再現可能性問題に対する諸関係領域の動向

大久保街亜
専修大学人間科学部
専修大学社会知性開発研究センター・
心理科学研究センター

あらまし

- 心理学における再現可能性
- 諸関連領域では再現可能性を高めるための動きがある。
- 日本の社会心理学ではどうか？
- 再現可能性を高めるための具体策は？

- 例数設計と停止規則に焦点

Matt Motylのはなし (Nuzzo, 2014; Nosek et al., 2012)

- Matt Motyl ヴァージニア大の大学院生
- 研究1
  - 1979人の参加者
  - 政治的中道派は，右派や左派より明るさを正確に見分けられる ($p < .01$)。
- “The hypothesis was sexy and the data provided clear support (Nuzzo, 2014, p 150). ”
  - Psychological Science? Nature? Science?

Matt Motylのはなし (Nuzzo, 2014; Nosek et al., 2012)

- 再現性を重視し追試（研究2）
- 1300 人，検定力 = .995
- $p = .59$

心理学における再現可能性

- あまり高くない。
- 理由
  - 測定ノイズ
  - 手続き・対象の不一致
  - ねつ造
  - 統計の誤用，知識不足
    - P-hacking (Simmons et al., 2011)
    - 大きすぎる標本サイズ（例，Matt Montyl）

Data detective (Simonsohn, 2013)

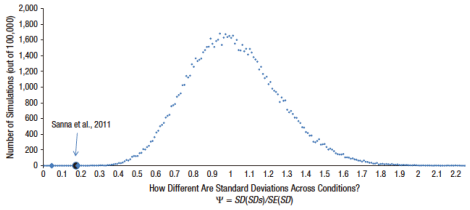
- 平均値と標準偏差からデータのねつ造を検知

Table 1
Charitable contributions, helping, compassion, and cooperating and moods by physical (vertical) height.

Study/measure	Physical (vertical) height		
	High	Low	Control
Study 1			
Proportion contributing	.16 (59/368)	.07 (26/391)	.11 (37/350)
Study 2			
Mean time helping (minutes)	11.36 (2.82)	6.77 (2.75)	8.74 (2.96)
Study 3			
Mean compassion (hot sauce grams)	39.74 (25.09)	85.74 (24.58)	65.73 (25.65)
Study 4			
Mean cooperating (fish returned)	32.93 (9.24)	20.60 (9.54)	23.66 (9.82)
Mean moods	5.70 (1.13)	5.46 (1.19)	5.59 (1.11)

Note. Proportions rounded to nearest decimal with numbers contributing and totals in parentheses for Study 1; standard deviations in parentheses for Studies 2–4.

シミュレーション結果
―そんなことまずない―



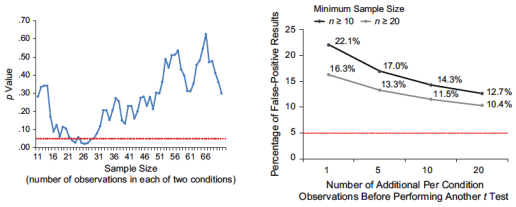
ねつ造対策

- Data detective
- 研究情報の公開
  - ローデータの提出（出版後のデータベース化）
  - データの共有
  - 実験手続きの共有
  - 論文のオープンアクセス化
- 倫理教育の徹底

擬陽性とp-hacking
(Simmons, Nelson, & Simonsohn, 2011)

- 「N増し」と擬陽性の関係
- 非意図的なミスコンダクト

N増しと擬陽性の関係
(Simmons et al. 2011)



ただし，実際の研究場面では擬陽性の確率はもっと減る(Murayama et al., in press)

p-hacking

- 望んだ結果が出るまでデータを取ること
  - 例：*N*増し
  - 無自覚に行われることもある。
- 基本的な問題
  - 多重の検定による有意水準の上昇
  - 1回の検定：α= 1 - .95=.05
  - 5回の検定: α= 1 - (.95)*(.95)*(.95)*(.95)*(.95) = .23

大きすぎる標本サイズの問題

- 標本サイズを増やせば，どこかで必ず有意になる。
- The effects of A and B are always different—in some decimal place—for any A and B. Thus asking 'are the effects different?' is foolish (Tukey, 1991, p. 100)
  - 必ず違いはある。

