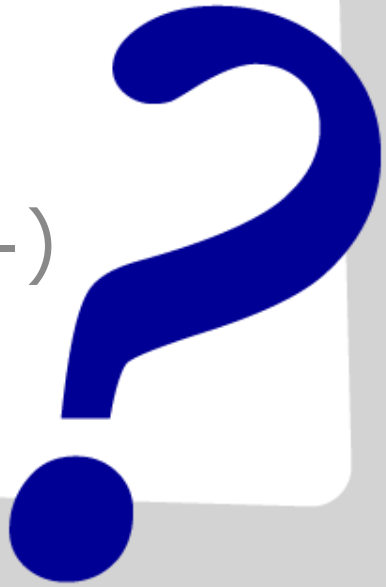
日本教育心理学会第56回総会@神戸国際会議場（2014.11.9）
自主シンポジウムJH01　文系学生に対する心理統計教育

“効果”を解釈する
－分散分析における効果量－

井関龍太
（理研BSI-トヨタ連携センター）



※この発表は個人の見解に基づくものであり，所属する
組織の公式見解を示すものではありません

近年の動向

- 各種ジャーナルの編集方針
  - 帰無仮説検定の結果だけではダメ
  - 効果量や信頼区間を使って効果の度合いを評価することを要求
- 効果量を報告する論文の増加
  - 2009～2010年のJEP:G掲載論文の58%が効果量を報告（Fritz et al., 2012）
    →懸念される事態：「心理統計の単位を取ったのに，多くの論文で報告されている効果量の意味がわからない！」

概　　要

- 効果量って何だろう
  - 効果量という特別な統計量がある？
  - サンプルサイズとは独立？
  - 効果量と検定とは別物？
- 効果量をどうしよう
  - 効果量が大きければ有意でなくても効果を主張できる？
  - 効果量をどう解釈する？

効果量という特別な統計量がある？

- 必ずしもそうではない
  - 関心のある問題に答えるために使える，現象を反映するあらゆる数量が効果量になりうる（Kelley & Preacher, 2012）

素朴な効果量の例

- 例：ある学習法がテストの成績を向上させる効果を調べる
  - 効果量１：平均差
    - ✓学習法を実施したときは，しなかったときよりも平均５点成績が上がった
  - 効果量２：成績が上がった人の割合
    - ✓学習法実施により実施前よりも成績が上がった人は全体の56%であった
- ただし，これらの量はばらつきを考慮していないし，測定単位に依存する

非標準化効果量と標準化効果量

- 非標準化効果量：測定した元の単位に依存する
  - 平均差
  - 割合　　など
    - ✓解釈がしやすい
- 標準化効果量：元の単位に依存しない
  - d 族：Cohen の d，Hedges の g
  - r 族：η^2，ω^2　　など
    - ✓特定の測定法を超えて結果を一般化できる

d族とr族

- d族：標準化平均差

$$d = \frac{\text{平均}_1 - \text{平均}_2}{\text{標準偏差}}$$

  - t検定の効果量としてよく使われる（2群の比較）

- r族：分散説明率 ← 今回はこちらを取り上げる

$$\eta^2 = \frac{\text{平方和}_{\text{要因A}}}{\text{平方和}_{\text{全体}}}$$

  - 分散分析の効果量としてよく使われる（群の数に限定されない）

効果量という特別な統計量がある？

- 必ずしもそうではない
  - 関心のある問題に答えるために使える，現象を反映するあらゆる数量が効果量になりうる（Kelley & Preacher, 2012）
- 単に“効果量”というときには，標準化効果量を指すことが多い
  - ただし，非標準化効果量が不適切なわけではないし，文脈によってはこちらのほうが有用なこともある

サンプルサイズとは独立？

- 理想：よい効果量の指標はサンプルサイズとは独立に効果の大きさを表す
- 実際：いろいろな理由から効果量の指標はサンプルサイズと無関係ではない

- $\eta^2 = \frac{SS_{Effect}}{SS_{Total}}$

✓ サンプルサイズが小さいときにバイアスが大きい

- $\omega^2 = \frac{SS_{Effect} - df_{Effect} \times MS_{Error}}{\left(SS_{Effect} - df_{Effect} \times MS_{Error}\right) + N \times MS_{Error}}$

