日本パーソナリティ心理学会
アフター・カンファレンス企画（統計セミナー）
（日本社会心理学会共催）

2016/9/16

認知心理学研究における サンプルサイズ設計

井関 龍太
(大正大学)

サンプルサイズ設計の重要性

- 認知心理学
  - 実験研究が中心（→以下，実験を想定）
- 事前の設計の意義
  - 計画の見通しが立つ（設備・労力・資金）
  - 実験協力の負担を減らす（人類にやさしい？）
  - ミスコンダクトを防ぐ（科学的に正しい）
    →実験研究においても，他の形態の研究と変わらず有益

認知心理学における現状

- 認知心理学者はサンプルサイズ設計しているか
  - 投稿案内に事前の設計を行うよう明記してあるジャーナルでも，掲載論文のほとんどには何の記載もない
  - たまに“Cohenのガイドラインにしたがって G*Powerで計算した”と書いてある
    - しかし，どの効果（主効果・交互作用）を検出しようとしているのかといった，肝心の情報が述べられていない（効果量への言及すらないことも）

サンプルサイズ設計したものの……

- “検定力分析で10人が必要と出た”
  - 実際に募集したら９人しか集まらなかった（あるいは，１人欠席・除外となった）
    →何がなんでもあと一名追加すべき？
- 現実的には，よく起こる状況
  - 実験設備・施設の使用期間
  - 実験者・参加者へ支払う謝金
    →検定力分析の結果は，これらの現実的制約を乗り越えてでも達成すべき基準なのか

ぴったりでなくていいじゃないか確率だもの

- 無理矢理数値を合わせることにあまり意味はないのでは？
  - G*Powerの整数の出力はミスリーディング
- 以下のような報告の仕方があってもよいのでは？
  - “.95の検定力を得るためには，18名のサンプルが必要であるとの事前の分析の結果が得られた。この結果に基づき募集したところ，実際に参加したのは17名であった。17名の検定力を計算したところ，.92であった。”

検定力分析≠サンプルサイズ設計

- サンプルサイズ設計の基盤を統計的な手法（検定力分析）に限定する必要はないはず
  - 実験デザイン（カウンターバランスしている場合は，その倍数の人数が必要）
  - 最低ラインのコンセンサス（検定力は十分でも，N = 3では通らなそう）
  - 現実的制約（いくら重要な研究でも5,000人以上を対象に実験することは難しい）

何のための設計か

- 統計改革
  - 再現性を高め，ミスコンダクトを防ぐために，機械的な帰無仮説検定の適用をするのはやめよう
    - 検定力分析の推奨もその一環と捉えられる
- サンプルサイズ設計
  - 検定力分析の結果だけを見て機械的に決めるのでは，統計改革以前に逆戻り
  - より柔軟かつ多面的であるべき

柔軟なサンプルサイズ設計の例（１）

- “先行研究の2.4倍にしてみた”

A total of twenty-four observers took part in this experiment. Fourteen observers were tested (seven males), mean age 30 (18–52 years old) with rising–falling patch quality pattern. Ten observers were tested (five males), mean age 32 (18–52 years old) for falling–rising patch quality pattern. In the absence of an obvious way to perform a power calculation (i.e. unknown effect size), we used a relatively large number of observers. Given that Wolfe (2013) used 10 observers for the basic version of a foraging experiment, we used 2.4× that number to look for a modulation of that basic foraging behavior. All observers gave informed consent approved by Brigham and Women's Hospital and consistent with the Declaration of Helsinki. All were paid $10/h for their time. All had vision corrected to at least 20/25 and passed the Ishihara color vision screen.

Zhang et al. (2015)

柔軟なサンプルサイズ設計の例（２）

- “各条件の各刺激を３人ずつが評価する”

for a $5 Starbucks gift card. We then recruited 162 people (mean age = 36.86 years, *SD* = 15.01; 80 males) visiting the Museum of Science and Industry in Chicago to evaluate the candidates in exchange for a food item. The sample size for evaluators was predetermined by our goal to have 3 people evaluate each of the 18 candidates in each of the three conditions (i.e., slightly more than 50 evaluators per condition). We did not know what effect size to predict in this first experiment, but we arrived at this number because it was feasible at our laboratory, offered multiple evaluators for each target, and was our best guess of the sample size needed to detect an effect of interest. Fifty participants per condition yields 80% power to detect a medium-sized effect.

