From Frequentist Problems Towards Bayesian Solutions

Part 1: 心理学における再現性の危機
(と統計学との関係)

Jorge N. Tendeiro

10 August 2019

Slides at: https://rebrand.ly/Nagoya2019-Part1-JP
GitHub: https://github.com/jorgetendeiro/Nagoya-Workshop-10-Aug-2019

Today

- Fraud
- Questionable research practices
- But why
- Ireproducibility
- Didnt we see this coming
- P values p
- Confidence intervals
- Publication policies
- What do statistical associations advice statistical associations
- What to avoid
- Bayesian statistics
- References

Fraud

定義

Fraud = 科学における不正行為.

- データの偽造または加工.
- 意図的であり、故意的でないもの.
- すべての科学的研究が 問われる.
- QRPsとは際立って異なるもの (next).

有名な例

- Diederik Stapel, 社会心理学者. 2011年 停職. データの偽造と加工.
- Marc Hauser, ハーバード大学心理学者. 2011年　辞職. 科学における不正行為.
- Jens Förster, 社会心理学者. 2017年　辞職. データ改ざん.

今日は科学における不正行為 *そのもの*については 詳しくお話しません.

もっと発見しにくくて撲滅するのが難しいものについてお話します:

Questionable research practices (QRPs).

Questionable research practices
問題ある研究活動

QRPs（問題ある研究活動）

John, Loewenstein, & Prelec (2012) によって作られた用語.
See also Simmons, Nelson, & Simonsohn (2011).

- 必ずしも fraud(科学における不正行為） ではない.
- 実際に*許容される範囲内の* 研究活動とその悪用も含む.
- QRPs（問題ある研究活動）に関する問題点:
  - バイアスを加えてしまう (典型的には, 研究者の意図を*支持する*方向にバイアスが加えられる...).
  - 第一種過誤(Type I error)の確率が上がるのを犠牲に($\gg 5\%$) power(検出力) がつり上げられる.
  - 結果が 再現されない.

QRPs（問題ある研究活動）の例

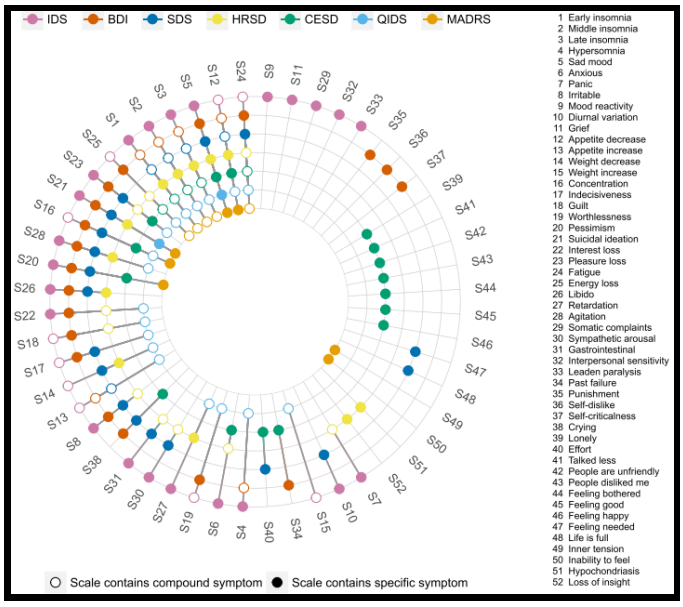
(John et al., 2012; Schimmack, 2015).

- いくつかの従属変数（dependent variables）を取り除く.
- いくつかの条件を取り除く.
- 覗き見(peeking): 逐次試験 — 結果を見ながら決める:
  - $p > .05$: もっとデータを集める.
  - $p < .05$: データ収集を止める.
- $p < .05$の結果のみ報告する.
- p-hacking: 例,
  - $p < .05$になるかどうかに基づいて外れ値を取り除く.
  - $p = .054 \longrightarrow p = .05$.
- HARKing (Kerr, 1998): 探索的に得られた結果を研究課題に変えること.
- ...

研究者の自由度

- 研究者は 多数 の決断を下さなければならない (実験デザイン, データ収集, 分析手法); Wicherts et al. (2016), Steegen, Tuerlinckx, Gelman, & Vanpaemel (2016).
- 研究者にとって*好ましい*結果になるよう操作することは充分、考えられることである.
- これらは *研究者の自由度*として知られている (Simmons et al., 2011).
- 結果的に: 誤検出による発見が増える (Ioannidis, 2005).

Fried (2017)



- 主に使われている７つのうつ病のスケールには52種類の症状が含まれている.
- これらは7つの異なるスケールに相当する.
- しかし、これらはすべて 'うつ病のレベル'として解釈される.

探索的分析から確認的分析へ

Bem (2004):

> “(…) [L]et us (…) become intimately familiar with (…) the data. Examine them from every angle. Analyze the sexes separately. Make up new composite indices. If a datum suggests a new hypothesis, try to find further evidence for it elsewhere in the data. If you see dim traces of interesting patterns, try to reorganize the data to bring them into bolder relief. If there are participants you don’t like, or trials, observers, or interviewers who gave you anomalous results, drop them (temporarily). Go on a fishing expedition for something– anything– interesting.”

これは探索的研究であることが*はっきり書かれていない限り*ダメである.

Daryl Bem氏は2011年の予知に関する有名な論文の著者である
(今日の後半部分でこのデータを使います).

最近の有名な例...

コーネル大学のBrian Wansink教授.

His description of the efforts of a visiting Ph.D student:

> I gave her a data set of a self-funded, failed study which had null results (…). I said, “This cost us a lot of time and our own money to collect. There’s got to be something here we can salvage because it’s a cool (rich & unique) data set.” I had three ideas for potential Plan B, C, & D directions (since Plan A had failed). I told her what the analyses should be and what the tables should look like. I then asked her if she wanted to do them.

> Every day she came back with puzzling new results, and every day we would scratch our heads, ask “Why,” and come up with another way to reanalyze the data with yet another set of plausible hypotheses. Eventually we started discovering solutions that held up regardless of how we pressure-tested them. I outlined the first paper, and she wrote it up (…). This happened with a second paper, and then a third paper (which was one that was based on her own discovery while digging through the data).

