

効果量と信頼区間： p 値だけでは不十分

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

実証的研究はすべて帰無仮説検定を 採用する(Hubbard & Ryan, 2000)

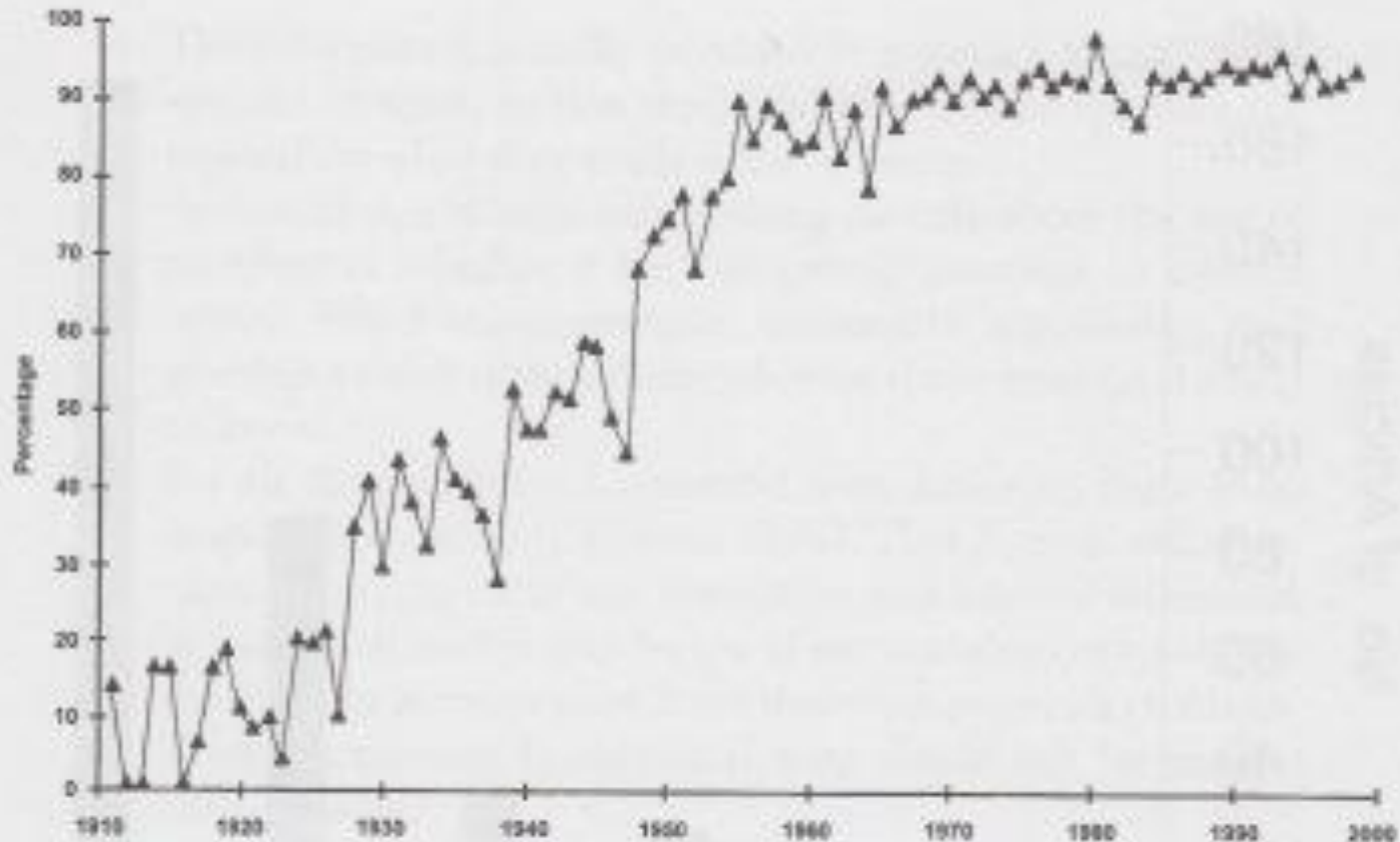
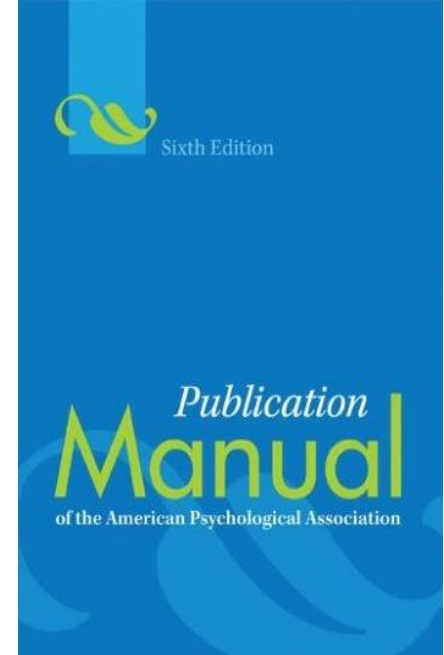


Figure 1.1. Percentage of articles reporting results of statistical tests in 12 journals of the American Psychological Association from 1911 to 1998. From "The Historical Growth of Statistical Significance Testing in Psychology—And Its Future Prospects," by R. Hubbard and P. A. Ryan, 2000, *Educational and Psychological Measurement*, 60, p. 665. Copyright 2001 by Sage Publications. Reprinted with permission.

APA Publication Manual

- 1,000を超える学術論文誌がAPAマニュアルを採用している (APA, 2001, p. xxi)
- 心理学の研究・教育のスタイルに強い影響力
- 最新版は第6版 (APA, 2009)
 - 日本語訳も (前田・江藤・田中訳, 2011)



APA Publication Manual (p. 33)

APA stresses that NHST is but a starting point and that additional reporting elements such as effect sizes, confidence intervals, and extensive description are needed to convey the most complete meaning of the results.

- 帰無仮説検定は分析の始まりにすぎません。結果の意味を十分に伝えるためには、効果量、信頼区間、そして詳しい記述が不可欠であるとAPAは強く主張します。

However, complete reporting of all tested hypotheses and estimates of appropriate effect sizes and confidence intervals are the minimum expectations for all APA journals.²

- すべてのAPAの論文誌において、著者は仮説検定の結果および効果量と信頼区間を適切に報告してください。これらは最低限求められることです。

実際の例：効果量と信頼区間

Repeated measures ANOVA² with the factors emotional expression (angry, fearful, happy, neutral) and rotation angle (-8° , -6° , -4° , -2° , 0° , 2° , 4° , 6° , or 8°) revealed significant main effects of emotional expression, $F(3, 153) = 36.59$, $p < .001$, $\eta_p^2 = 0.42$, and of rotation angle, $F(2.24, 114.46) = 199.58$, $p < .001$, $\eta_p^2 = 0.80$. The interaction between emotional expression and rotation angle was significant, $F(9.12, 465.25) = 2.66$, $p = .005$, $\eta_p^2 = 0.05$.

In the vertical task, by contrast, a majority of both left-handers (89%) and right-handers (83%) drew the good animal in the top box and the bad animal in the bottom box (left-handers: sign test on 8 top vs. 1 bottom, $p_{\text{rep}} = 0.93$; right-handers: sign test on 88 top vs. 18 bottom, $p_{\text{rep}} = 0.99$; see Figure 1d). By Fisher's exact test and binary logistic regression, there was no reliable association between the handedness of the participant and the top-bottom placement of the good and bad animals (p_{rep} for the exact association = 0.74, OR = 0.61, 95% CI = 0.07–5.19).

