日本教育心理学会第５６回大会自主企画シンポジウムJH01

文系学生に対する心理統計教育

～分散分析の理論と実践について～

話題提供１．文系学生に対する分散分析の数理の教育

寺尾敦　青山学院大学

　　　　話題提供２．”効果” を解釈する―分散分析における効果量―

　　　　　　　　　　　　　　井関龍太　理化学研究所

　　　　話題提供３．「心理学の分散分析」の問題点とモデル理解の重要性

　　　　　　　　　　　　　　星野崇宏　東京大学

　　　　指定討論　　　　　　千野直仁　愛知学院大学

　　　　　　　　　　　　　平成26年11月9日

　　　　　　　　　　神戸国際会議場（担当校：神戸大学）

**指定討論資料**

愛知学院大学心身科学部　千野直仁

討論内容

1. 当シンポジウムの意義
2. 話題提供者の話題
3. 分散分析の理論と実践についての各種の問題
4. 効果量の簡潔な定義（教義の効果量の定義？）
5. 各指定討論者の発表へのコメント
6. 効果量と今後の課題

引用文献

1. **当シンポジウムの意義**

分散分析の理論と実践は心理学をはじめとする多くの分野で古くて新しい問題であり、分散分析をきちんと理解することは統計学の基礎知識の学習にとっては格好の材料と考えらえる。

ただし、これを文系の学生にどのように教えるべきか、さらには一般の必ずしも統計学的方法を専門にしていない多くの心理学関連の研究者に対してこれをどのようにして普及させるべきか、は悩ましい問題であり、その意味でこの問題をタイトルに据えたシンポジウムを今回企画された山田、村井、杉澤、寺尾先生に敬意を表したい。

1. **話題提供者の話題**

寺尾先生は、文系学生に対する心理統計教育では、１）数学からなるべく逃げない、２）数理の理解を助けるために統計ソフトを活用する、ことを提案している。

一方、井関先生は、分散分析における効果量を用いて（要因の）効果を解釈することの意味とその解釈の重要性について述べている。

最後の星野先生は、いわゆるr族の効果量について、１）分散分析を回帰分析モデルとして理解するときの両者の違いや、２）交互作用の有無が異なる研究間での効果量の比較の問題を議論している。

ともに、大変興味深い内容で、大いに考えさせられた。

1. **分散分析の理論と実践についての各種の問題**

　　分散分析は、内外の心理学関係の学術雑誌の多くでいまだ頻繁に利用されている。それにも拘わらず、分散分析の理論のみならず実践においても、いまだに十分周知されていない問題や誤解が幾つか存在する。それらの主要なものをあげると次のようになる：

１）全体的交互作用が有意な場合の下位検定としての単純交互作　用検定の問題点と、処理・対比交互作用検定 (Kirk, 1982, 1995)。

1. 主効果と交互作用のどちらを先に検定するかの問題（MANOVA では内外のテキストでほぼ、交互作用項の検定が先になされるが、ANOVA では一貫していない）
2. 全体的交互作用が有意な場合の主効果の意味の問題（広津、1981; Scheffé, 1959)
3. 分散分析と重回帰分析の基本的な違いー無作為割付けの有無
4. 反復測定デザイン ANOVA における球形（球面性）仮定（sphericity assumption）のやぶれと、F統計量の p値のインフレ
5. 反復測定デザインANOVA における、大局的・局所的球形（球面性）仮説 (global, local sphericity hypothesis) (Crowder & Hand, 1990; Kirk, 1982; Mendoza, Toothaker, & Crain, 1976) や多標本球形（球面性）仮説 (multi-sample sphericity hypothesis) (Harris, 1984; Mendoza, 1980; Huynh and Feldt, 1970; Kirk, 1982; Sugiura and Nagao, 1968) の無視（例えば、心理学研究誌でもしばしば、Psychophysiology 誌でも未だにほぼすべての論文で、これら２つの概念の記述はない！）
6. 効果量の定義の多様性とあいまいさ（例えば、Kelley & Preacher, 2012）
7. **効果量の簡潔な定義**

Kelley ら (2012) が指摘する効果量の定義の多様性 (variability) や曖昧さ (ambiguity) は、どこから来るのであろうか。

筆者は、それは効果量の問題を初期に指摘した Cohen (1962) 論文が計量心理学や数理心理学分野の論文ではなく、**Journal of Abnormal and Social Psychology** 誌であったことによるのでは、と推察する。つまり、Cohen はこの雑誌の読者が文系である可能性を考え、寺尾先生の言葉を援用すると、「数学から逃げた」とまでは言えないであろうが、「数学を避けた」、のではないかと推察する。