これは *クリエイティブまたは 型にはまらない考え方*といったものではない.

これは QRPing　（問題ある研究活動）である.

Wansink教授はどうなったか?

- かなり批判され, 彼の研究は精査された (e.g., van der Zee, Anaya, & Brown, 2017).
- 100以上におよぶ間違い が 4 本の論文から発見された…
- 現在では 40本 (!!) の論文が 撤回されている (as of July 2019).
- 1 年に及ぶ国際的な調査の結果, 彼は 辞職へと追い込まれた.

これって本当に そこまでいけない事なの?...

もちろんいけません.

- Martinson, Anderson, & Vries (2005): “Scientists behaving badly”.
- Fanelli (2009): Meta-analysis shows evidence of science misconduct.
- John et al. (2012): Evidence for QRPs in psychology.
- Mobley, Linder, Braeuer, Ellis, & Zwelling (2013): Reported evidence of pressure to *find* significant results.
- Agnoli, Wicherts, Veldkamp, Albiero, & Cubelli (2017): Evidence of QRPs, now in Italy.
- Fraser, Parker, Nakagawa, Barnett, & Fidler (2018): In other fields of science.

興味深いことに、研究における不正行為は長い間懸念されてきた (see Babbage, 1830).

参考までに:
数名の研究者は、現在の研究における現状は*それ程悪くない*としている(e.g., Fiedler & Schwarz, 2016).

研究の事前登録(preregistration)をすれば QRPs（問題ある研究活動）はなくなるのか?...

残念ながら，（まだ）そうはならない.

ちなみに、(2019年7月の）日本の研究グループ(九州大学）でも研究の事前登録がなされています：

> Ikeda, A., Xu, H., Fuji, N., Zhu, S., & Yamada, Y. (2019). *Questionable research practices following pre-registration* [Preprint]. doi: 10.31234/osf.io/b8pw9

But *why*?...

なぜQRP（問題ある研究活動）は蔓延しているのか?

それはインセンティブ（誘因）と深く関係がある (Nosek, Spies, & Motyl, 2012; Schönbrodt, 2015).

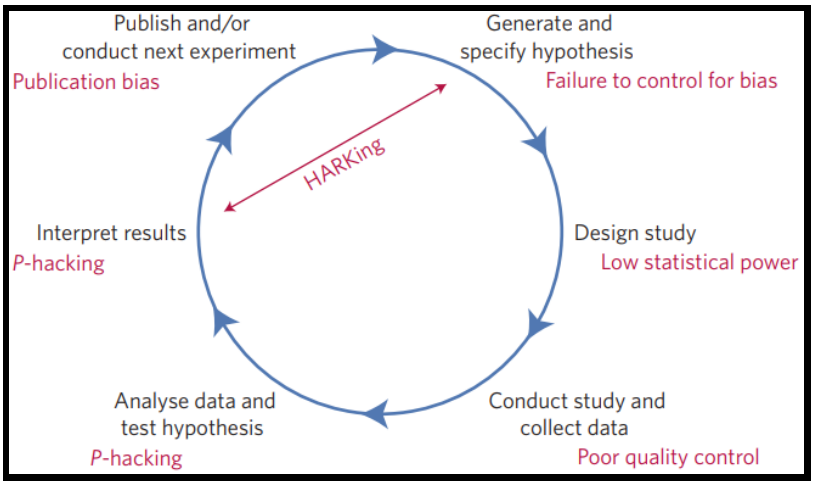
- “Publish or perish”（論文などを書かない学者は消滅する）:
  より多くの論文を、より権威あるジャーナルに出版しなければならないという現実.
  - ジャーナルは提出されたうちのごく一部の論文しか出版しない.
  - ジャーナルは否定的な結果(予期されていない結果)を出版したがらない...
- 終身雇用のポジションを得るため.
- 研究費を得るため.
- 名声 (賞, マスコミに注目される等).
- ...

しかし, 忘れてはならないのは, 研究者の最善の意図を持ってしても問題ある研究活動は起こり得るということである.

- 不十分な統計科目の教育 (そう、統計学者はこの点を理解すべきであると思います!...).
- 各分野にある永続的な伝統.

(I)reproducibility
再現性

再現できる研究への脅威



Munafò et al. (2017)

- 研究における仮説演繹法.
- 赤での記述: このモデルにおける潜在的な脅威.

再現実験（追試）の欠如

つい最近(Makel, Plucker, & Hegarty, 2012).

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Replications in Psychology Research: How Often Do They Really Occur?

Perspectives on Psychological Science
7(6) 537–542

DOI: 10.1177/1745691612460688
http://pps.sagepub.com
SAGE

Matthew C. Makel[1], Jonathan A. Plucker[2], and Boyd Hegarty[3]
[1]Duke University, [2]University of Connecticut, and [3]University of New Hampshire

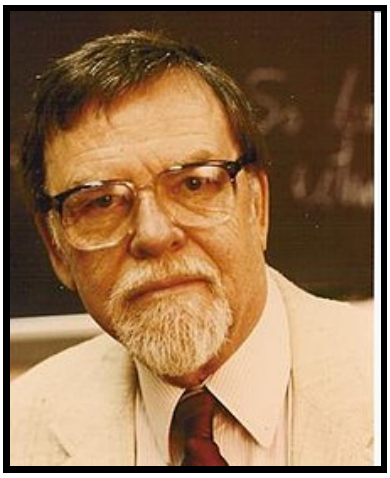
- 心理学における再現率が非常に低い (推定1%未満).
- 2012年までは, 主な再現実験は うまくいっていた!!
- しかし, 多くのケースにおいて、オリジナルの研究も追試研究も同日研究者によって報告されていた…
- Conflict of interest(利益の衝突)?…

有名な 追試の失敗

- マシュマロテスト (Watts, Duncan, & Quan, 2018)
- 自己消耗 (Friese, Loschelder, Gieseler, Frankenbach, & Inzlicht, 2019; Hagger et al., 2016; Vadillo, Gold, & Osman, 2018)
- パワーポージング (Ranehill et al., 2015)
- スタンフォード監獄実験 (Griggs, 2014; Reicher & Haslam, 2006)
- 表情フィードバック仮説 (Wagenmakers et al., 2016)
- Newborn babies' imitation (Oostenbroek et al., 2016)
- ブロッキング効果 (Maes et al., 2016)
- ステレオタイプ・スレット (Flore, Mulder, & Wicherts, 2019)
- 表情 (Gendron, Crivelli, & Barrett, 2018)
- ESP, of course! (Galak, LeBoeuf, Nelson, & Simmons, 2012; Ritchie, Wiseman, & French, 2012)
- The Mozart Effect (McKelvie & Low, 2002; Steele, Bass, & Crook, 1999)
- …

Didn’t we see this coming?
これは予測できる事ではなかったのか?