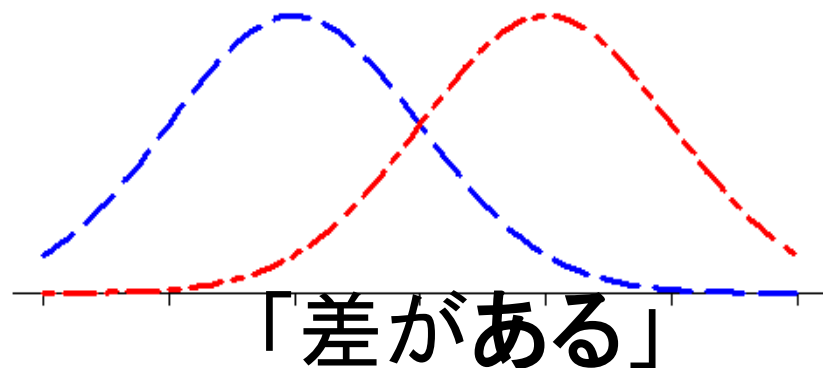
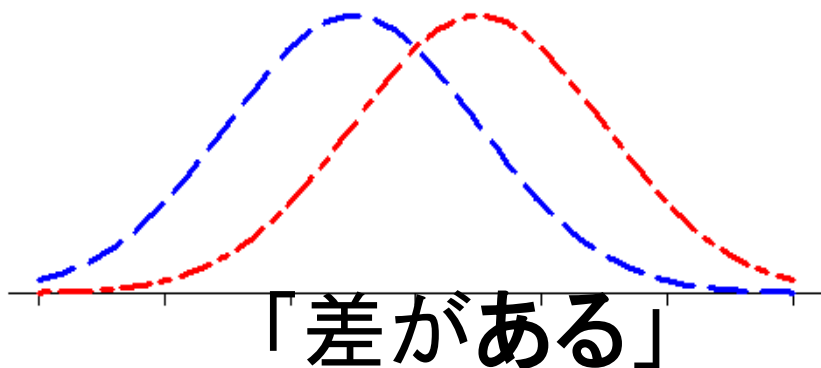
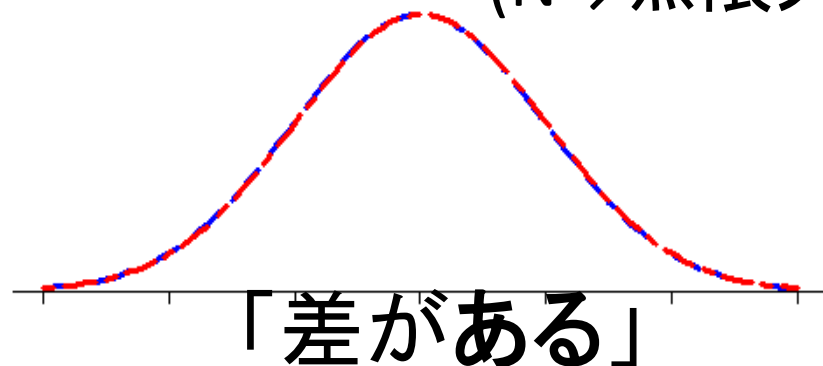
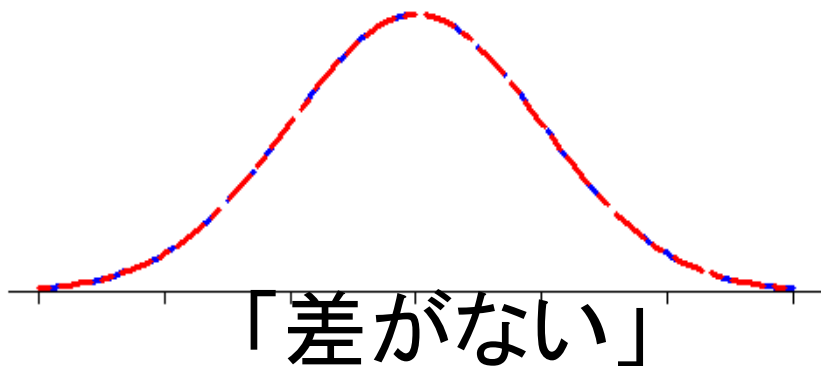
New Statistical Guidelines for Journals of the Psychonomic Society (8/23/2012)

- 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。
 - 効果量, 信頼区間などを併せて報告せよ。
- 検定力を考慮し, 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。
 - などなど。

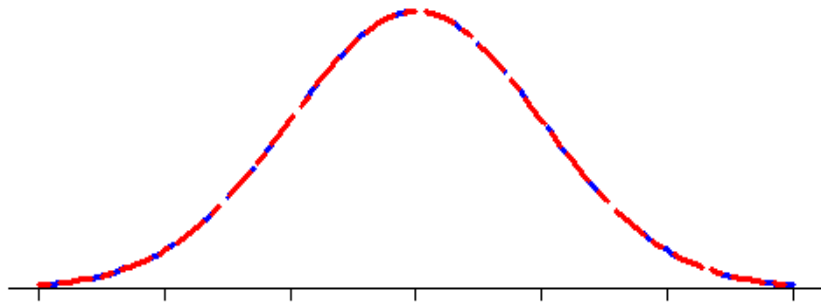
p値で考える(帰無仮説検定)

$$H_0: \mu_1 = \mu_2 \text{ vs } H_1: \mu_1 \neq \mu_2$$

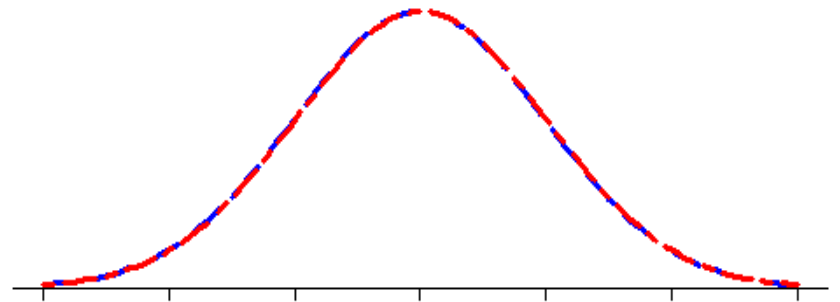
($N \rightarrow \infty$)



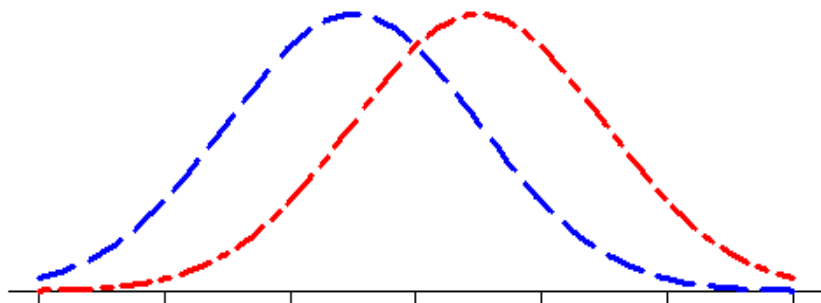
効果量で考える



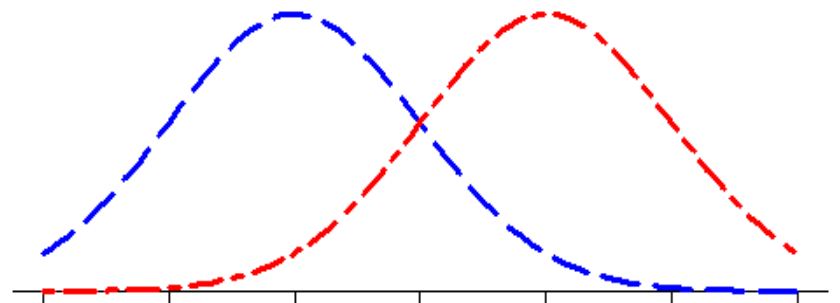
「差がない」



「0.01SDだけ差がある」



「1SDだけ差がある」



「2SDだけ差がある」

効果量

d族の効果量

- 2群の平均値差
(平均値差の指標)