✓サンプルサイズの影響を統制するため，式にサンプルサイズが含まれる

効果量と検定とは別物？

- 検定と共通する統計量を用いる
  - 平方和，サンプルサイズなど
- 指標によっては検定が解釈に関係する
  - η^2：効果が有意な場合のみ解釈可能；有意でない場合は基本的に 0 になるため（Fay & Boyd, 2010）
- デザインの影響を受ける
  - 分散分析の効果量では，要因計画が大きな影響を及ぼすことがある

複数の要因があることの影響

ロ一要因の
被験者間計画

- $SS_A = 100$
- $SS_e = 900$
- $SS_T = 1{,}000$
  →$\eta^2 = .10$

ロ二要因の
被験者間計画

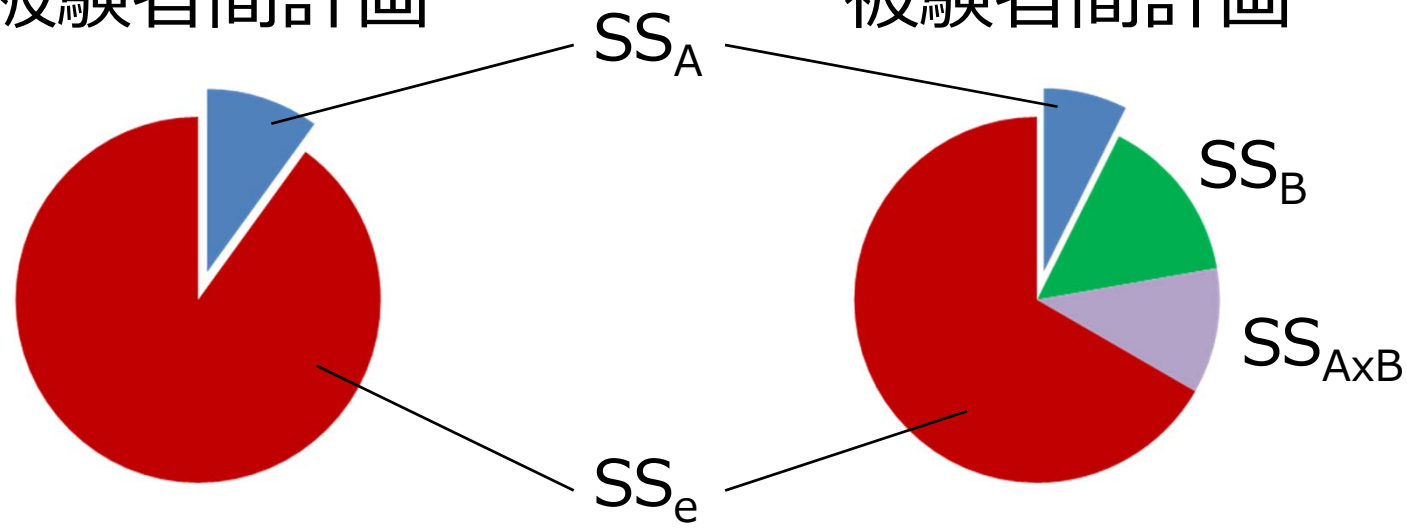
- $SS_A = 100$
- $SS_B = 200$
- $SS_{AxB} = 150$
- $SS_e = 900$
- $SS_T = 1{,}350$
  →$\eta^2 = .07$