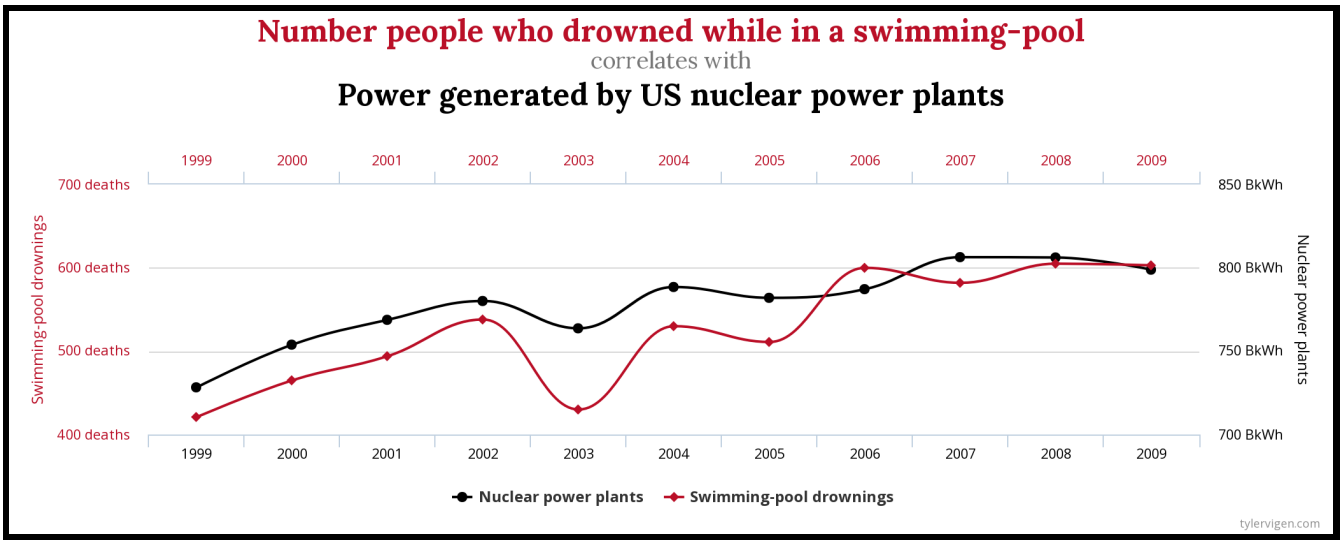
Meehl (1967)



我々がいかに不完全に仮説を立てているか (see Gelman):

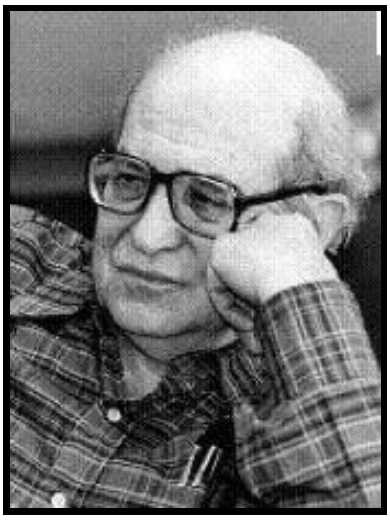
> "It is not unusual that (…) this *ad hoc* challenging of auxiliary hypotheses is repeated in the course of a series of related experiments, in which the auxiliary hypothesis involved in Experiment 1 (…) becomes the focus of interest in Experiment 2, which in turn utilizes further plausible but easily challenged auxiliary hypotheses, and so forth. In this fashion a zealous and clever investigator can slowly wend his way through (…) a long series of related experiments (…) without ever once refuting or corroborating so much as a single strand of the network."

Say what?...



http://www.tylervigen.com/spurious-correlations

Cohen (1962)

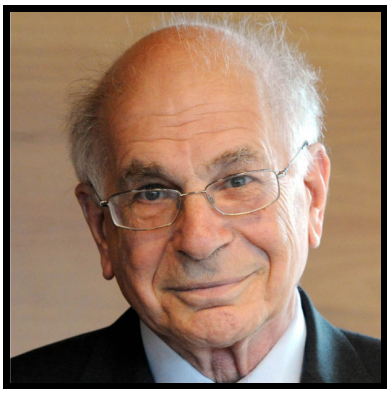


Low-powered experiments (検出力の低い実験):

> "(…) It was found that the average power (probability of rejecting false null hypotheses) over the 70 research studies was .18 for small effects, .48 for medium effects, and .83 for large effects. These values are deemed to be far too small."
>
> "(…) it is recommended that investigators use larger sample sizes than they customarily do."

Kahneman (2012)

See here.

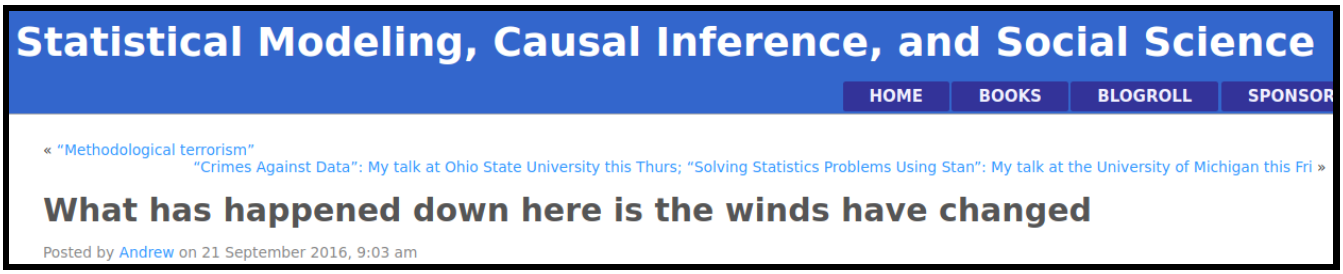


ノーベル賞受賞者, 2002.