高すぎる精度の問題

- 非常にわずかな効果でも検出されてしまう。
  - 心理学は測定ノイズが大きくしかも多種多様
    →危険
  - (一般論から言って)効果量の軽視は問題

停止規則を決め例数設計をしよう

- p-hacking (Simmons et al., 2011)
- 大きすぎる標本サイズ（例，Matt Montyl）
  ↓
- 適切な停止規則と例数設計とで避けられる。

停止規則と例数設計

- 停止規則：データ取得を停止する事前規則
- 例数設計：信頼できるデータを得るために必要な標本サイズの推定

基本的な例数設計

- 検定力を基準にしたもの
- 信頼区間を基準にしたもの
- 適応的基準を用いるもの

基本的な例数設計

- 検定力を基準にしたもの
- 信頼区間を基準にしたもの
- 適応的基準を用いるもの

検定力とは？

- 差があるときに，あると言える確率
  - 「有意差を見つける力」
  - 「第2種の誤りをおかさない確率」

研究結果	真の結果 差なし	真の結果 差あり
有意差なし	正しい判断(1-α)	第2種の過誤(β)
有意差あり	第1種の過誤(α)	正しい判断(1-β) 検定力

研究場面における検定力

- 検定力，効果量，有意水準，標本サイズ
  - 相互に影響し合う
- *Power* = *f*(*ES*, α, *N*)
- 標本サイズのみ能動的にコントロール可能
  - 他は研究者が決めることが困難

適切な検定力の重要性

- 低すぎる検定力
  - 真の差や効果を見逃す可能性
- 高すぎる検定力
  - 参加者に負担→倫理的な問題
  - 労力，資金，時間に負担→経済的な問題
  - （心理学では）大きすぎる標本サイズの弊害も

適切な検定力

- 5-80ルール （Cohen, 1988）
  - αはβの4倍の慎重さが必要
  - α=.05 β= .20 1-β= .80
- 第2種の過誤が多大な影響をもたらすときは厳しくするべき（例，環境問題，人命に関わる問題）
  - 検定力 .95にすべき場合もある

検定力に基づく例数設計

- 効果量＝先行研究から推定
  - わからないときは中程度 (Cohen, 1962, 1994)
  - Hedge's *g*, Cohen' s *d* = .50
- 有意水準は固定＝.05
- 検定力＝通常は.80

  ↓
- 必要な標本サイズ決定

基本的な例数設計

- 検定力を基準にしたもの
- 信頼区間を基準にしたもの
- 適応的基準を用いるもの

信頼区間

- 信頼区間＝ある確率で（母数の）代表値が存在する区間
  - 例：平均値の95%信頼区間＝95%の確率（100個の標本についてうち95回）で母平均が存在する区間

$$95\%CI = M \pm z_{95\%} \times SE$$

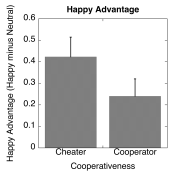
- 指標によって計算方法が異なる。詳しくは

大久保・岡田(2012) 伝えるための心理統計：効果量，信頼区間，検定力

- 心理学における統計改革
  - 帰無仮説検定に偏ったデータ解析の是正
- 効果量・信頼区間・検定力

信頼区間に基づく例数設計（正確度分析）

- 信頼区間が設定した値になるよう標本サイズを決定する。
- 測定値そのものが重要なときに用いられる。
  - 緩解率，有病率，視聴率，内閣支持率



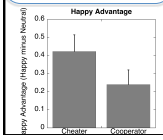
基本的な考えかた

$$95\%CI = M \pm z_{95\%} \times SE$$

$$95\%CI = M \pm z_{95\%} \times \frac{s}{\sqrt{N}}$$

誤差範囲(*ME*)を決めると必要な*N*が求まる

$$ME = z_{95\%} \times \frac{s}{\sqrt{N}}$$



$$N = \left(\frac{z_{95\%} \times s}{ME} \right)^2$$

適応的基準を用いるもの

- 事前に設定した規則をデータに当てはめ，データ取得の停止を決める。
- COAST(composite open adaptive stopping rule, Frick, 1998)
- CLAST (composite limited adaptive stopping rule; Botella et al., 2006)
- Variable-criteria sequential stopping rule (Fitts, 2010)