- 要因Aはどちらのデザイン
- でも同じだけの分散を説明している
- 誤差平方和も同じ
  →しかし，効果量の大きさが違う

複数の要因があることの影響

ロ一要因の
被験者間計画

ロ二要因の
被験者間計画

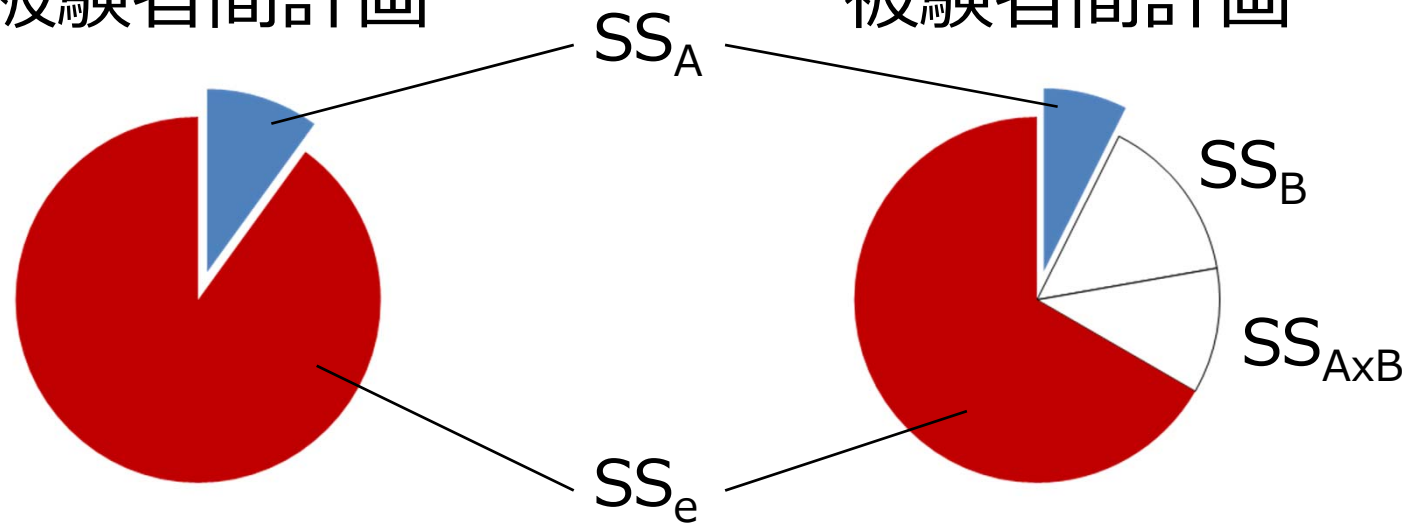


- η^2の計算では，関心のある効果（要因A）の平方和が全体に占める割合のみ考慮する
→他の要因が増えるほど全体平方和は大

他の効果の影響を除外する

ロ一要因の
被験者間計画

ロ二要因の
被験者間計画



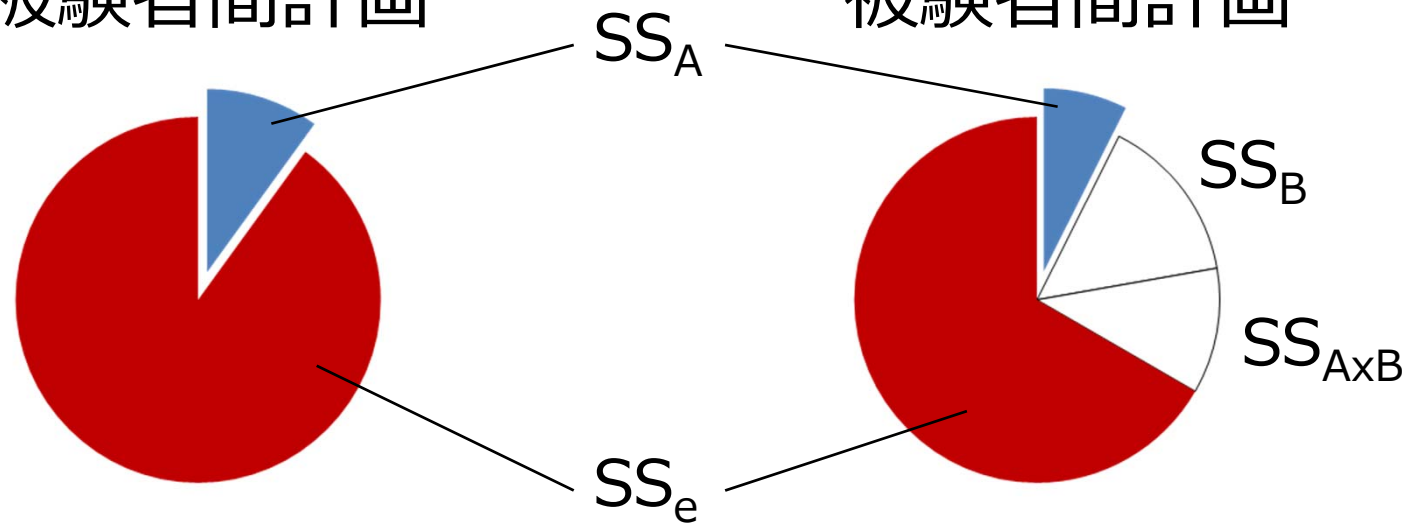
$$\eta_p^2 = \frac{SS_{Effect}}{SS_{Effect} + SS_{Error}}$$