r族の効果量

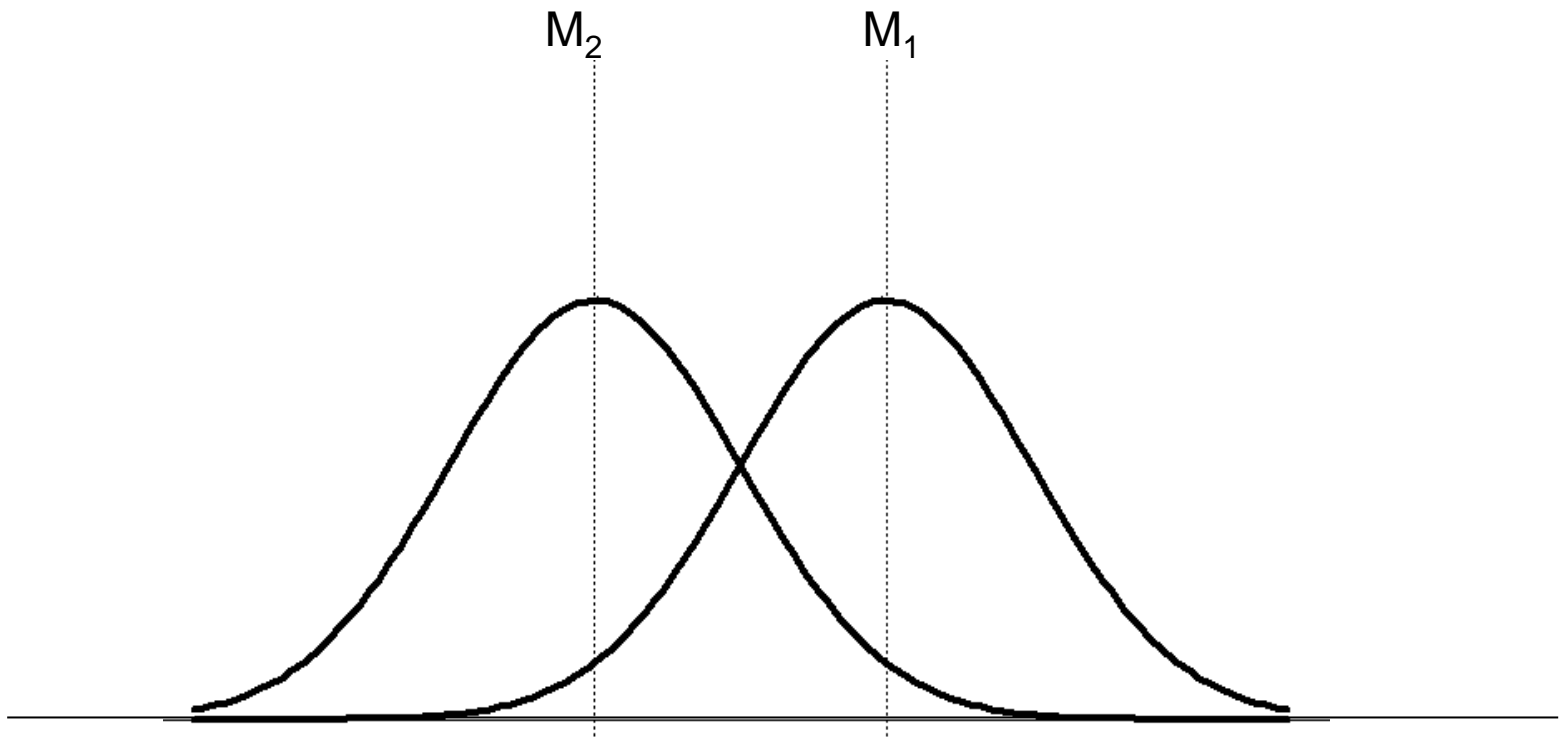
- 相関係数
- 分散説明率
(POV; percent of variance explained)

(関連の大きさの指標)

d族の効果量

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

– S_p はプールした標準偏差



計算方法

- エクセルでも可。手計算でも。
- 簡易なweb サイトも
 - 例
<http://www.cognitiveflexibility.org/effectsize/>

The screenshot shows a web browser window titled "Effect Size Calculator". The address bar shows the URL "www.cognitiveflexibility.org/effectsize/". The page content includes the title "Effect Size Calculator", a section "What It Does" explaining the calculator's purpose, and two main sections: "Between-subjects Studies" and "Within-subjects Studies". The "Between-subjects Studies" section contains a "Calculator for Means and SDs" with input fields for Mean 1, SD 1, Mean 2, and SD 2, and a "Correlation between means" field. The "Within-subjects Studies" section contains a "Calculator for *t* scores and Sample Sizes" with input fields for *t* score, n for Mean 1, and n for Mean 2. Both sections have "Submit" and "Reset Form" buttons. The page also includes a "Choice of SD Term" section at the bottom.

Effect Size Calculator

What It Does

This calculator evaluates the effect size between two means (i.e., Cohen's *d*; Cohen, 1988), which is the difference between means divided by standard deviation.

Between-subjects Studies

Enter the two means, plus SDs for each mean. To compute effect size using pooled or control condition SD, only enter one SD (computed as appropriate, of course).

Within-subjects Studies

For within-subjects studies, one must correct for dependence among means in order to make direct comparisons to effect sizes from between-subjects studies. To do this, you also need to enter the correlation between the two means, so that Morris and DeShon's (2002) equation 8 can be applied.

Calculator for Means and SDs

Mean 1: SD 1:

Mean 2: SD 2:

Required. SD 2 is optional but recommended. See description below.

Correlation between means:

Use for within-subjects data only. See description above.

Submit Reset Form

Calculation Using *t* score

Alternatively, enter the *t* score and sample size for each condition. In this case, Ray and Shadish's (1996) equation 2 will be used, which produces an equivalent effect size to using pooled SD. See the paper, and please don't combine effect sizes using different SD terms.

Calculator for *t* scores and Sample Sizes

t score:

n for Mean 1: n for Mean 2:

Required. Use for between-subjects data only.

Submit Reset Form

Choice of SD Term

I prefer to use the average of each mean's individual SD, as opposed to pooled or control condition SD. This decision is based on reading more than