プライミング効果について (かなり一般的な所見…):

> “The storm of doubts is fed by (…) the recent exposure of fraudulent researchers, general concerns with replicability (…), multiple reported failures to replicate salient results (…), and the growing belief in the existence of a pervasive file drawer problem (…).”
>
> “My reason for writing this letter is that I see a train wreck looming.”
>
> “I believe that you should collectively do something about this mess.”

Timeline of a train wreck


Statistical Modeling, Causal Inference, and Social Science

HOME BOOKS BLOGROLL SPONSOR

« "Methodological terrorism"
"Crimes Against Data": My talk at Ohio State University this Thurs; "Solving Statistics Problems Using Stan": My talk at the University of Michigan this Fri »

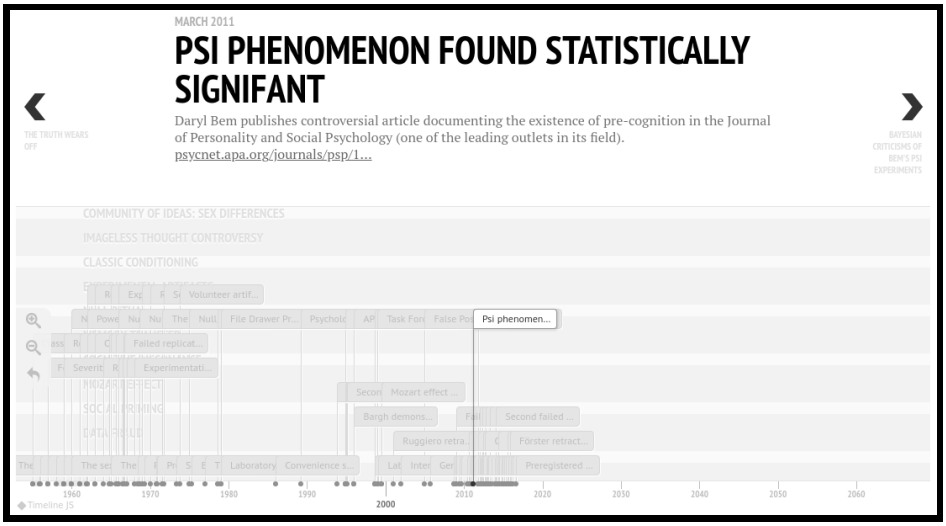
What has happened down here is the winds have changed

Posted by Andrew on 21 September 2016, 9:03 am

- Gelman教授はブログに印象深い再現危機についての見解を公開しています[timeline](#).
- Gelman教授の批判の批判も含め、このブログは一見の価値があります!
  (versus Susan Fiske's [position](#)).

Timeline of a train wreck

See also this impressive dynamic plot:
https://psyborgs.github.io/projects/replication-in-psychology/



p-values
p一値

定義

> Probability of an effect at least as extreme as the one we observed, *given that* $\mathcal{H}_0$ *is true*.
> ($\mathcal{H}_0$ *が正しいとして*、測定された効果(effect)が少なくともそれ同等かそれ以上に観測される確率)

$$\boxed{p\text{-value} = P\left(X_{\text{obs}} \text{ or more extreme} | \mathcal{H}_0\right)}$$

この定義、わかりにくいですよね?...

実際に自分で試してみましょう

以下を読んでみましょう(Falk & Greenbaum, 1995; Gigerenzer, Krauss, & Vitouch, 2004; Haller & Kraus, 2002; Oakes, 1986):

> *Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say, 20 subjects in each sample). Furthermore, suppose you use a simple independent means t-test and your result is significant (* $t = 2.7$, $df = 18$, $p = .01$*). Please mark each of the statements below as "true" or "false."* False *means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct.*

自分で試してみよう

(1) You have absolutely disproved the null hypothesis (i.e., there is no difference between the population means).	☐ True	False ☐
(2) You have found the probability of the null hypothesis being true.	☐ True	False ☐
(3) You have absolutely proved your experimental hypothesis (that there is a difference between the population means).	☐ True	False ☐
(4) You can deduce the probability of the experimental hypothesis being true.	☐ True	False ☐
(5) You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.	☐ True	False ☐
(6) You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.	☐ True	False ☐

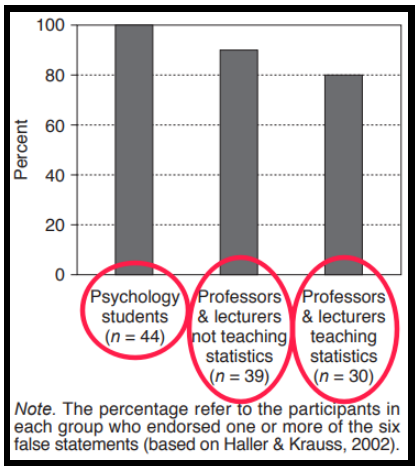
試してみよう!: rebrand.ly/pvalue

結果

すべての 文は 間違っています.

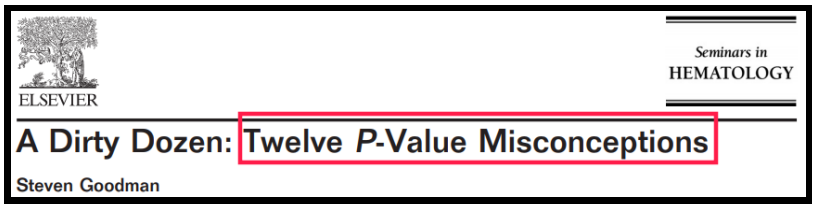
結果

では生徒と先生達はこれらの見方をどのように理解しているでしょうか？



これは2004年のものです. でもそれ以降、改善は見られていません...