- η_p^2（偏イータ二乗）の計算では，他の効果の平方和を含めない

他の効果の影響を除外する

□一要因の
被験者間計画

□二要因の
被験者間計画



要因Aの効果量
- η^2 ＝.10
- η_p^2 ＝.10

⟷

要因Aの効果量
- η^2 ＝.07
- η_p^2 ＝.10

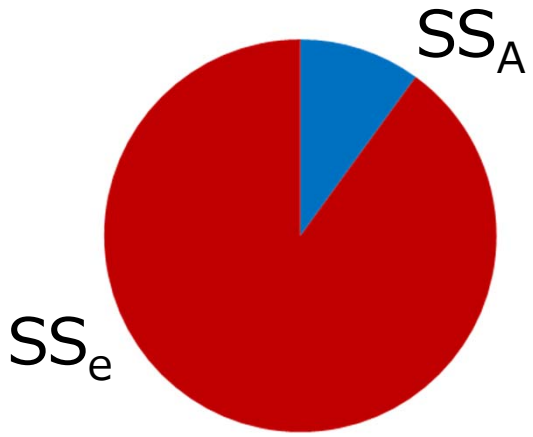
他の要因の存在にかかわらずη_p^2は同じ値

ただし，η_p^2の使いどころは難しい

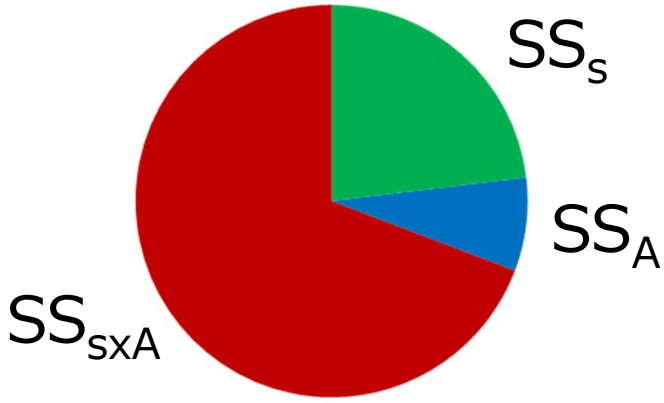
- 同じデザインの中の複数の要因の効果の比較には使えない
  - ベースラインとなる分母が異なる
    （η^2は和が 1 になるが，η_p^2はならない）
- デザインが違う研究との比較には使えない
  - 被験者間計画，被験者内計画，混合要因計画のいずれに当たるかが違うと一般化できない