Schroeder & Epley (2015)

柔軟なサンプルサイズ設計の例（３）

- “うちのラボではこのくらいがよかった”

for a $5 Starbucks gift card. We then recruited 162 people (mean age = 36.86 years, *SD* = 15.01; 80 males) visiting the Museum of Science and Industry in Chicago to evaluate the candidates in exchange for a food item. The sample size for evaluators was predetermined by our goal to have 3 people evaluate each of the 18 candidates in each of the three conditions (i.e., slightly more than 50 evaluators per condition). We did not know what effect size to predict in this first experiment, but we arrived at this number because it was feasible at our laboratory, offered multiple evaluators for each target, and was our best guess of the sample size needed to detect an effect of interest. Fifty participants per condition yields 80% power to detect a medium-sized effect.

Schroeder & Epley (2015)

これでいいのだ

- 一見，いい加減なように見える
  - しかし，これこそが心理学者が実際に行なっていることを反映しているのでは？
    →比較考量のプロセスが述べられている
- 定型的な検定力分析以外の要因
  - 経験知
    - 「だいたい一群20人にしておけばいい」
      →理論的根拠はなくても，経験的には支持されるかもしれない
  - 実験デザインへの適合

経験を蓄積することの重要性

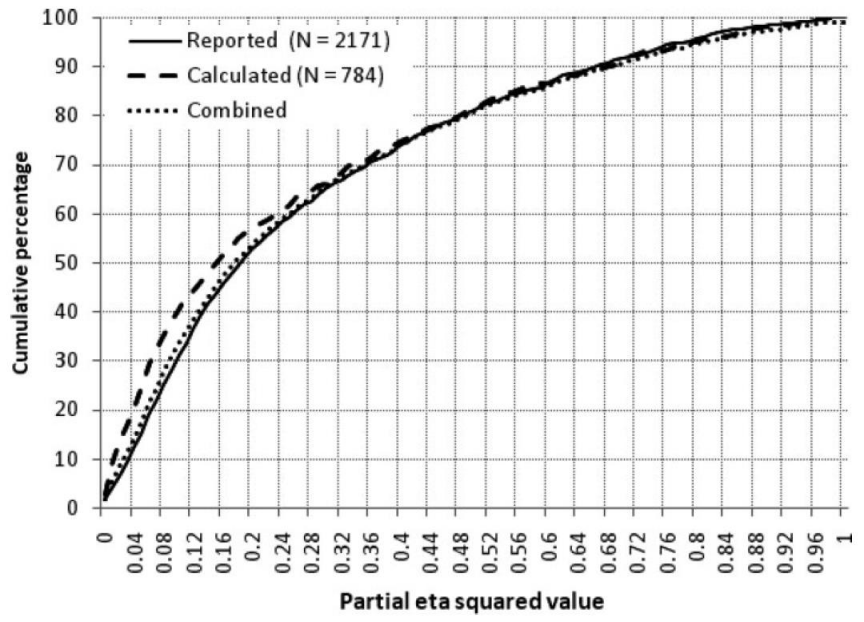
- 研究者は経験に基づく直感を持っている
  - 経験を共有し，説得力のある方法へと洗練させることはできないか？

たしかに，研究者の頭の中には「○○のタイプの研究であれば，サンプルサイズは△△くらいが適当であろう」という考えがあるように思う。このような「経験則」については，調査してみたいところである。具体的には，架空の研究例を分野別に各種呈示して，適当だと思うサンプルサイズを回答させ，その平均をとる，さらにそれを回答者の専門分野別に検討する，という研究である（杉澤〈1999〉[38]が参考になろう）。