Goodman (2008)



ELSEVIER

Seminars in HEMATOLOGY

A Dirty Dozen: Twelve *P*-Value Misconceptions

Steven Goodman

Table 1 Twelve *P*-Value Misconceptions

1	*If P = .05, the null hypothesis has only a 5% chance of being true.*
2	*A nonsignificant difference (eg, P ≥.05) means there is no difference between groups.*
3	*A statistically significant finding is clinically important.*
4	*Studies with P values on opposite sides of .05 are conflicting.*
5	*Studies with the same P value provide the same evidence against the null hypothesis.*
6	*P = .05 means that we have observed data that would occur only 5% of the time under the null hypothesis.*
7	*P = .05 and P ≤.05 mean the same thing.*
8	*P values are properly written as inequalities (eg, "P ≤.02" when P = .015)*
9	*P = .05 means that if you reject the null hypothesis, the probability of a type I error is only 5%.*
10	*With a P = .05 threshold for significance, the chance of a type I error will be 5%.*
11	*You should use a one-sided P value when you don't care about a result in one direction, or a difference in that direction is impossible.*
12	*A scientific conclusion or treatment policy should be based on whether or not the P value is significant.*

Greenland et al. (2016)

Eur J Epidemiol (2016) 31:337–350
DOI 10.1007/s10654-016-0149-3

CrossMark

ESSAY

Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations

Sander Greenland[1] · Stephen J. Senn[2] · Kenneth J. Rothman[3] · John B. Carlin[4] · Charles Poole[5] · Steven N. Goodman[6] · Douglas G. Altman[7]

この論文は Goodman (2008) を拡張したもので、 25 個の誤った解釈について詳しく述べられています.

The American Statistician (2019)

43本の論文による特別号を出版(Wasserstein, Schirm, & Lazar, 2019).

"$p < .05$"の世界の向こう側

Confidence intervals
信頼区間

他のより良い選択肢はあるのか？

- 信頼区間 (Confidence Intervals; CIs)は仮説検定の代替推定手段としてよく推奨される.
- 例）the Wilkinson Task Force (Wilkinson & Task Force on Statistical Inference, 1999):

> “(...) it is hard to imagine a situation in which a dichotomous accept–reject decision is better than reporting an actual p value or, better still, a confidence interval.”

- でも、信頼区間は本当に代替案として優れているのか?

定義

例） Hoekstra, Morey, Rouder, & Wagenmakers (2014).

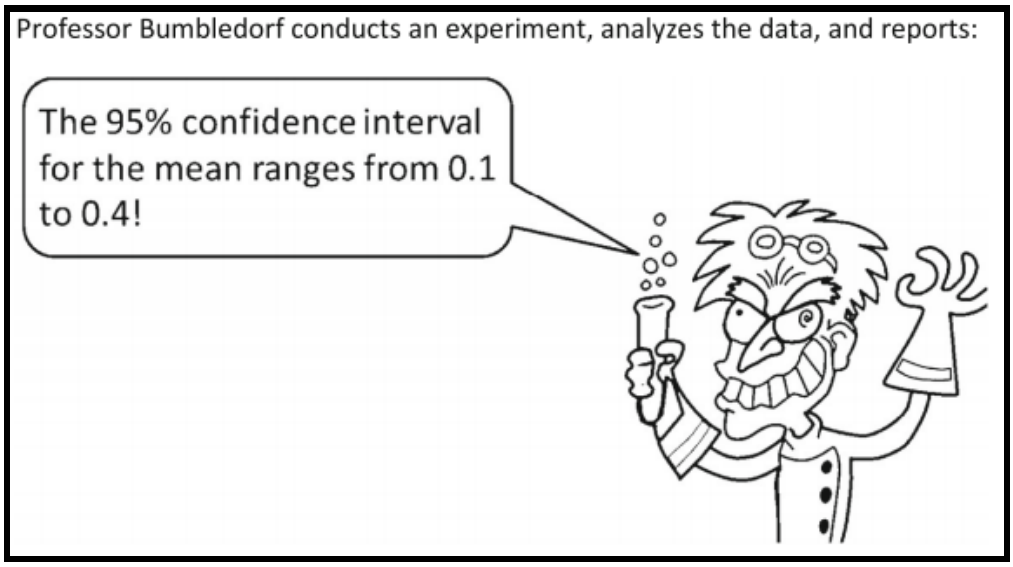
> A (say) 95% CI is a numerical interval found through a procedure that, if repeated across a series of hypothetical data, leads to an interval covering the true parameter 95% of the times.

- 信頼区間は 手法(procedure)のパフォーマンスの特性を示すものとして算出される:
  長期的にみて、 手法(procedure)がどのようなパフォーマンスをみせると予測されるか？- パラメーターの信頼区間は パラメーターの推定値付近に構成される.
- しかし、信頼区間は推定されたパラメーターの特性を直接的に示すものでは(まったく!)ない！

頭の中が混乱していますか?
ほとんどの心理学者も同じです...

自分で試してみましょう

Hoekstra et al. (2014) を参考に, Gigerenzer et al. (2004) によるp一値の研究を模倣した.



自分で試してみましょう

Please mark each of the statements below as “true” or “false”. False means that the statement does not follow logically from Bumbledorf’s result. Also note that all, several, or none of the statements may be correct:

1. The probability that the true mean is greater than 0 is at least 95%. ☐True ☐False
2. The probability that the true mean equals 0 is smaller than 5%. ☐True ☐False
3. The “null hypothesis” that the true mean equals 0 is likely to be incorrect. ☐True ☐False
4. There is a 95% probability that the true mean lies between 0.1 and 0.4. ☐True ☐False
5. We can be 95% confident that the true mean lies between 0.1 and 0.4. ☐True ☐False
6. If we were to repeat the experiment over and over, then 95% of the time the true mean falls between 0.1 and 0.4. ☐True ☐False

Please indicate the level of your statistical experience from 1 (no stats courses taken, no practical experience) to 10 (teaching statistics at a university): _____

試してみよう!: rebrand.ly/confint

結果

すべての文は 間違っています.