被験者間か，被験者内か

- 一要因の**被験者間**計画



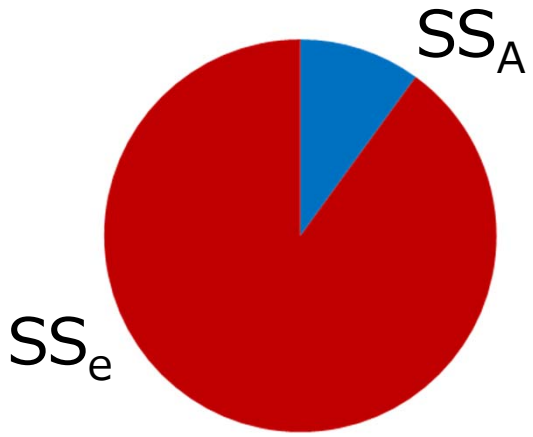
- 一要因の**被験者内**計画



- 被験者内計画では，被験者による誤差項（個人差の影響；SS_S）を誤差平方和（SS_e）から分離する

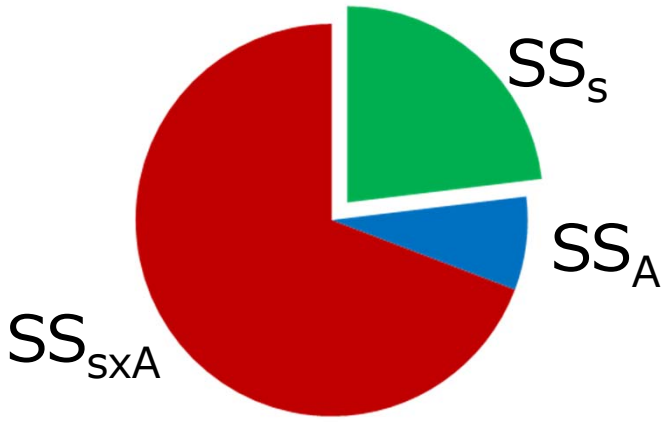
η_p^2を計算すると……

- 一要因の**被験者間**計画



$$\eta_p^2 = \frac{SS_A}{SS_A + SS_e}$$

- 一要因の**被験者内**計画

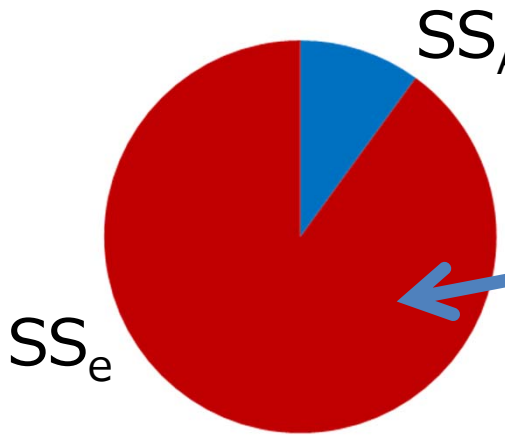


$$\eta_p^2 = \frac{SS_A}{SS_A + SS_e}$$

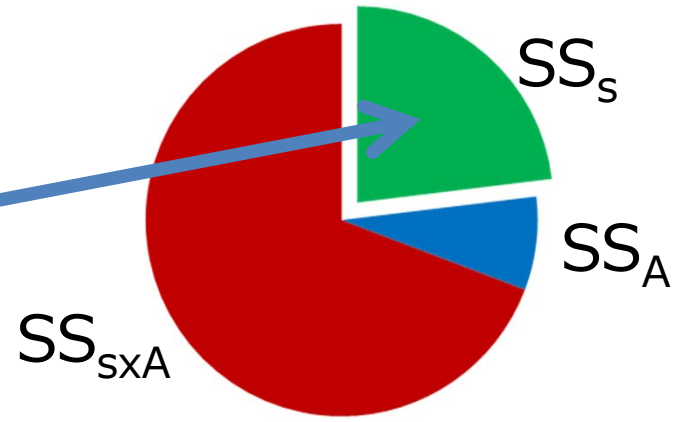
式は同じだが，被験者内計画でのみSS_Sが除外されている

η_p^2を計算すると……

- 一要因の**被験者間**計画



- 一要因の**被験者内**計画



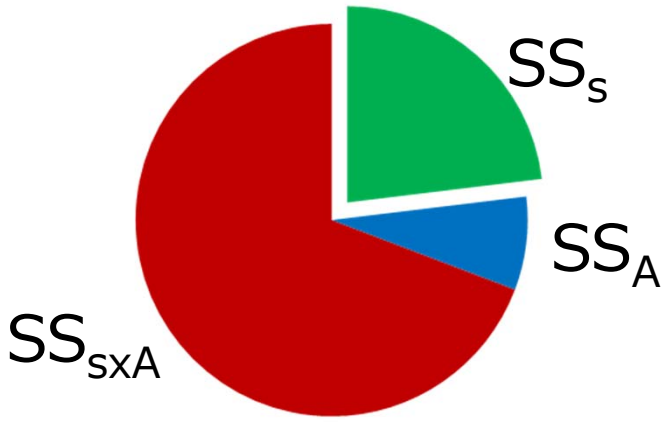
- しかし，被験者間計画でも個人差による変動は分離されていないだけで，SS_eに潜在的に含まれている
  →被験者間と被験者内のη_p^2は比較不能