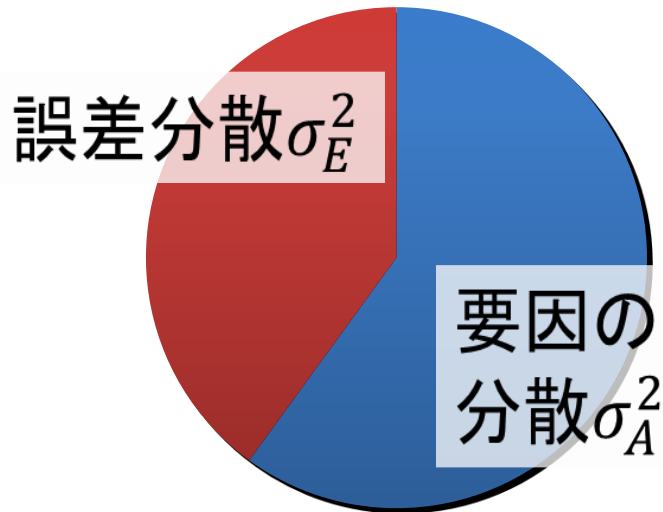
r族の効果量

$$\eta^2 = \frac{\sigma_A^2}{\sigma_T^2}$$

- η^2 = 全分散に対する当該要因の分散の割合

1 要因被験者間分散分析

$$\text{全分散 } \sigma_T^2 = \sigma_A^2 + \sigma_E^2$$



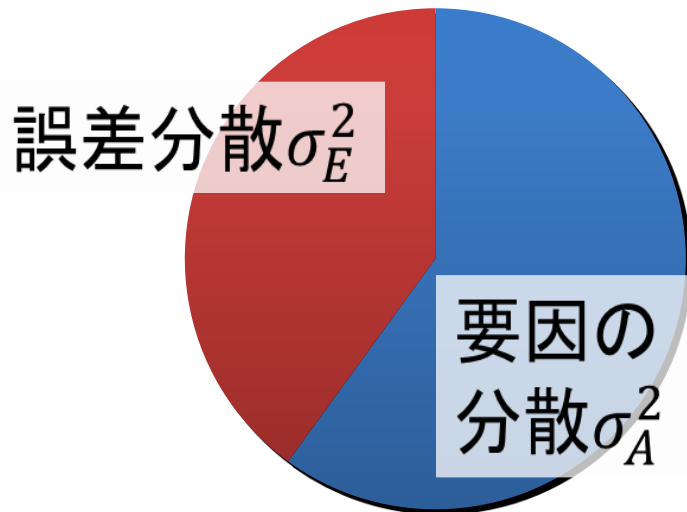
r族の効果量(その2)

$$\eta_p^2 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2}$$

- η_p^2 = 誤差分散に対する当該要因の分散の割合

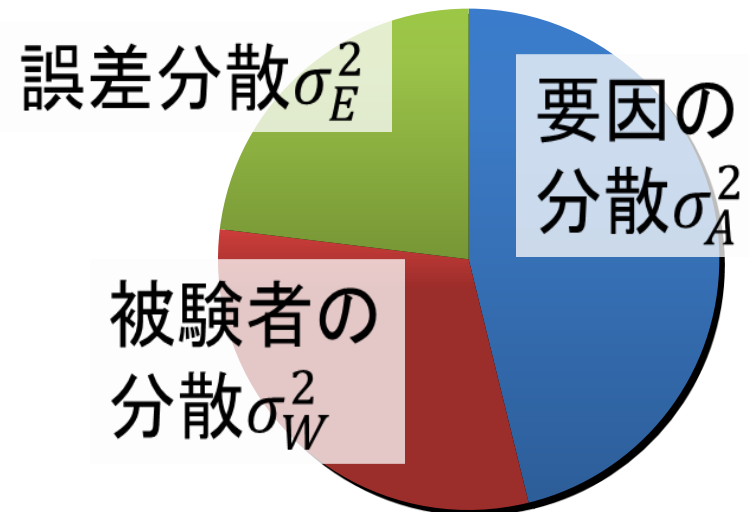
1要因被験者間分散分析

$$\text{全分散 } \sigma_T^2 = \sigma_A^2 + \sigma_E^2$$



1要因被験者内分散分析

$$\text{全分散 } \sigma_T^2 = \sigma_A^2 + \sigma_W^2 + \sigma_E^2$$



具体例： η^2 と η_p^2 の比較

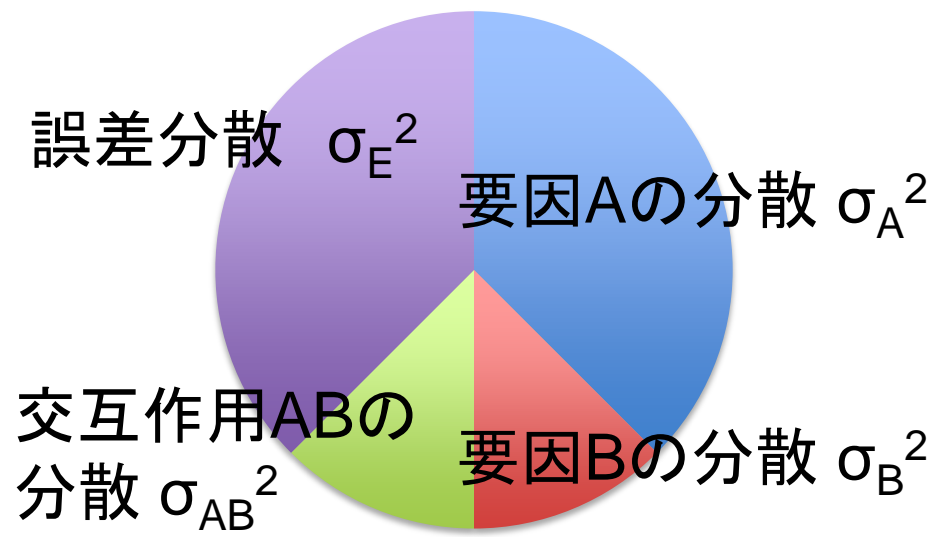
SS	SS	df	Ms	F	p
要因A	69.80	2	34.90	9.27	.001
要因B	120.00	1	120.00	31.86	.000001
交互作用 A X B	8.60	2	4.30	1.14	.336
誤差 E	90.40	24	3.77		
全体 T	288.80				

要因A

$$\eta^2 = 69.80/288.80 = .24$$

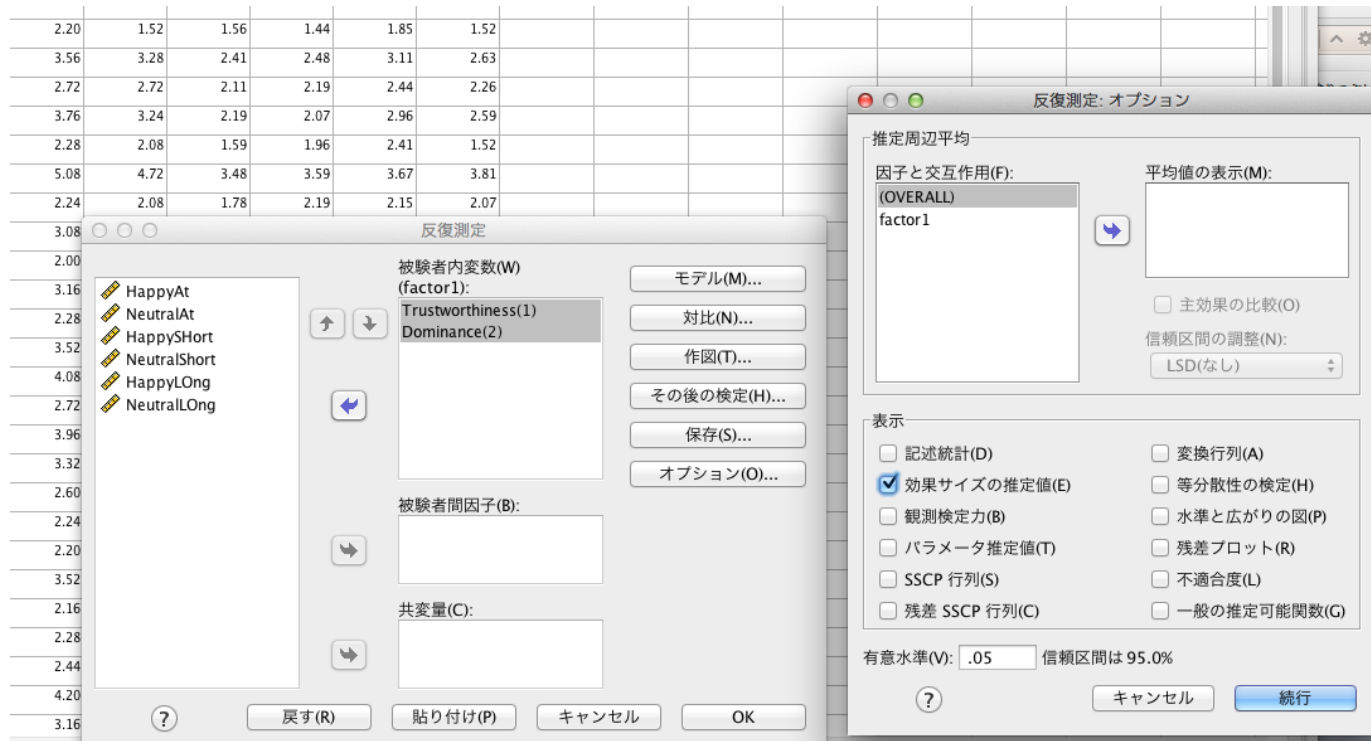
$$\eta_p^2 = 69.80/(69.80+90.40) = .44$$