結果

でもどのように生徒と先生達はこれらのステートメントを理解しているでしょうか？

Table 1 Percentages of students and teachers endorsing an item

Statement	First Years ($n = 442$)	Master Students ($n = 34$)	Researchers ($n = 118$)
The probability that the true mean is greater than 0 is at least 95 %	51 %	32 %	38 %
The probability that the true mean equals 0 is smaller than 5 %	55 %	44 %	47 %
The "null hypothesis" that the true mean equals 0 is likely to be incorrect	73 %	68 %	86 %
There is a 95 % probability that the true mean lies between 0.1 and 0.4	58 %	50 %	59 %
We can be 95 % confident that the true mean lies between 0.1 and 0.4	49 %	50 %	55 %
If we were to repeat the experiment over and over, then 95 % of the time the true mean falls between 0.1 and 0.4	66 %	79 %	58 %

何が正解なの?...

> “If we were to repeat the experiment over and over, then 95% of the time the confidence intervals contain the true mean.”

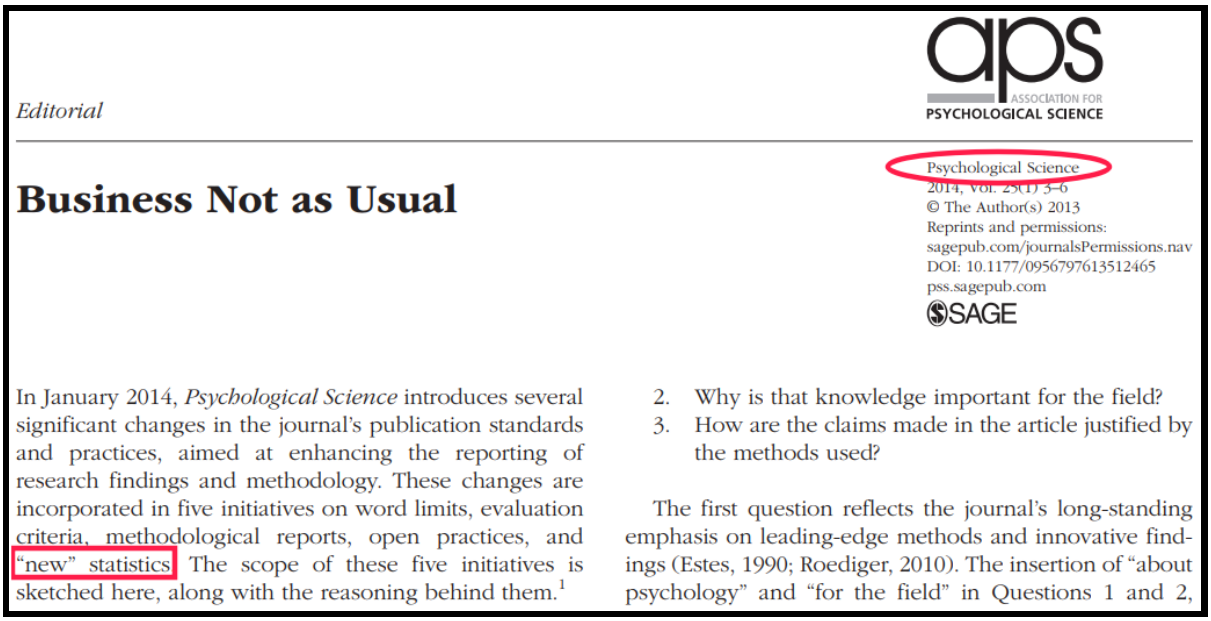
とっても参考になりますよね?!

心に留めておくべきこと:
今日のワークショップの第２部で扱う ベイズの信用区間を解釈する時にこの事を覚えておいて下さい!

ちなみに、全員がHoekstra氏の研究(García-Pérez & Alcalá-Quintana, 2016; Miller & Ulrich, 2016; see also a reply by Morey, Hoekstra, Rouder, & Wagenmakers, 2016)を支持している訳ではない.

Publication policies
出版の方針

Psychological Science (Eich, 2014)



Editorial

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Business Not as Usual

Psychological Science
2014, Vol. 25(1) 3–6

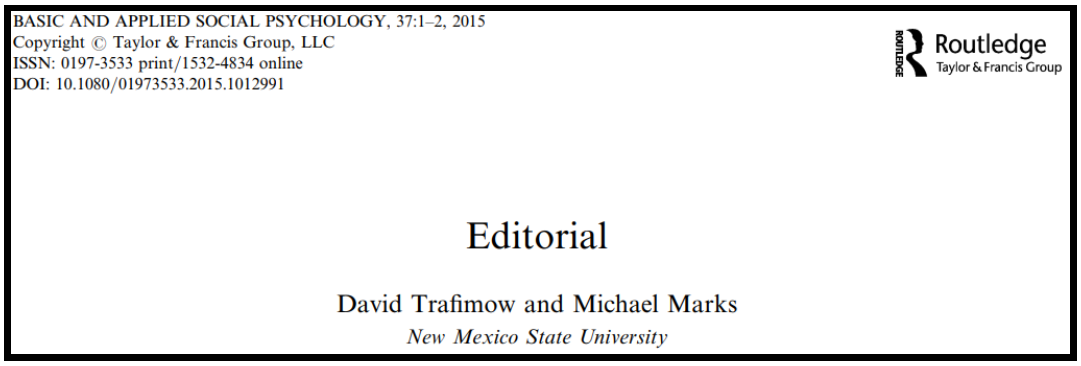
DOI: 10.1177/0956797613512465
pss.sagepub.com
SAGE

In January 2014, *Psychological Science* introduces several significant changes in the journal's publication standards and practices, aimed at enhancing the reporting of research findings and methodology. These changes are incorporated in five initiatives on word limits, evaluation criteria, methodological reports, open practices, and "new" statistics. The scope of these five initiatives is sketched here, along with the reasoning behind them.[1]

2. Why is that knowledge important for the field?
3. How are the claims made in the article justified by the methods used?