デザインを通して一般化

- 個人差などの変動を生じる成分を誤差に含めなおす

$$\eta_G^2 = \frac{SS_A}{SS_A + SS_s + SS_e}$$

- 一要因の**被験者内**計画

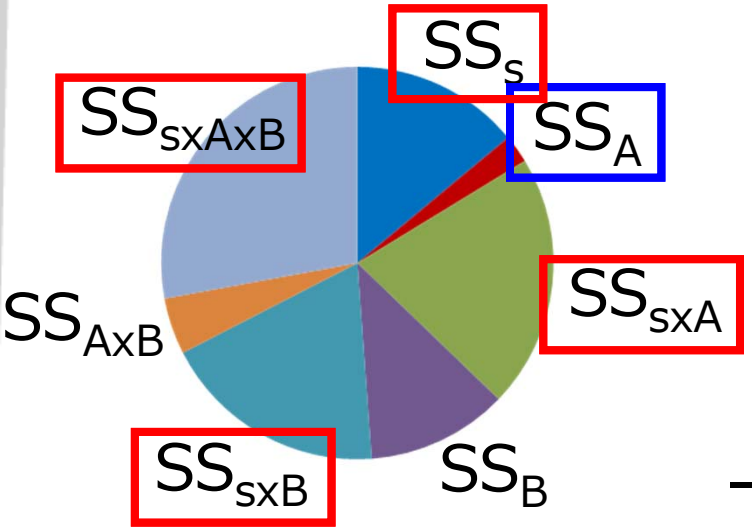


- η_G^2（一般化イータ二乗）の計算では，被験者内計画の場合にSS_Sを分母に加える
→被験者間計画のη_G^2と比較可能

二要因への拡張

ロ二要因の被験者内計画

$$\eta_G^2 = \frac{SS_A}{SS_A + SS_s + SS_{s\times A} + SS_{s\times B} + SS_{s\times A\times B}}$$



- 関心のある効果に加えて個人差による変動をすべて分母に含めていることがわかる例

一般式は少々複雑

$$\eta_G^2 = \frac{SS_{Effect}}{\delta \times SS_{Effect} + \sum SS_{Meas} + \sum SS_k}$$

分散分析の標準化効果量のまとめ

	イータ	オメガ	特　　徴
無印	η^2	ω^2	実験内比較に適している （共通の分母での比較）
偏	η_p^2	ω_p^2	同一デザインの実験間比較にのみ対応
一般化	η_G^2	ω_G^2	実験間比較に適している （デザインの違いを考慮）
	標本 推定値	母集団 推定値	※どの指標も効果量ｆに 変換できる

そのつど目的にあった指標を選び，必要であれば
複数の指標を併用することが重要

効果量が大きければ有意でなくても効果を主張できる？

- たぶん，そんなことはない
  - 効果量が大きければ，検定結果が有意であることが多い
  - 検定が有意だったとしても，効果量が大きいとは限らない

新たな**公刊バイアス**の源となる？

	効果量大	効果量小
有意	理想的	疑問視の的
非有意	検定力不足	問題外

効果量が小さくても効果が重要な場合

□実質的な影響が大きい

- 「メールサービスの障害により，８％のアカウントに影響があった」
  →約400万人のユーザーに問題が生じた
- タイプAパーソナリティの人は，タイプBパーソナリティの人に比べて，約束の時間よりも平均3.85分早く来る（到着時間の分散の約1.5%を説明；$\eta^2 = .015$）
  →従業員1,000名の会社なら3.85分の遅れは年間$140,000，一時間$10の損失
  （Yeaton & Sechrest, 1981）

効果量が小さくても効果が重要な場合

- 間接的・累積的に重大な結果を引き起こす
  - 風評・省エネ効果など
- 理論的に意味がある
  - 理論的にありえないと思われた現象・効果の発見など
- 小さいことを示すことに意味がある
  - 回避不能な現象や副作用の程度を見積もるなど（信頼区間を併用すべきかも）