2要因被験者間分散分析



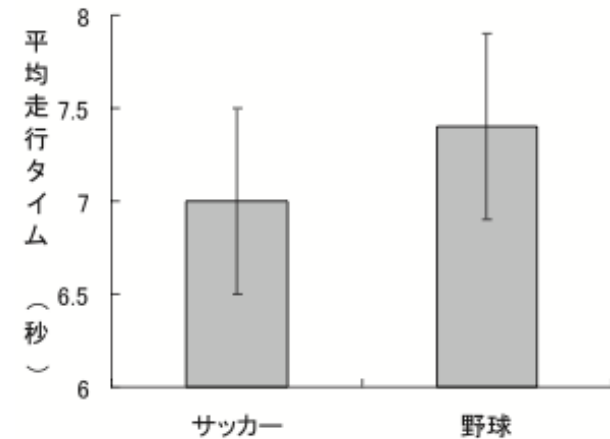
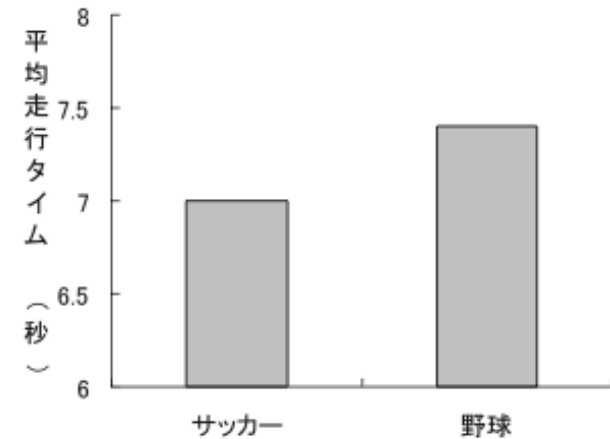
計算方法

- エクセルでも，手計算でも。
- η_p^2 はSPSSで簡単に出力できる。
- R でanovakunを使うともっと多様なことが。。。。



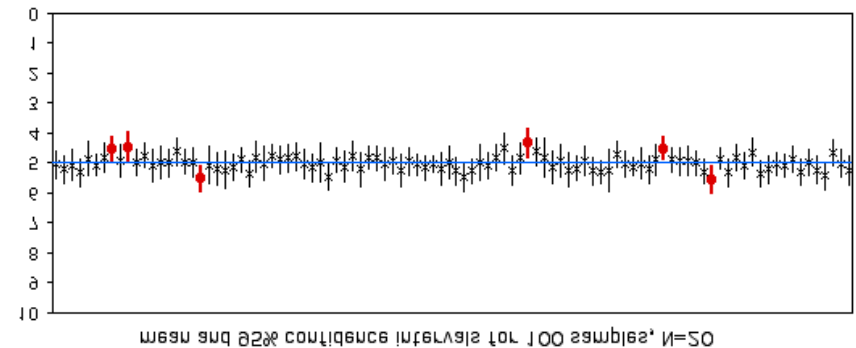
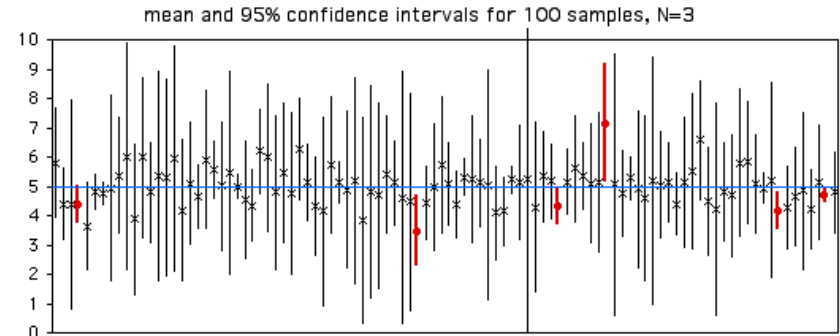
信頼区間：点推定と区間推定

- 点推定
 - 母集団を代表する点の推定：
平均値，中央値
- 区間推定
 - 点推定のばらつきを推定。
 - ある確率で(母数の)代表値
が存在する区間を推定する
- ある確率＝信頼水準
- ある信頼水準において推定
された区間＝信頼区間



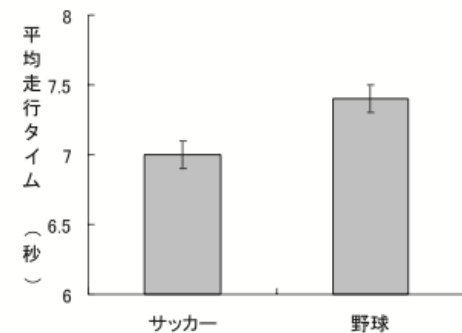
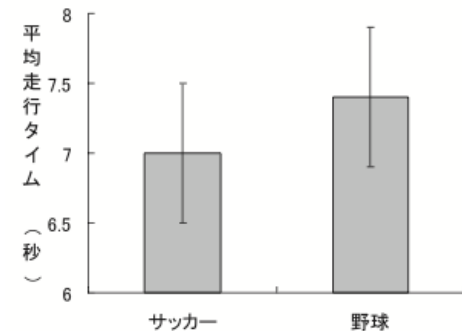
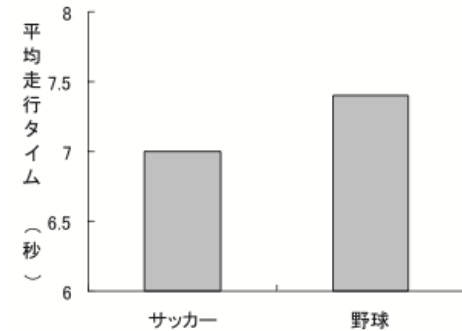
信頼区間 (Confidence Interval, CI)

- 95%信頼区間 (95% CI)
 - 慣習的にこの値が用いられる
 - $95\%CI = M \pm SE \times t_{95\%}$
 - $t_{95\%}$ = t分布の95%点
- 100回の推定を行えば、95回はその範囲に母数が含まれる



