $.01<p\leq.05$ 」「 $p\leq.01$ 」の 4 群に分類した。また、同じデータで効果量の区間推定も行った。多数回の反復に対して、効果量の 95% 信頼区間は理論どおりに真値を 100 回中 95 回程度含んだ。一方、 p 値は上記の各カテゴリのすべてにわたって広くばらついた。とくに「 $p>.10$ 」が選択されることも多く、 p 値だけに基づく議論は再現性の観点から疑問を呈されうることが見てとれた。

まとめと議論

心理学の主要欧文誌では、すでに多数の論文で効果量の報告がなされるようになってきている。一方で、標本効果量の値が「単に書かれている」もしくは「機械的に基準を適用して解釈を与えたことにしている」といった報告が少なくないのも現状である。

² 次の URL からダウンロードできる。

<http://www.latrobe.edu.au/psy/research/projects/esci>

効果量やその信頼区間を分析結果として用いるにあたっては、得られたその実現値に基づき何が言えるのかを、実践的・臨床的・理論的な各観点からそれぞれ論じていかなければならない。この意味で、心理学における統計改革は、研究者が主体的に考えることを要求する。前述のとおり、結果の解釈には、関連する先行研究との比較が有用である。またメタ分析は効果量を統合して解釈するための手続きを定型化した方法論であると考えることができる。

統計的分析と言えば仮説検定だけで事足りる時代はたしかにあった。しかし、もはやそうではない。本セミナーで紹介した効果量とその信頼区間をはじめとする、検定を補完する方法論を利用して、データから得られた効果やその意味について知恵を絞ることが求められる。

指定討論

南風原 朝和

効果量の relevance と統合的理解

たとえば不眠治療の効果を調べる場合、「平均して何分くらい、睡眠時間が延びたか」を確かめないと、費用に見合う効果の大きさかどうか判断ができない。このような場合は、自然に効果量が注目され、その値を報告することはおおに情報的価値がある。

これに対し、たとえば記憶のメカニズムについての仮説があり、それが正しいならば条件Aのもとでの反応時間が条件Bのもとでの反応時間より短くなると予測されるという場合、反応時間に何 msec の差があったか、それが標準偏差の何倍であったかを報告しても、さほど情報的価値があるとは感じられない。

このように、効果量を算出し報告することの価値や relevance は、研究によって異なると考えられる。「統計改革」を進めるにあたっては、このことも考慮しないと、単なる儀式的強制になりかねない。

また、効果量の普及のためには、公式の表現においても、理解を促進するものを用いたい。

たとえば、多要因の場合の (8) 式の η^2 は、他のすべての要因がモデルに含まれているときに、当該の要因をモデルに追加したときの決定係数の増分

$$\Delta R^2 = R^2_{\text{after}} - R^2_{\text{before}} \quad (\text{D1})$$

であり、重回帰分析においてなじみのあるものである（「部分決定係数」と呼ばれる）。また、このような定式化により、セルの大きさが不揃いのアンバランスデザインにも柔軟に対応できる。

同様に、(9) 式の η^2_p は、上記の決定係数の増分が、当該の要因をモデルに追加する前の非決定係数（説明できていなかった分散の割合）に対してどれだけの比率になるかを示すもので、

$$R^2_{\text{partial}} = \frac{\Delta R^2}{1 - R^2_{\text{before}}} \quad (\text{D2})$$

と表される「偏決定係数」である。

このような統合的理解を助ける表現が、効果量の「理解を伴う、意味ある活用」に有効であると考えられる。

引用文献

- American Psychological Association (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington DC : APA.
- Cumming, G. (2012). *Understanding the new statistics : Effect sizes, confidence intervals, and meta-analysis*. New York : Routledge.
- 南風原朝和 (1991). 有意性検定からの脱却は可能か 日本教育心理学会第 33 回総会発表論文集, L 17-L 18.
- 南風原朝和 (1995). 教育心理学研究と統計的検定 教育心理学年報, 34, 122-131.
- 南風原朝和 (2002). 心理統計学の基礎—統合的理解のために— 有斐閣
- Hedges, L. V., & Olkin, L. (1985). *Statistical methods for meta-analysis*. New York : Academic Press.
- 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.
- McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree : The case of *r* and *d*. *Psychological Methods*, 11, 386-401.
- 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
- 孫 媛 (2010). 教育心理学研究の明日のために 教育心理学年報, 49, 109-119.
- Thompson, B. (2002). “Statistical,” “practical,” and “clinical” : How many kinds of significance do counselors need to consider? *Journal of Counseling and Development*, 80, 64-71.
- 山田剛史・井上俊哉 (2012). メタ分析入門—心理・教育研究の系統的レビューのために— 東京大学出版会