大きな効果量が小さな意味しか持たないこともある

- ある会社が従業員を平均30秒早く来させる制度を採用したとする（Yeaton & Sechrest, 1981）
  - もともと到着時間の分散は小さい
  - 制度採用前後の比較によって到着時間の多くの分散が説明された（$\eta^2 > .50$）
    →従業員1,000名の会社では，年間$18,000の節約にしかならない
- 従属変数のばらつきと平均差がともに小さいと効果量の値は大きくなる

効果量をどう解釈する？

- 効果量を報告する論文は増えたが，解釈しているものはほとんどない（Fritz et al., 2012）
  - 多くの研究者が頭を悩ませるところ
- 解釈のしやすさは研究の性格にもよる
  - 応用的な研究：何らかの利得と結びつけて解釈しやすい
  - 理論的な研究：研究上の意義と量的な評価が乖離することが多い

非標準化効果量を使う

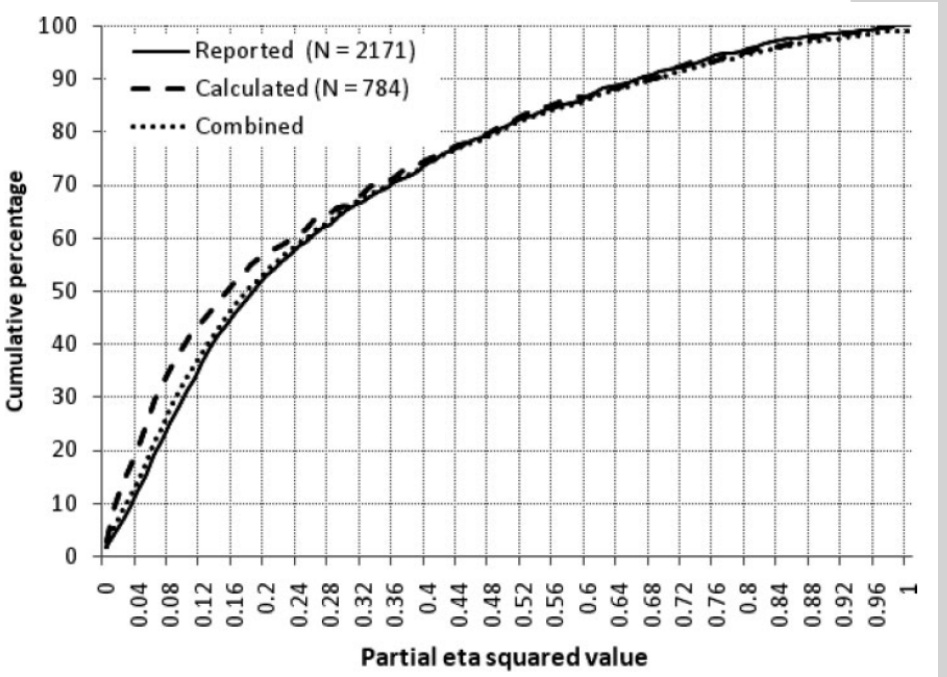
- 測定している単位に意味がある場合は非標準化効果量のほうが解釈しやすい
  - 学習法Aの効果は，$\eta^2 = .40$（統制条件との比較）
  - 学習法Aによって平均 5 点成績が上がった（統制条件との比較）
- 標準化効果量や信頼区間を併記することで，ばらつきの情報を加えられる
  - 指標をひとつに限る必要はない

Cohenのベンチマーク

- 効果量 f に基づく目安
  - f ：小＝.10，中＝.25，大＝.40
  - η_p^2：小＝.01，中＝.05，大＝.14
- 絶対的な基準ではない
  - × criterion，standard
  - ○ benchmark，guideline
- 他に参照するものがどうしても見つからないときの“last resort”（Ellis, 2010）
  - 大きさの目安は，文脈・研究領域によっても異なることが想定されている

データに基づくベンチマーク

- 記憶研究領域の効果量の累積割合（Morris & Fritz, 2013）
  - 2010年の224本の論文からの約3,000のη_p^2に基づく
- Cohenよりも高め
  - 第一四分位＝.08
  - 中　央　値＝.18
  - 第三四分位＝.41