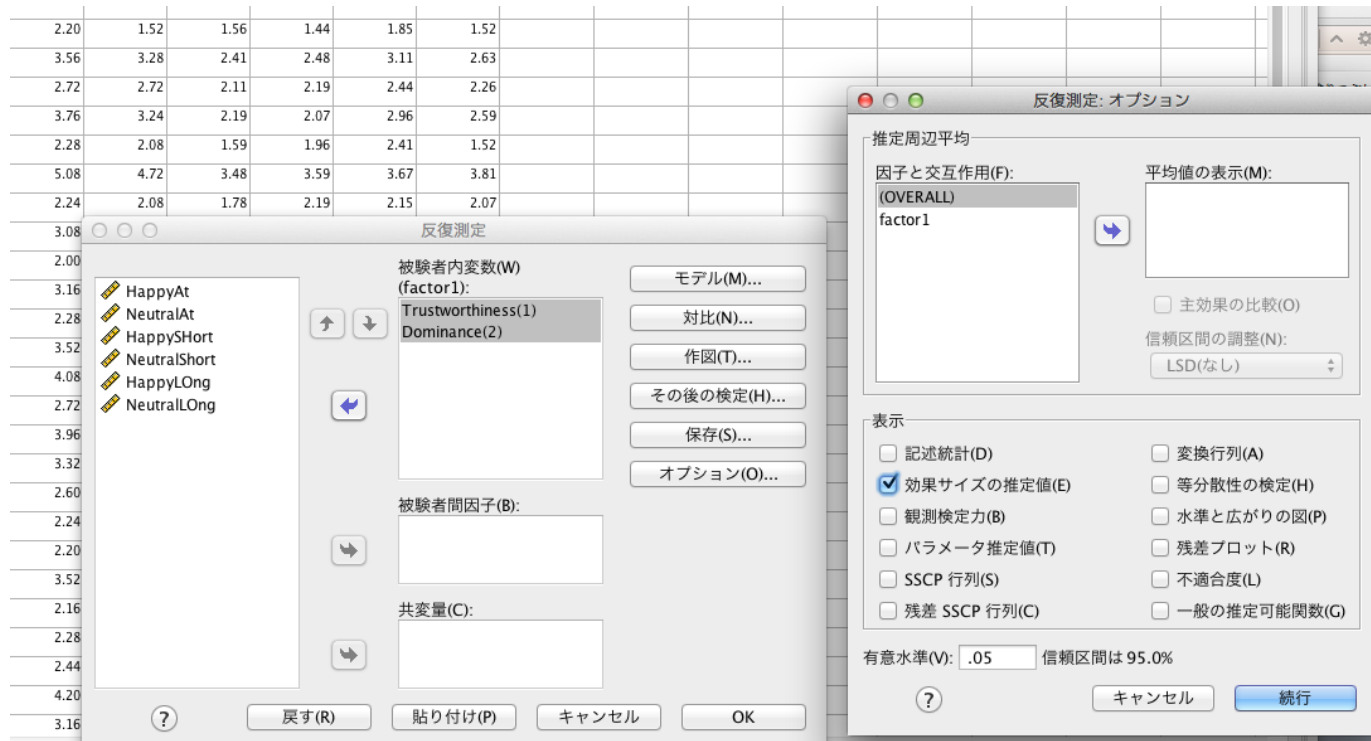
信頼区間で なにがわかるか？

- 母数の含まれる範囲
 - 応用的には重要
- データの精度
 - 効果・誤差
 - 標本サイズ
- 効果の大きさ
- 検定力
- 「差がない」仮説のエビデンス



計算方法

- エクセルでも，手計算でも。
- SPSSで簡単に出力できる。



最近の論文における効果量の報告 (Fritz et al., 2012)

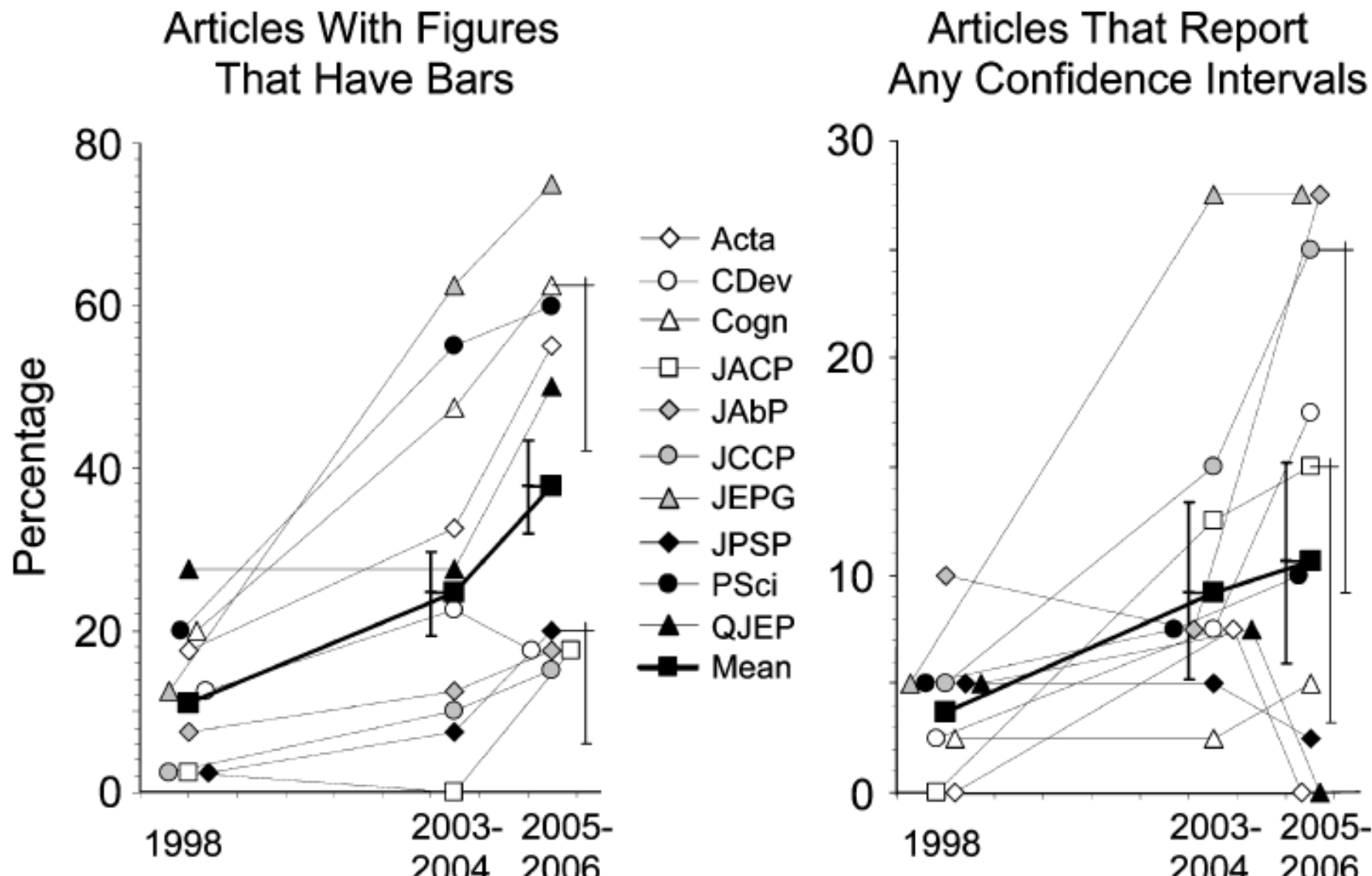
Table 2

Number (and Percentage) of Articles Reporting Effect Size Estimates Associated With ANOVA

Year	Articles with ANOVA	Any ES measure	η^2	η_p^2
2009	27	18 (67)	1 (6)	17 (94)
2010	32	15 (47)	5 (33)	9 (60)
Overall	59	33 (56)	6 (18)	26 (79)