COASTを用いたデータ取得と停止

1. 少数標本（2-3）で有意性検定
   - *I.* $p < .01$（下限基準）なら，α＝.05で有意と判断。データ取得停止
   - *II.* $p > .36$（上限基準）なら帰無仮説を採択。データ取得停止
   - III. $.01 < p < .36$ ならデータ取得を継続
2. 事前に決めた*N*の増分ごとに 1. のプロセスを繰り返す。

比較

- 検定力を基準にしたもの
  - 心理学では多い
  - 標本サイズが大きくなりがち
- 信頼区間を基準にしたもの
  - 医学や応用分野で多い
  - 標本サイズが大きくなりがち
- 適応的基準を用いるもの
  - 動物実験などで多い
  - 基準について批判もある

例数設計は関連領域では不可欠

- 疫学（観察研究の指針）
  - STrengthening the Reporting of OBservational studies in Epidemiology(STROBE)
- 医学（医学における効果研究）
  - Consolidated Standards of Reporting Trials (CONSORT)
- 医学では例数設計を含んだ情報を事前登録しなければ論文化できないこともある（臨床試験登録）
  - UMIN Clinical Trials Registry
  - ClinicalTrials.gov

ほかにも

- 生命科学全般，実験心理学でも例数設計は当然のことになりつつある。
- そもそも。。。

APA Publication Manual 6th edn. (p. 30)

Sample size, power, and precision. Along with the description of subjects, give the intended size of the sample and number of individuals meant to be in each condition, if separate conditions were used.

- 参加者の記述ともに想定した標本サイズについて書け。

State how this intended sample size was determined (e.g., analysis of power or precision). If interim analysis and stopping rules were used to modify the desired sample size, describe the methodology and results.

- 想定した標本サイズをどのように決めたか述べよ（検定力分析，正確度分析）。中間分析や停止規則を適用し，想定した標本サイズを変更したなら，その方法と結果を記述せよ。

Nature Editorial (2013,463)

事前に設定した効果量をもとにどのようにサンプルサイズ決めたか報告を求める

ANNOUNCEMENT

Reducing our irreproducibility

Over the past year, *Nature* has published a string of articles that highlight failures in the reliability and reproducibility of published research (collected and freely available at go.nature.com/huhbyr). The problems arise in laboratories, but journals such as this one compound them when they fail to exert sufficient scrutiny over the results that they publish, and when they do not publish enough information for other researchers to assess results properly.

From next month, *Nature* and the Nature research journals will introduce editorial measures to address the problem by improving the consistency and quality of reporting in life-sciences articles. To ease the interpretation and improve the reliability of published results we will more systematically ensure that key methodological details are reported, and we will give more space to methods sections. We will examine statistics more closely and encourage authors to be transparent, for example by including their raw data.

Central to this initiative is a checklist intended to prompt authors to disclose technical and statistical information in their submissions, and to encourage referees to consider aspects important for research reproducibility (go.nature.com/vlokqp). It was developed after discussions with researchers on the problems that lead to irreproducibility, including workshops organized last year by US National Institutes of Health (NIH) institutes. It also draws on published concerns about reporting standards (or the lack of them) and the collective experience of editors at Nature journals.

The checklist is not exhaustive. It focuses on a few experimental and analytical design elements that are crucial for the interpretation of research results but are often reported incompletely. For example, authors will need to describe methodological parameters

we will commission statisticians as consultants on certain papers, at the editor's discretion and at the referees' suggestion.

We recognize that there is no single way to conduct an experimental study. Exploratory investigations cannot be done with the same level of statistical rigour as hypothesis-testing studies. Few academic laboratories have the means to perform the level of validation required, for example, to translate a finding from the laboratory to the clinic. However, that should not stand in the way of a full report of how a study was designed, conducted and analysed that will allow reviewers and readers to adequately interpret and build on the results.

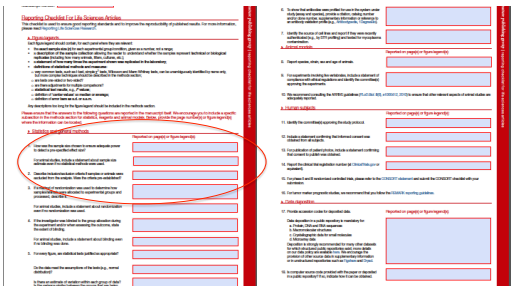
To allow authors to describe their experimental design and methods in as much detail as necessary, the participating journals, including *Nature*, will abolish space restrictions on the methods section.