The first question reflects the journal's long-standing emphasis on leading-edge methods and innovative findings (Estes, 1990; Roediger, 2010). The insertion of "about psychology" and "for the field" in Questions 1 and 2,

Basic and Applied Social Psychology


BASIC AND APPLIED SOCIAL PSYCHOLOGY, 37:1–2, 2015

ISSN: 0197-3533 print/1532-4834 online
DOI: 10.1080/01973533.2015.1012991

Routledge
Taylor & Francis Group

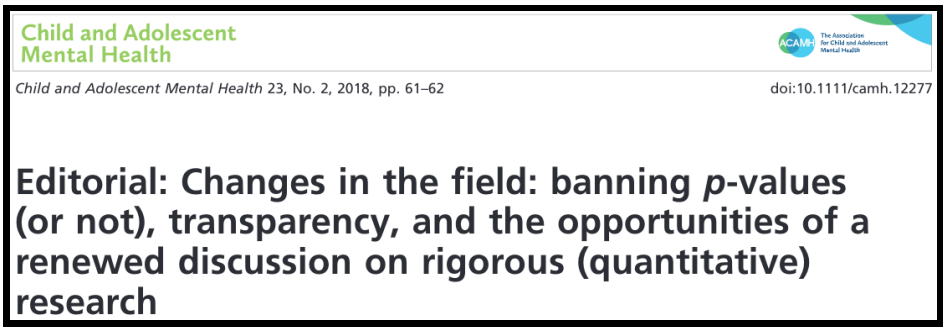
Editorial

David Trafimow and Michael Marks
New Mexico State University

> “The Basic and Applied Social Psychology (BASP) (…) emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid (…). From now on, BASP is banning the NHSTP.”

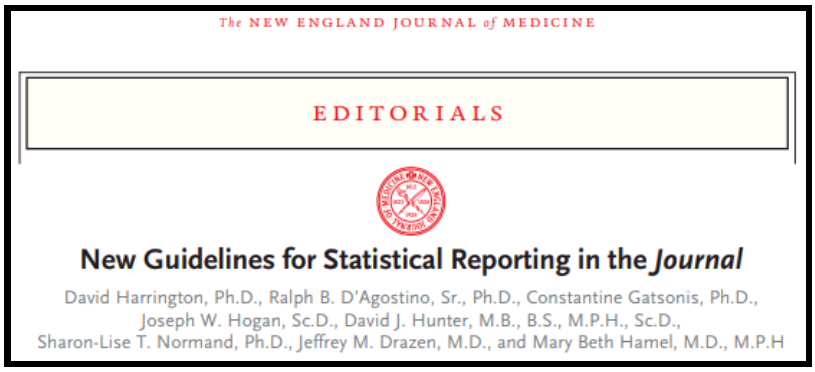
実際は うまくいったのか？see Fricker, Burke, Han, & Woodall (2019).

Child Adolescent Mental Health



> (…) I will encourage authors to provide replication syntax and data through public repositories. Moreover, I will encourage the journal to focus on a manuscript's research design and the author's justification thereof, rather than the results, with the aim of ensuring that transparent studies that explore a research question with equipoise, will be published.

The New England Journal of Medicine


The NEW ENGLAND JOURNAL of MEDICINE

EDITORIALS

New Guidelines for Statistical Reporting in the *Journal*

David Harrington, Ph.D., Ralph B. D'Agostino, Sr., Ph.D., Constantine Gatsonis, Ph.D., Joseph W. Hogan, Sc.D., David J. Hunter, M.B., B.S., M.P.H., Sc.D., Sharon-Lise T. Normand, Ph.D., Jeffrey M. Drazen, M.D., and Mary Beth Hamel, M.D., M.P.H

Editorial (Harrington et al., 2019).

> "(…) a requirement to replace p values with estimates of effects or association and 95% confidence intervals"

What do statistical associations advice?
statistical associations(統計学協会)はどうアドバイスしているのか?

Wilkinson Task Force 1999

多くのアドバイスの一部抜粋、

- pー値のみに注目しない.
- 効果サイズ(effect sizes)を記述する.
- 検出力分析(power analyses)を記述する.
- モデルの仮定(model assumptions)を調べる.

> “Novice researchers err either by overgeneralizing their results or, equally unfortunately, by overparticularizing.”

ASA 2016 (Wasserstein & Lazar, 2016)


THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108
Taylor & Francis
Taylor & Francis Group

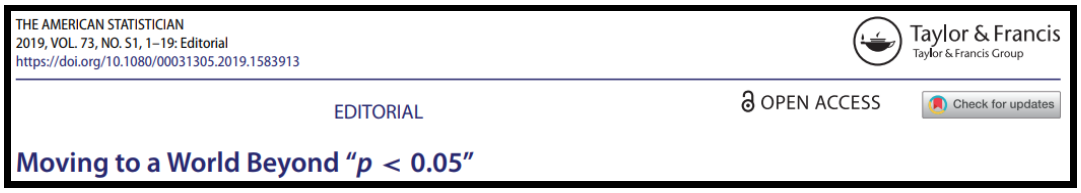
EDITORIAL

The ASA's Statement on *p*-Values: Context, Process, and Purpose

6つの原則:

1. p-値は特定の統計モデルに対して、いかにデータが不適合であるかを示せる.
2. p-値は研究仮説の正しさを示す確率、および、データがランダムによってのみ得られる確率を示すものではない.
3. 科学的結論、ビジネスまたは政策の判断を p-値が特定の値を越えるかどうかのみによって決めるべきではない.
4. 適切な推定をするためには、すべてを報告し、透明性を保たなければならない.
5. p-値または統計的有意性は効果の程度(effect size)や結果の重要性を示す指標ではない.
6. p-値自体はモデルや仮説の適合性を示すものではない.

ASA 2019 (Wasserstein et al., 2019)



これは43(!!)本の論文からなる特別号の編集者のコメントである.

主なアイデア:

- 禁止するだけでは充分ではない– いくつかの*具体案*が示されている.
- ただ…「統計的有意差」という表現は **使うべきではない** .

> "(…) it is time to stop using the term "statistically significant" entirely. Nor should variants such as "significantly different," "$p < 0.05$," and "nonsignificant" survive, whether expressed in words, by asterisks in a table, or in some other way."

But:

> "Despite the limitations of p-values (…), however, we are not recommending that the calculation and use of continuous p-values be discontinued. Where p-values are used, they should be reported as continuous quantities (e.g., $p = 0.08$). They should also be described in language stating what the value means in the scientific context."

- これだけ行えば良いという解決法はない:

> "What you will NOT find in this issue is one solution that majestically replaces the outsized role that statistical significance has come to play."