- JEP:General 2009-2010に掲載された論文
- ANOVAを用いた論文でおよそ5-6割が効果量を報告
- その大半が η_p^2

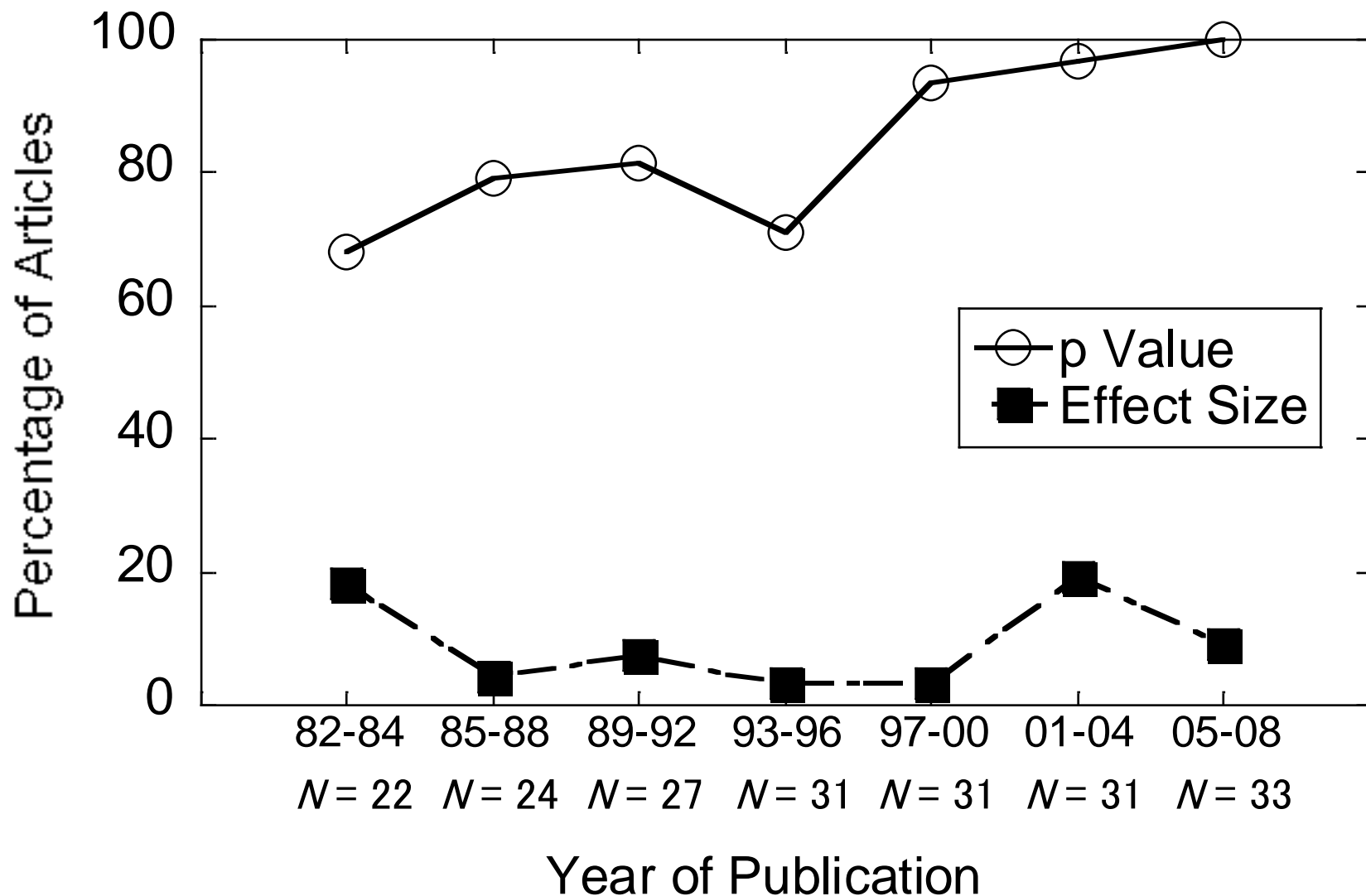
信頼区間報告に関する変化(Cumming et al., 2007)



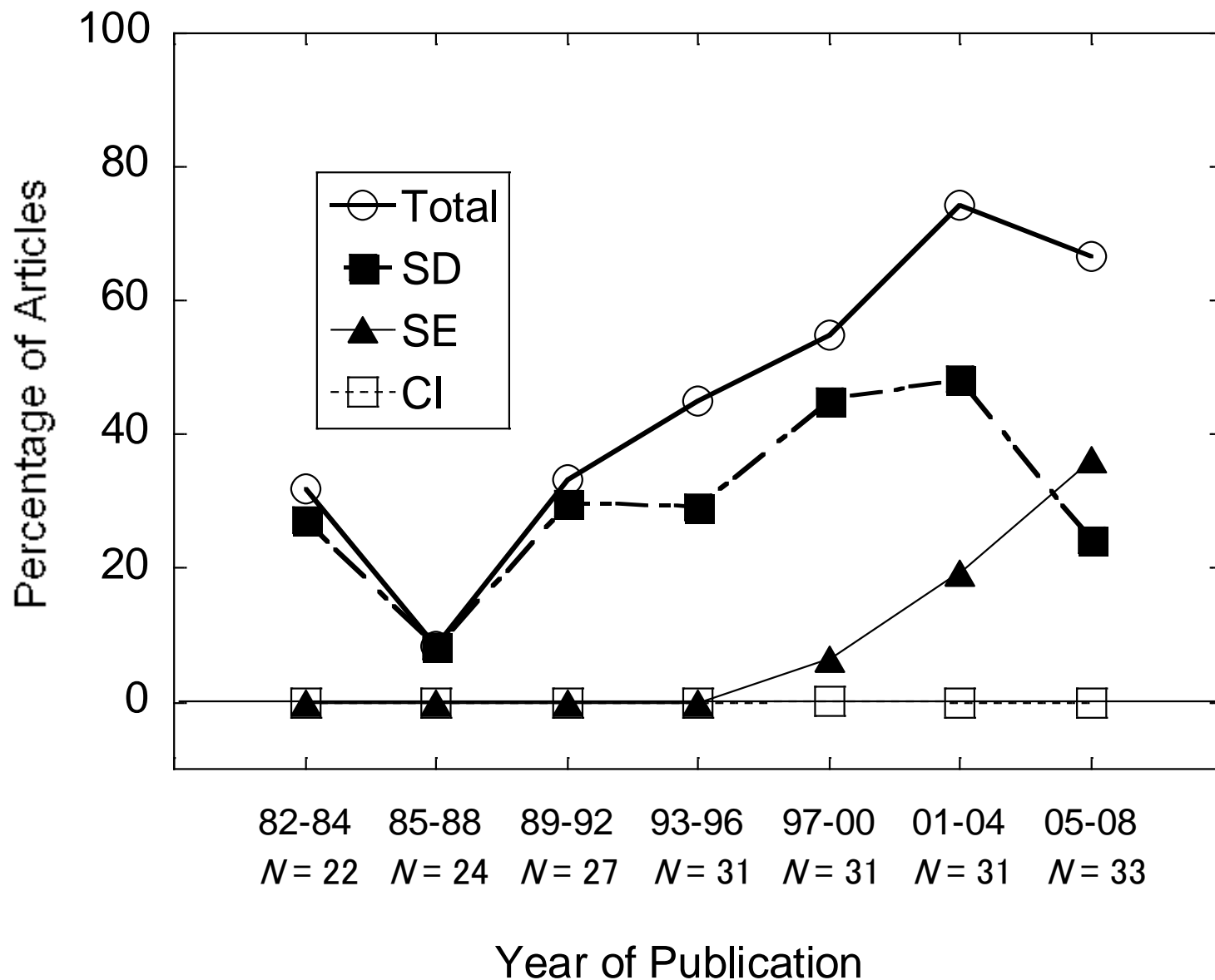
日本の現状

- 大久保 (2009)
 - 論文誌「基礎心理学研究」を対象
 - 199報の論文（1982 -2008年）
- 効果量と信頼区間の報告数

効果量の報告＝ほぼなし(相関係数, 回帰係数, 決定係数のみ)