To further increase transparency, we will encourage authors to provide tables of the data behind graphs and figures. This builds on our established data-deposition policy for specific experiments and large data sets. The source data will be made available directly from the figure legend, for easy access. We continue to encourage authors to share detailed methods and reagent descriptions by depositing protocols in Protocol Exchange (www.nature.com/protocolexchange), an open resource linked from the primary paper.

Renewed attention to reporting and transparency is a small step. Much bigger underlying issues contribute to the problem, and are beyond the reach of journals alone. Too few biologists receive adequate training in statistics and other quantitative aspects of their subject. Mentoring of young scientists on matters of rigour and transparency is inconsistent at best. In academia, the ever increasing pressures to publish and chase funds provide little incentive to pursue studies and publish results that contradict or confirm previous papers. Those who document the validity or irreproducibility of a published piece of work seldom get a welcome from journals and funders, even as money and effort are wasted on false assumptions.

Tackling these issues is a long-term endeavour that will require

Reporting check lists for life science articles



Reporting check lists for life science articles: Statistics and general Methods

1. How was the sample size chosen to ensure adequate power to detect a pre-specified effect size?

   For animal studies, include a statement about sample size estimate even if no statistical methods were used.

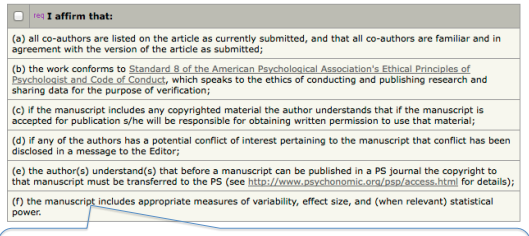
2. Describe inclusion/exclusion criteria if samples or animals were excluded from the analysis. Were the criteria pre-established?

New Statistical Guidelines for Journals of the Psychonomic Society (8/23/2012)

- 検定力を考慮せよ。また，どのように標本サイズを決めたか報告せよ。
- 検定の繰り返しは重大な過誤をもたらす。
- データを選択して報告するべからず。
- 豊富な記述はデータの理解を助ける。多面的な指標を用いよ。
  - などなど。

実験心理学における評価の高い論文誌を発行

投稿時に明示的に宣言

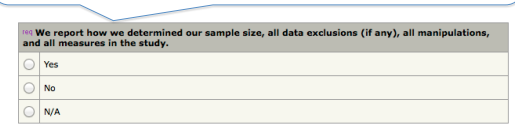


The manuscript includes appropriate measures of variability, effect size, and (when relevant) statistical power.

Psychonomic Society 以外でも

- 論文誌 ”Cognition and Emotion”

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.



実際の論文でも

- Psychological Science
- Psychonomic Bulletin and Review
- Journal of Experimental Psychology

- 標本サイズの決定方法について記述が増えてきた。

ただし，検定力の低い研究も多い

- 実際の研究場面では，まだ，充分に考慮されていない。
  - 日本では最近の論文でも(心理学研究2007–2008年)，半数は適切な範囲に達していない（鈴川・豊田,2012）。
  - 海外でもあまり違いはない。
- 心理学における例数設計の徹底は不十分なのかもしれない。

日本の社会心理学の現状 —「社会心理学研究」を対象に—

- 論文誌「社会心理学研究」に掲載された研究を対象に現状を検討
- 方法
  - 2006年から2013年までに「社会心理学研究」に掲載された論文
  - 対応のない*t*検定を行った101の検定結果を対象

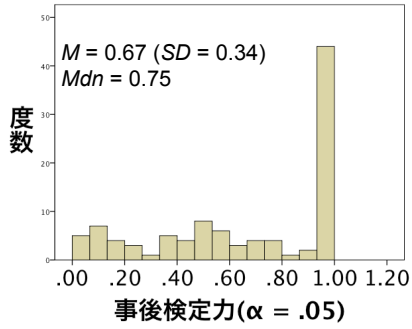
今回の例数設計

- 相関係数の検定について，検定力 .80 で中程度の効果量を想定し，81 が適切な標本サイズ。
- 論文中の記述不足や記述間違いによる20 程度の除外データが予想された。
  - 実際は記述から推定した値を入力。
  - 除外データは無し。