Morris & Fritz（2013）Figure 4

自分の分野のベンチマークがないときは……？

- Morris & Fritz（2013）ほどの規模でなければ，比較的簡単に自分で作れる
  - η_p^2は一般的な分散分析の報告結果から計算できる

$$\eta_p^2 = \frac{df1_{effect} \times F_{Effect}}{df1_{Effect} \times F_{Effect} + df2_{Effect}}$$

  - 自分の研究に近い分野の先行研究を参照すれば，相対的な大きさを比較できる
    - ✓学習法Aは先行研究の学習法Bよりも効果が大きいなどの想定があるときには，より説得的

該当領域の画定の問題

- 自分の研究に近い先行研究がない or 何が近いかわからない場合は？
  - 「新規に開発したトレーニングでとっさの判断能力を向上させる」
  - 「脳の特定部位を電気刺激すると刺激検出課題の成績が上がる」
- どのくらい違いがあれば効果があったといえるのかわからない……
  - 背後にある理論が明確な量的仮定をもたない

“効果”とは何か

- 何が起こったら“効果”があったといえるのかを想定しておくことの重要性
  - 「判断力に５点くらい差があればトレーニングは効果があったといえるだろう」（平均0.001点の差は意味がない）
    - ✓ただし，この例では尺度の意義も明確にしておく必要あり（１点＝どの程度の能力か）
  - 「脳のこの部位が刺激検出に関係しているかどうかを決めよう」（完全に影響がなければ０のはずだと仮定する）

効果の概念の難しさ

- 従属変数は量的であることが多い
  - 刺激検出率は連続的な値をとるので，どの程度の差なら意味があるのかを問える
- 効果は２条件の比較に限定されない
  - ３条件の比較では，どのようなパターンを効果ありと想定するのか
    - ✓線形関係なら簡単だが，そうとは限らない
  - 交互作用の場合には，程度を問えるのか
    - ✓そもそも複数の要因の効果の程度が異なる程度を表していると思われる

大小を比べるだけ？

- 研究上のロジックの拡大
  - 「信頼区間がCohenのベンチマークでいえば小～大の範囲にまたがるのでこの効果は不安定」（Olejnik, 2010）
  - 効果量とその信頼区間を参照して「先行研究 or 類似現象と同程度の効果がある」などの主張ができるのではないか
    - ✓これまでの研究の枠組みでは主張しづらかった「～と同等」「～より大きな効果」などを積極的に検証できるかもしれない

まとめ

- 効果量は既存の統計量と深い関わりがある
  - 標準化効果量だけが効果量ではない
  - サンプルサイズやデザインの影響を受ける
- 効果量を解釈するための努力が必要
  - 実利的な数値や先行研究と照らし合わせる
  - 興味のある現象を反映する効果とは何かを量的に想定する

統計教育にとっての意義

- 検定結果のみに注目しがちな態度を改める方向づけになる
  - 検定だけに着目すると，F値とp値のみに注目しがち
  - 効果量を扱うことで，平方和やサンプルサイズの情報も活用
- 研究で扱おうとしている“効果”について考えなおす機会になる
  - 何が起これば，期待通りなのか，仮説を検証したことになるのか

参考文献

- Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.
- Fay, K., & Boyd, M. J. (2010). Eta-squared. In N. J. Salkind, D. M., Dougherty, & B. Frey (Eds.), *Encyclopedia of research design (vol. 1)*. Thousand Oaks, CA: SAGE Publications. pp. 422-425.
- Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, **141**, 2-18.
- Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*, 137-152.
- Morris, P. E., & Fritz, C. O. (2013). Effect sizes in memory research. *Memory*, **21**, 832-842.
- Olejnik, S. (2010). Omega squared. In N. J. Salkind (Ed.), *Encyclopedia of research design (vol. 2)*. Los Angeles: SAGE Publications. pp. 963-967.
- Yeaton, W. H., & Sechrest, L. (1981). Meaningful measures of effect. *Journal of Consulting and Clinical Psychology*, **49**, 766-767.

※発表時の資料を若干修正しています。ご意見を下さいました先生方に心よりお礼申し上げます。