実際、彼は同論文中の効果量 (そこでは、彼は効果量をSize of Effectと呼んでいる) を説明するために、**検定力（検出力）分析**という言葉、および主として数理統計学の分野の検出力関連の引用文献をTable 1 にきちんと引用はしているが、Table 2 として数種類の母数としての効果量とそれらの大中小の大雑把な基準を示しているのみで、それらの母数がどのようにして導かれるのか、またそれらは検出力概念との関連ではどんな位置づけなのか、などについては完全に省略している。

一方、同じく心理学関係の雑誌 Psychological Bulletin 誌に掲載された Cohen (1992, p.156) では、やはり数式表現ではないが、言葉で効果量の定義を明確に記述している。すなわち、

 “In the Neyman-Pearson method …. **The degree to which H0 is false is indexed by the discrepancy between H0 and H1 and is called the ES**.” もちろん、H0は**帰無仮説**を、H1 は**対立仮説**を、またES は**効果量**をさす。

筆者は、Cohen のこの定義が、効果量を数学を用いず言葉で定義するときの、もっとも簡潔な効果量の定義であると考える。もっとも、この定義だけでは、「**H0 とH1 の乖離**」が何を意味するのかがあいまいである。明らかに、この乖離は、H1 のもとでの統計量の分布の形を決めるパラメータのうちの**非心母数** (non-central parameter) であり、効果量はこの値からサンプル数にかかわる項を除いたものである。母数の中にサンプル数が入るのは、個々の検定における統計指標（例えば、平均差のt値やANOVAの要因の効果のF値等）が、標本の関数としてのいわゆる**統計量**だからである。

これらのことを図で示すと、次のようである。



　ちなみに、平均の差のt検定では、各群の母平均をμx μy、両群に共通の母分散をσ2、両群のサンプルサイズをそれぞれ Nx、Nyとすると、帰無仮説H0: μx =μy に対して、H1すなわち平均に差があるという対立仮説のもとでの非心t分布の非心母数λは、

　　　　　　　　　　$λ=\frac{μ\_{x}-μ\_{y}}{σ} \sqrt{\frac{N\_{x}N\_{y}}{N\_{x}+N\_{y}}}$

であり、効果量は右辺第１項である。ただし、上の図は片側検定の場合の危険率αや、検出力1－βを示してある。

つぎの図は、自由度 10の中心t分布（帰無仮説H0のもとでの分布）に対して、同一自由度の非心t分布（対立仮説H1のもとでの非心t分布）を、非心母数λを1から4まで変えてプロットしたものである。この図から、H0 からH1への乖離の大きさの変化がわかる。

彼は、さらに、つぎのように述べている：

“Each statistical test has its own ES index.”

この表現も、しかしながら、数式を用いないと曖昧さを免れない。しかし、このことは（パラメトリックな場合の）統計検定では対応する統計量の分布が、たとえ統計量が同一分布（t とかFなどの）であっても、モデル、サンプルサイズ、帰無仮説、対立仮説が異なれば、当然非心分布は異なるものとなるので、当然である。

　最後に、**各種分析ごとの効果量と対応する非心分布の非心母数の値を明記してあるのが**、Faul et al. (2007) とそのマニュアルであり、

効果量等計算の**ウエブ上のフリーソフト** G\*Power 3 も、大変便利なものである。

1. **各指定討論者の発表へのコメント**

うえの議論をもとに、本シンポジウムの指定討論者の話題について若干のコメントをすれば、つぎのようになる。

　まず、寺尾先生の話題の主張の１つである、文系の統計教育において、必要な部分は数学から逃げない、という方針はやはり大変重要であると思われる。一方、もう一つの主張である、数理の理解を助けるために統計ソフトを活用する、という方針については、重要ではあると思われるが、あまり「バカチョン」方式の便利すぎるソフトは、大変危険性をもはらんでいるのではなかろうか。しかし、一方では、統計ソフトでもプログラミング言語に近い簡易言語を用いなければならないソフトは、文系の多くの学生にはそれをマスターすることがそれほど容易ではない、という難点もあるものと思われる。