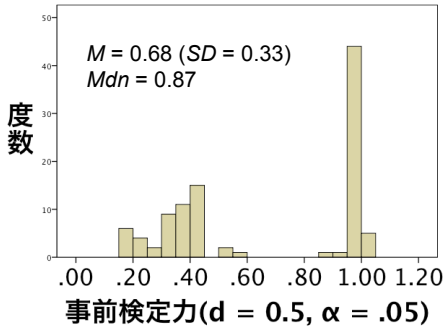
「社会心理学研究」における例数設計の記述

- 0/101 ＝ 0.0
- まったくなし。

事後検定力の分布



中程度の効果量を想定した事前検定力の分布



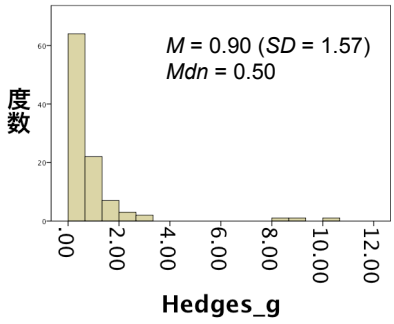
効果量の基準（Cohen, 1988）

	効果量の強さ		
	小	中	大
Hedges's *g*	.20	.50	.80
Pearson's *r*	.10	.30	.50

$$g = \frac{M_1 - M_2}{S_p}$$

プールした標準偏差

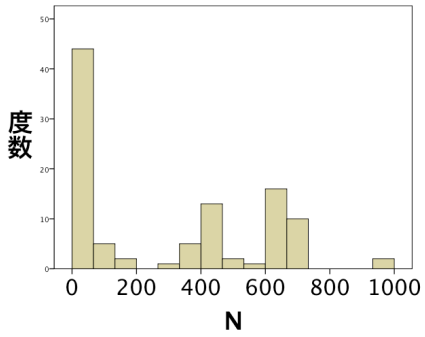
効果量（Hedges's *g*）の分布



ふた山の事前検定力分布

- 中程度の効果量を想定した場合，多くの研究は高すぎるか，低すぎる。
- 適切な検定力（.80）の周辺にほとんどデータがない。
- 適切な例数設計がなされていない。

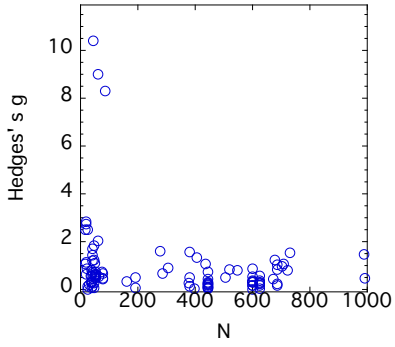
*N*の分布



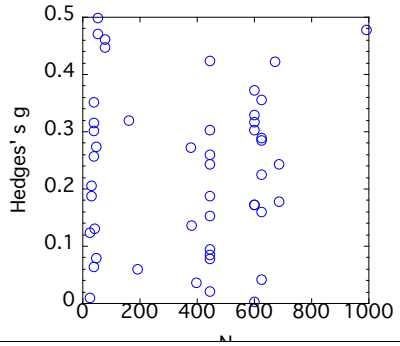
対応のない *t* 検定に必要な標本サイズ (power = .80)

	d = 0.2（大）	d = 0.5 （中）	d = 0.8 (小)
対応なし（両群）	788	128	52
対応なし（一群）	394	64	26

効果量と標本サイズの関係



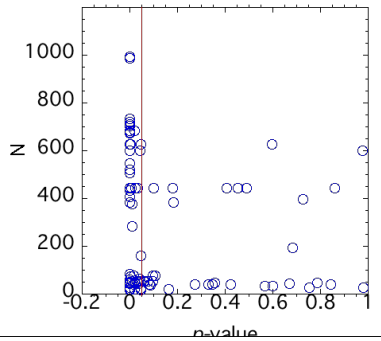
効果量と標本サイズの関係 (g < 0.5)



過度に大きい標本サイズの問題

- 高すぎる検定力
  - 資源のムダ（人的，時間的，金銭的）
  - 倫理的な問題（参加者に不必要な負担）
- （心理学では）大きすぎる標本サイズの弊害
  - 有意になりやすい
  - 効果量の軽視

標本サイズとp値の関係（全体）



標本サイズと有意差

- 標本サイズが大きいと有意になりやすい。
- $p < .05$ で過度に大きな標本サイズがある。
- $p < .05$ でp ≥.05より標本サイズが大きい