- 不確定性を受け入れる (しつこいようですが!).
  よく考え、オープンにし、謙虚である必要がある.
- ジャーナルの編集, 教育, そしてその他の組織のシステムを変える必要がある.
  その為には: ジャーナル、研究資金提供機関、教育、そしてキャリアシステムを変える必要がある.
- （当然、時間はかかりますが）再現性、実験材料やデータを公表し、信頼できるシステムを評価し、現在の「"publish or perish"（論文を出版するか滅びるか）」というシステムを変える必要がある.

ASA 2019: ベイズ統計の推奨もしている

Ruberg, S., Harrell, F., Gamalo-Siebers, M., LaVange, L., Lee J., Price K., and Peck C., Inference and Decision-Making for 21st Century Drug Development and Approval

1. Apply Bayesian paradigm as a framework for improving statistical inference and regulatory decision making by using probability assertions about the magnitude of a treatment effect.
2. Incorporate prior data and available information formally into the analysis of the confirmatory trials.
3. Justify and pre-specify how priors are derived and perform sensitivity analysis for a better understanding of the impact of the choice of prior distribution.
4. Employ quantitative utility functions to reflect key considerations from all stakeholders for optimal decisions via a probability-based evaluation of the treatment effects.
5. Intensify training in Bayesian approaches, particularly for decision makers and clinical trialists (e.g., physician scientists in FDA, industry and academia).

What to avoid
何を避けるべきか

いじめ

・ ブログ、ツイッターそしてジャーナルにおける議論は激しくなる事も度々ある.
・ 当然、批判することは、研究の 一部であるべきである.
・ もちろん、いじめるために(特にしっかりした理由なく)批判すべきではない(例 Wansink).
・ 私見ですが、時に批判することだけに 夢中になりすぎている場合もある.

NYT, 2017

(興味深い記事: 最近の反論 in Psychological Science.)

警察気取りの活動について一言

恐らく、我々自身も知られたくない内輪の秘密があることでしょう.

我々も皆、今日紹介した問題点のいくつかに該当するところがあるのではないでしょうか.

正直に告白すると:

> 私自身も該当するところがあります!!

つまり:

> *完璧な人などいない.*

Brian Nosekの言葉を借りると (as quoted here):

> “We’re not here to *be right*. We’re here to *get it right*.”

No time today for...
今日は時間の都合上できませんが...

今日は時間の都合上、カットした内容...

- Replications projects(追試プロジェクト)
- Registered reports（登録されたレポート）
- Preregistrations（研究の事前登録）
- Education(教育)
- ...

(でもこれらの事に興味があればお話することもできます!...)

今日は*統計に限定し*てお話しします.

Bayesian statistics
ベイズ統計

統計的推定における代替的アプローチ

休憩後:

ベイズ統計 への簡単なイントロダクション

References

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among italian research psychologists. *PLOS ONE*, *12*(3), e0172792. doi: 10.1371/journal.pone.0172792

Babbage, C. (1830). *Reflections on the Decline of Science in England: And on Some of Its Causes*. Retrieved from http://www.gutenberg.org/files/1216/1216-h/1216-h.htm

Bem, D. J. (2004). Writing the empirical journal article. In *The compleat academic: A career guide, 2nd ed* (pp. 185–219). Washington, DC, US: American Psychological Association.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145–153. doi: 10.1037/h0045186

Cuddy, A. J. C., Schultz, S. J., & Fosse, N. E. (2018). P-Curving a More Comprehensive Body of Research on Postural Feedback Reveals Clear Evidential Value for Power-Posing Effects: Reply to Simmons and Simonsohn (2017) - Amy J. C. Cuddy, S. Jack Schultz, Nathan E. Fosse, 2018. *Psychological Science*. doi: 10.1177/0956797617746749

Eich, E. (2014). Business Not as Usual. *Psychological Science*, *25*(1), 3–6. doi: 10.1177/0956797613512465

Falk, R., & Greenbaum, C. (1995). Significance Tests Die Hard - the Amazing Persistence of a Probabilistic Misconception. *Theory & Psychology*, *5*(1), 75–98. doi: 10.1177/0959354395051004

Fanelli, D. (2009). How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. *PLOS ONE*, *4*(5), e5738. doi: 10.1371/journal.pone.0005738

Fiedler, K., & Schwarz, N. (2016). Questionable Research Practices Revisited. *Social Psychological and Personality Science*, *7*(1), 45–52. doi: 10.1177/1948550615612150

Flore, P. C., Mulder, J., & Wicherts, J. M. (2019). The influence of gender stereotype threat on mathematics test scores of Dutch high school students: A registered report. *Comprehensive Results in Social Psychology*, 1–35. doi: 10.1080/23743603.2018.1559647

Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PLOS ONE*, *13*(7), e0200303. doi: 10.1371/journal.pone.0200303

Fricker, R. D., Burke, K., Han, X., & Woodall, W. H. (2019). Assessing the Statistical Analyses Used in *Basic and Applied Social Psychology* After Their *p* -Value Ban. *The American Statistician*, *73*(sup1), 374–384. doi: 10.1080/00031305.2018.1537892

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191–197. doi: 10.1016/j.jad.2016.10.019

Friese, M., Loschelder, D. D., Gieseler, K., Frankenbach, J., & Inzlicht, M. (2019). Is Ego Depletion Real? An Analysis of Arguments. *Personality and Social Psychology Review*, *23*(2), 107–131. doi: 10.1177/1088868318762183

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). *Correcting the Past: Failures to Replicate Psi* (SSRN Scholarly Paper No. ID 2001721). Retrieved from Social Science Research Network website: https://papers.ssrn.com/abstract=2001721

García-Pérez, M. A., & Alcalá-Quintana, R. (2016). The Interpretation of Scholars' Interpretations of Confidence Intervals: Criticism, Replication, and Extension of Hoekstra et al. (2014). *Frontiers in Psychology*, *7*. doi: 10.3389/fpsyg.2016.01042

Gendron, M., Crivelli, C., & Barrett, L. F. (2018). Universality Reconsidered: Diversity in Making Meaning of Facial Expressions. *Current Directions in Psychological Science*, *27*(4), 211–219. doi: 10.1177/0963721417746794