村井（2006）

データに基づく領域ごとの効果量

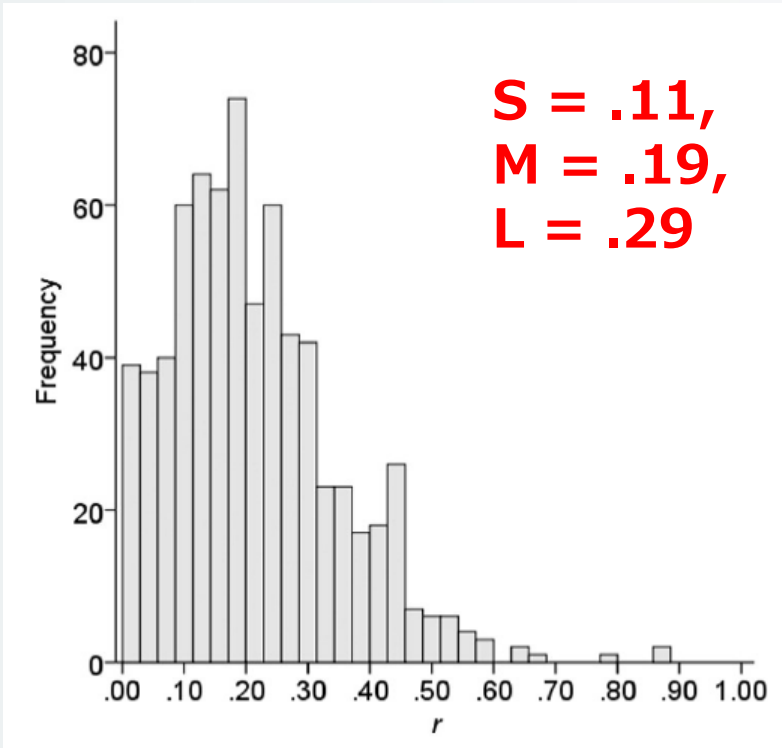
- 記憶研究（η_p^2）



S = .08, M = .18, L = .41

Morris & Fritz (2013) Fig. 4.

- 個人差研究（r）



Gignac & Szodorai (2016) Fig. 1.

経験知への２つのアプローチ

- メタ分析に基づく効果量の累積頻度（ボトムアップの経験知）
  - 分野の現状を表している
  - 実用的な効果の大きさを表すとは限らない
- 研究者の直感についてのメタ研究（トップダウンの経験知）
  - 分野を超えた広がりを持つ可能性
  - 経験に基づくバイアスの両面的価値
  - 臨床的効果の丁度可知差異

実験デザインを反映させた検定力分析

- 分散分析
  - 多くの実験研究が典型的に使用する方法
  - 検定力分析にとっては頭の痛い対象
- 分散分析は，複雑な実験デザインを伴うことが多い
  - 参加者間・参加者内
  - 主効果・交互作用
  - カウンターバランス
  - 同一条件の繰り返し　etc.

→
- いずれも検定力に影響
- 一般的な検定力分析には反映されない

条件の繰り返し（replicate）

- 実験研究では，同じ条件を 2 回以上測定することが多い
  - まったく同じ刺激と条件を繰り返す
  - 刺激を変えて同じ条件を複数回測定する

等

繰り返しに対する一般的な扱い

- 繰り返して集めたデータを個人ごとに平均する
  - ストループ条件の30試行の反応時間
  - 精緻化学習条件の10種類の画像に対する確信度評定

    など
- 平均の根拠
  - 個々の反応はノイズや外れ値の影響が大きいかもしれないが，平均では相対的に小さい
  - 個々の反応は正規分布していなくても，中心極限定理により平均値は正規分布に近づく