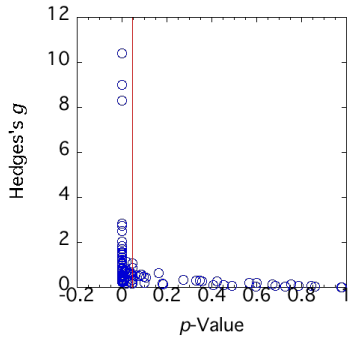
→p-hacking and/or 大きすぎる標本サイズ

p	N	平均値	標準偏差	中央値
< .05	69	362.1594	291.24972	435
>= . 05	32	180.0313	203.34152	49

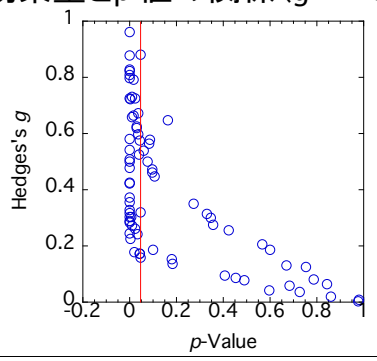
標本サイズと有意差

- 効果量の軽視
  - 標本サイズが多ければ必ずいつか有意になる
- 差や効果の大きさを考慮していない。
- *p*値のみが一人歩きをするかもしれない

効果量とp値の関係（全体）



効果量とp 値の関係（$g < 1.00$）

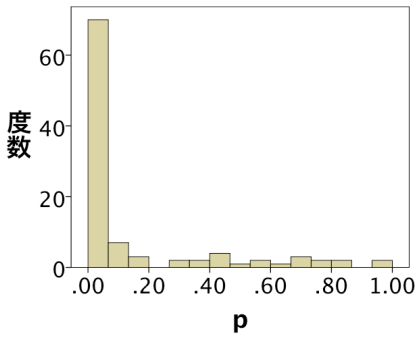


効果量と*p*値の乖離

- 中程度から小さな効果量→有意になったりならなかったり
- 小さな効果量でも有意

- 大きすぎる標本サイズ
  - 「意味のある」効果や差ではない？
  - 望んだ（検定）結果を望んだ結果？
  - p-hacking and/or 大きすぎる標本サイズ

p 値の分布



p < .05に偏った報告

- File drawer problem
  - 有意な結果しか報告されない。
  - 再現できないと報告されない。

- 見かけの再現性が高くなる。
- 事実により知識が（適切に）更新されない。

- メタ分析に影響

日本の社会心理学の現状

- 例数設計がなされていない。
  - 検定力を考慮していない。
  - 効果量も考慮していない。
- 標本サイズが大きすぎる（ものがある）。
- p-hackingの可能性。
- 有意な結果を選択的に報告。

- 再現可能性を高める改革が必要。

停止規則を決め例数設計をしよう

- p-hacking (Simmons et al., 2011)
- 大きすぎる標本サイズ（例，Matt Montyl）
  ↓
- 適切な停止規則と例数設計とで避けられる。

検定力を使った例数設計

- *Power* = *f*(*ES*, α, *N*)
  - 検定力，効果量，有意水準，標本サイズ
- 先行研究から効果量を推定
- 事前に定めたαとβに基づき標本サイズを算出。

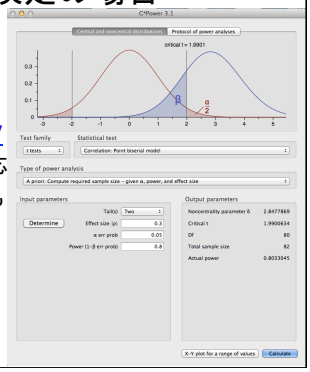
- 簡便には表を見る
  - 「伝えるための心理統計」

表を見て決める　表5.2 p. 154

	検定力				検定力		
効果量＝d	.70	.80	.90	効果量＝r	.70	.80	.90
.10	2471	3142	4205	.05	2467	3137	4198
.20	620	787	1053	.10	616	782	1046
.30	277	351	469	.15	273	346	462
.40	157	199	265	.20	153	193	258
.50	101	128	171	.25	97	123	164
.60	71	90	119	.30	67	84	112
.70	53	67	88	.35	49	61	81
.80	41	52	68	.40	37	46	61
.90	33	41	54	.45	29	36	47
1.00	27	34	45	.50	23	29	37