　つぎに井関先生の発表内容については、Kelley et al. (2007) を引用して、「効果量という特別な統計量があるというわけでは必ずしもない」とし、Kelley et al. らの「**現象を反映するあらゆる数量が効果量になりうる**」としているが、文系の学生や計量心理学や数理心理学、数理統計学等を専門としない一般の研究者に対しては、まさにKelley et al. 論文で指摘されている「効果量のイメージが多様でありあいまいな部分を持つ」という印象を与えるので、必ずしも好ましくないのではなかろうか。

　筆者は、効果量という概念の簡潔な定義は、Cohen (1992) の言葉で言えば、（パラメトリック検定の場合）各検定における個々の統計量に対するH0 とH1 のもとでの統計量の分布間の乖離であり、より具体的には、効果量は、対立仮説H1のもとでの統計量の分布（非心分布）の非心母数の中のサンプル数にかかわらない母数の部分、とするのがあいまいさを排除した定義であると考える。

　もちろん、効果量は母数なので通常標本から推定することになるので、モデルの違い、モデルの誤差分布の何らかの原因によるゆがみ、推定方法の違いなどが効果量の定義に多様性を生む。また、当然の帰結として標本効果量はサンプルサイズの影響を受ける。

　さらに、今回の指定討論者のお二人が、分散分析の（母）効果量として共にη2 （$=^{σ\_{A}^{2}}/\_{σ\_{T}^{2}}$）を用いているが、うえの議論での非心母数がらみの定義では$ f^{2}$（=$^{σ\_{A}^{2}}/\_{σ\_{E}^{2}}$）である。もっとも、両指標間の関係はどちらから見ても他方は一方の**単調増加関数**なので、その意味では Cohen (1992) 流の「乖離」の定義からは、どちらも効果量といえる。

　最後の星野先生の発表内容については、非心母数から得られる指標 f2 による効果量の定義を用いると議論は推定方法等の違いを除き単純なものになるが、先生の発表では分散分析を回帰分析モデルの範疇で捉えているため、効果量の指標としてη2 と関連指標を用いていることは自然であり、モデルを変えることにより標本（変量）の分散成分の違いによるいろいろな指標の検討は興味深い。

　また、発表資料の中に出てくる「本来実験では説明変数をランダムに変動させている」というときの「ランダム」の意味は、モデルが混合模型の場合の「**当該要因が確率変数である**」ことの意味で、母数模型の場合の要因の**各水準への（Fisher の実験計画法の３原則の中の）無作為割り付け**の意味ではないのではないか。

1. **効果量と今後の課題**

　最後に、APA の（論文）出版マニュアルで近年p値や検定結果以外に効果量（さらにはその信頼区間）を添えることが強く勧められているが、p値の場合と同様、しゃくし定規にこれを添えればよいという風潮が生まれるのであれば、大変安直な議論であると思われる。なぜならば、効果量はもともと帰無仮説に対して対立仮説を考える時の当該統計量の非心分布の非心母数で、サンプルサイズを除いた部分をさし、母数としてあくまでも仮説的構成概念と考えられるからである。

　このように考えるとすれば、効果量を標本から推定することは１つの方法ではあるが、あくまでも統計学の枠組みの中での話であり、効果量については、内外の効果量についての論文の著者の一部が指摘しているような当該現象に関する**substantive な意味**も考慮する必要があろう。さらにこの議論を一歩進めると、そのようなsubstantive な内容が拠って来るところの論拠としての当該分野の**理論法則**の構築も必要であると思われる。

**引用文献**

Cohen, J. (1962). The statistical power of abnormal-social psychological research:　A review. *Journal of Abnormal and Social Psychology*, **65**, 145-153.

Crowder , and Hand, (1990). *Analysis of Repeated Measures*. London: Chapman and Hall.

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

廣津千尋 (1981). 分散分析 教育出版

Harris, P. (1984). An alternative test for multisample sphericity. *Psychometrika*, **49**, 273-275.

Huynh, H., and Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact F-distributions. *Journal of the American Statistical Association*, **65**, 1582-1589.

Kelley, K., and Preacher, K. J. (2012). On effect size. *Psychological Methods*,

 **17**, 137-152.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences*. Monterey: Brooks/Cole.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences*. Third Edition. Monterey: Brooks/Cole.

Mendoza, J. L. (1980). A significance test for multisample sphericity. *Psychometrika*, **45**, 495-498.

Scheffé, H. (1959). *The Analysis of Variance*. New York: John Wiley & Sons, Inc

Sugiura, N., & Nagao, H. (1968). Unbiasedness of some test criteria for the equality of one or two covariance matrices. *The Annals of Mathematical Statistics*, **39**, 1686