繰り返しのデータを平均することの問題

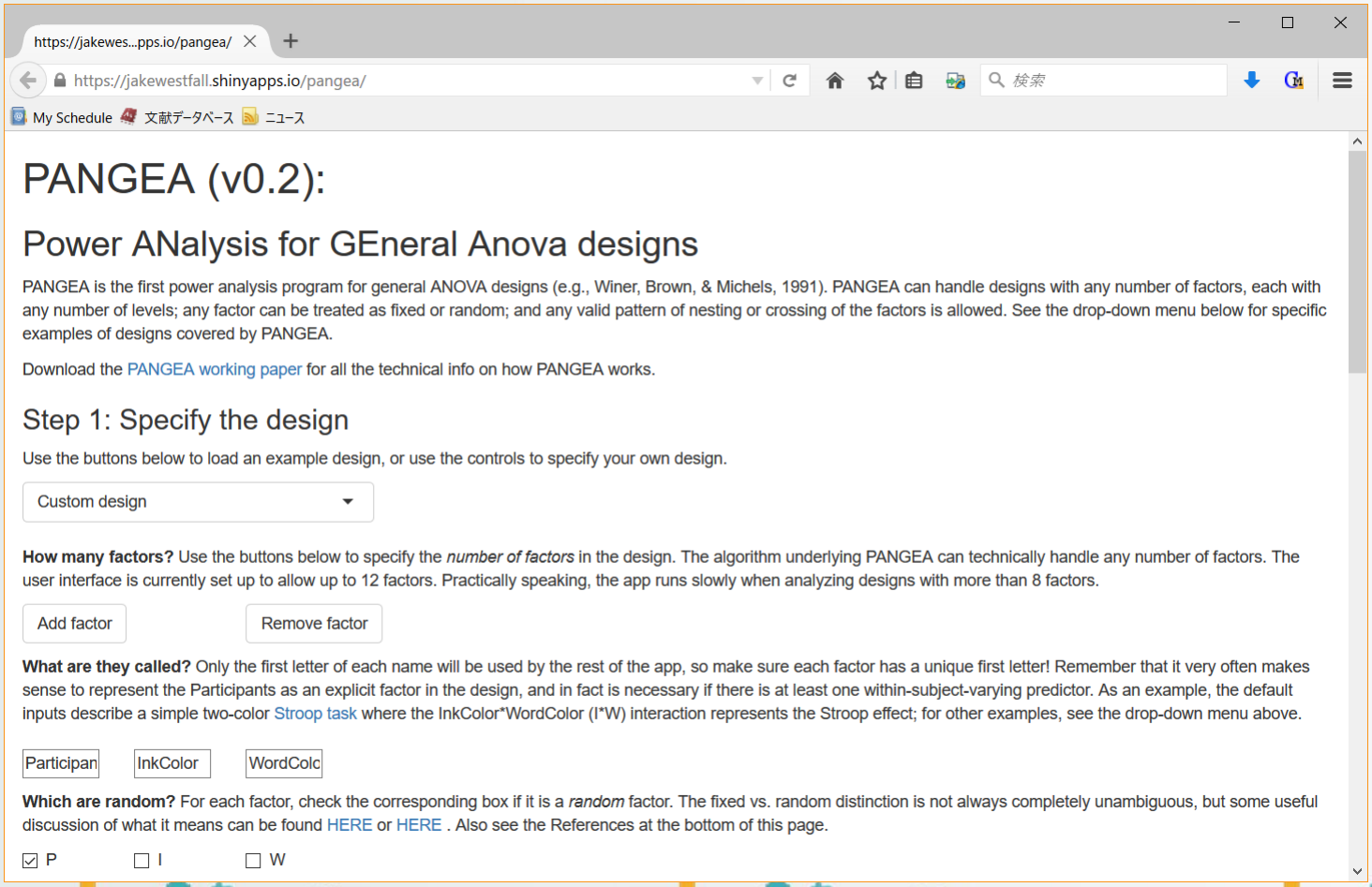
- 平均するとばらつきの情報が失われる
  - 分散が大きいデータも小さいデータも同じ平均として分析される
    →ノイズや外れ値の生じやすさ／生じにくさ（安定性）がむしろ反映されない
- 正規分布するのは特定の繰り返しの結果得られた平均
  - たとえば，30試行の反応時間の平均値は正規分布するかもしれないが，それに基づく推定では30試行の平均についての結論しか下せない（本当に興味ある対象は？）

一般的な分散分析デザインへの対応

- PANGEA (Power ANalysis for GEneral Anova designs)：あらゆる分散分析デザインを扱える検定力分析の枠組み（Westfall, in prep.）
  - 複雑な多要因デザインに対応（現在，12要因まで指定可能）
  - 繰り返しの問題を考慮
  - ウェブアプリ（https://jakewestfall.shinyapps.io/pangea/）
  - Rのソースコードもダウンロード可能

分析ツールとしてのPANGEA

ウェブアプリにアクセス



くわしくは……

- 村井潤一郎・橋本貴充 (編)
  心理学のためのサンプルサイズ設計入門
  講談社
  2017年２月頃刊行（予定）

まとめ

- 認知心理学者のサンプルサイズ設計
  - 伝統保守主義
  - G*Power絶対主義
    - 機械的な検定力分析の適用がサンプルサイズ設計ではないはず
- 柔軟で多面的なサンプルサイズ設計のために
  - 経験知の蓄積と活用
  - 実験デザインを反映させた検定力分析