複雑な検定の場合

- Rで
- G*Power で
  - www.gpower.hhu.de/
  - さまざまな検定に対応
  - MacにもWindowsにも
  - 操作も簡単
  - 検定力，効果量，標本サイズを簡単に



適切な検定力と膨大な*N*

- Power = .95を実現する標本サイズは膨大。
  - 中程度の効果量で対応のないt検定＝210人
- “Typically people are astonished at how many participants are required to achieve decent power, say 0.95. A typical reaction is to reduce the power to maybe 0.90, then finding the number of participants still too much, reducing further to .80, and settling for that (Dienes, 2008, p.64)”.
- 210→172→128

適応的基準を用いる

- 事前に設定した規則をデータに当てはめ，データ取得の停止を決める。
- COAST(composite open adaptive stopping rule, Frick, 1998)
- CLAST (composite limited adaptive stopping rule; Botella et al., 2006)
- Variable-criteria sequential stopping rule (Fitts, 2010)

信頼区間を使った例数設計（一例）

- 事前に信頼区間における誤差範囲を設定し，それに基づき標本サイズを決定
- 20 ms の誤差範囲で，*SD* = 60を想定し，平均反応時間を求める。

$$N = \left(\frac{z_{95\%} \times s}{ME} \right)^2$$

$$N = \left(\frac{1.96 \times 60}{20} \right)^2 = 34.5744 \cong 35$$

論文中の記述に必要な情報

- 標本サイズを決めた手段（e.g., 検定力，信頼区間，適応的停止規則）
- 基準
  - 検定力：検定力，効果量，有意水準（共変量などもデザインによっては必要）
  - 信頼区間：誤差，標準偏差
  - 停止規則：その内容
- 想定される標本サイズ

Barber and Mother (2013, Psychological Science, 24, p. 2424)

- The sample size used in these experiments was based on an a priori power analysis conducted in G*Power 3.1. Assuming an effect size of Cohen's d = 0.79—derived from the following previously published studies:(中略) —a significance level of α = .05, four participant groups, and one covariate (baseline task performance), we determined that a total sample size of 52 participants (n = 13 per group) would provide 80% power to detect effects. To exceed this criterion and achieve greater than 80% power, we recruited 56 participants (n = 14 per group).

まとめ

- 再現可能性を高めるには適切な例数設計が必要
- 諸関連領域では不可欠な手続きになりつつある
- 日本の社会心理学では，適切な例数設計はなされていない
- これからきちんと例数設計をして，再現可能性を高めていこう。

引用文献

- American Psychological Association (2009). *Publication manual of the American Psychological Association (6th edn)*.. Washington, DC: American Psychological Association.
- Barber, S. J., & Mather, M. (2013). Stereotype threat can enhance, as well as impair, older adults' memory. *Psychological Science, 24,* 2522-2529.
- Botella, J., Ximénez, C., Revuelta, J., & Suero, M. (2006). Optimization of sample size in controlled experiments: The CLAST rule. *Behavior Research Methods, Instruments, & Computers, 38,* 65-76.
- Cohen, J. (1988). Statistical power analysis for the behavioral sciences(2nd ed.). Hillsdale, NJ: Erlbaum.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145-153.
- Cohen, J. (1994). The earth is round (*p* <. 05). *American Psychologist, 49,* 997–1003.
- Dienes, Z. (2008). *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference*. Basingstoke, UK: Palgrave Macmillan.
- Fitts, D. A. (2010). The variable-criteria sequential stopping rule: Generality to unequal sample sizes, unequal variances, or to large ANOVAs. *Behavior Research Methods, 42,* 918-929.
- Frick, R. W. (1998). A better *stopping rule for conventional statistical tests. Behavior Research Methods, Instruments, & Computers, 30,* 690-697.
- Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*, 615-631.
- Nuzzo, 2014
- 大久保・岡田(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
- Simmons, J. P., Nelson, L. D., Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant, *Psychological Science, 22,* 1359-1366
- Simonsohn, U. (2013). Just Post it: The Lesson from Two Cases of Fabricated Data Detected by Statistics Alone. Psychological Science, 24, 1875-1888.
- 鈴川由美・豊田秀樹（2012）. "心理学研究"における効果量・検定力・必要標本数の展望的事例分析. 心理学研究, 83, 51-63
- Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6,* No. 1, 100–116.