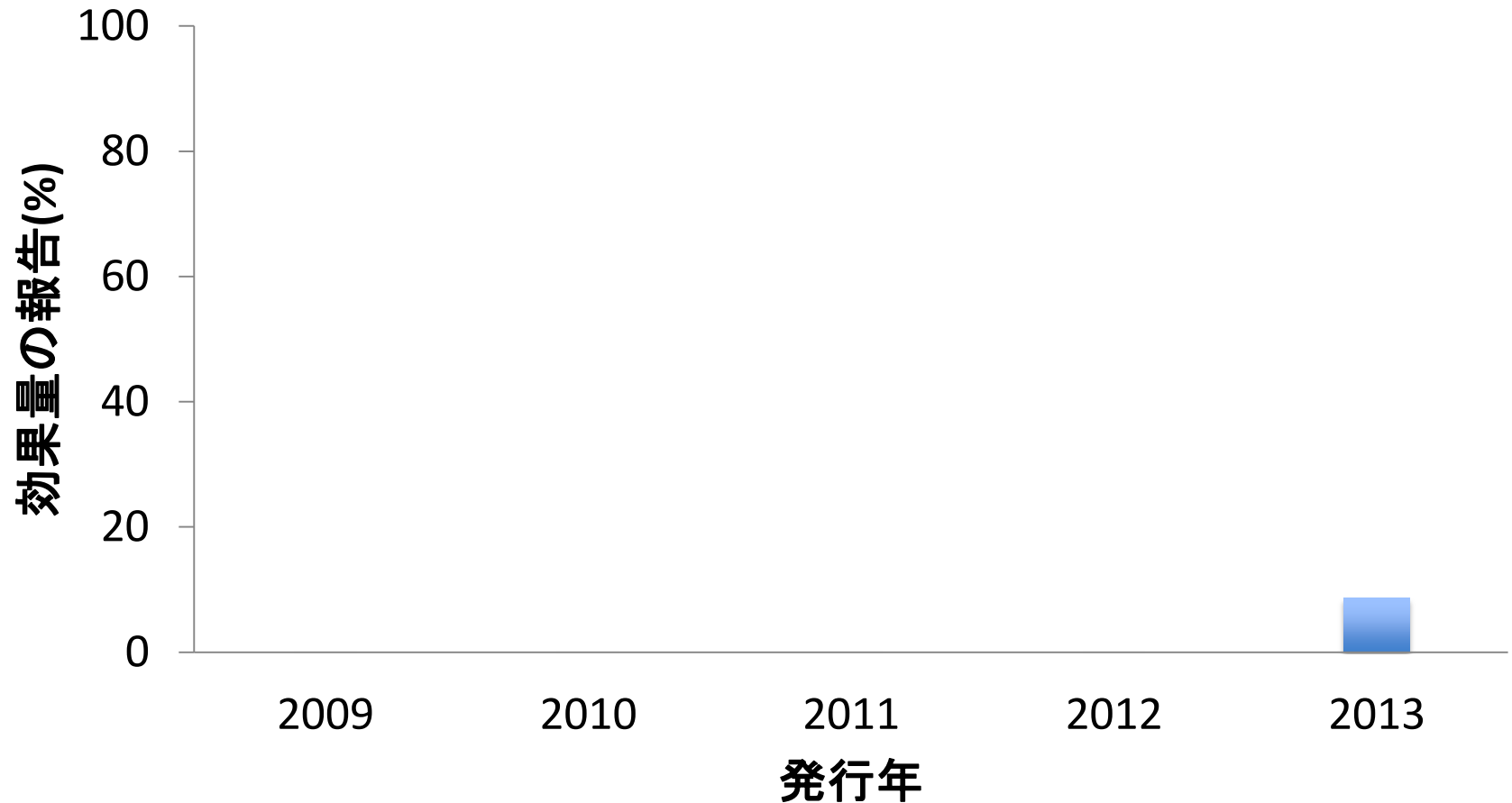
信頼区間の報告＝ほぼゼロ



論文誌「パーソナリティ研究」における 効果量の報告

- 対応のないt検定を対象
- 2009年から2013年
- CohenのdやHedgesのgなど差の効果量の報告のみを対象

論文誌「パーソナリティ研究」における 効果量の報告



n = 10 (2009), 40 (2010), 25 (2011), 17(2012), 23 (2013)

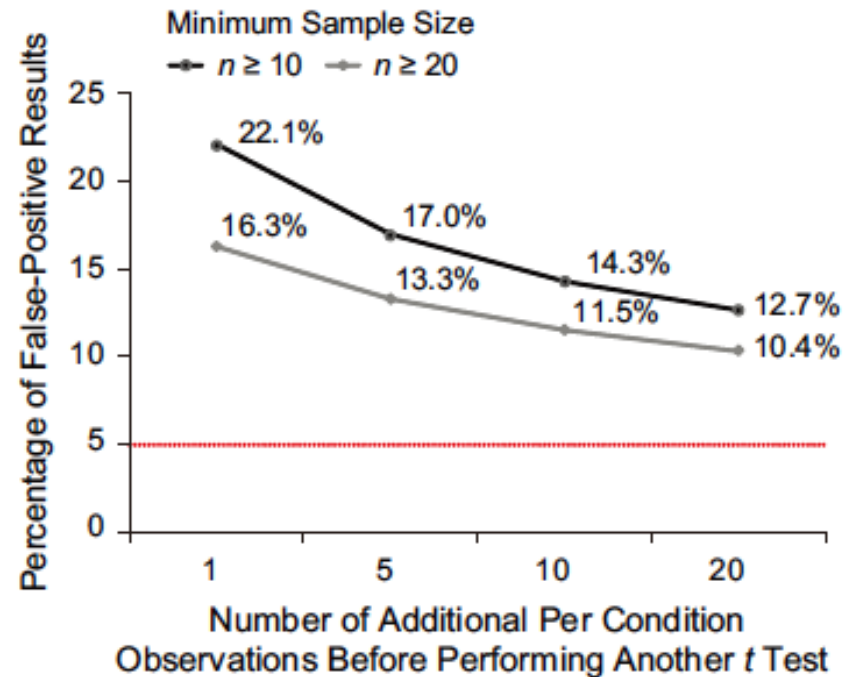
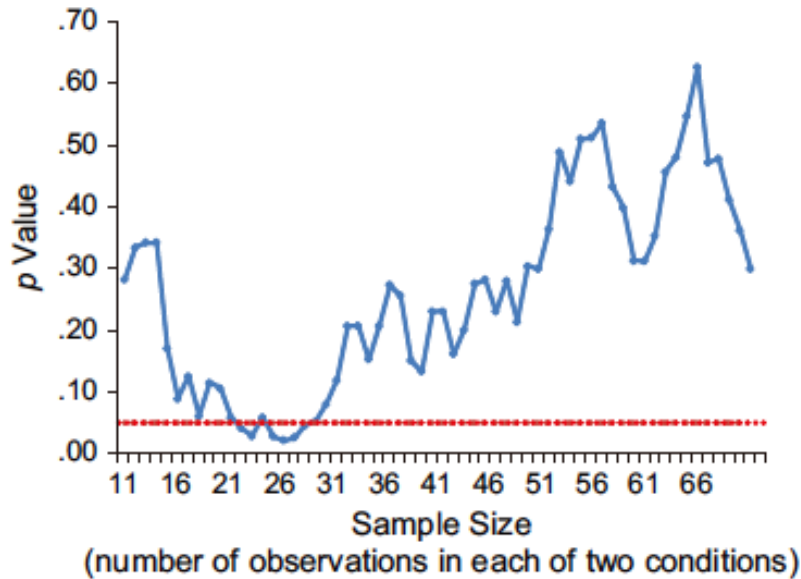
“false-positive psychology” 問題

(Simmons, Neson, & Simonsohn, 2011)

- Bem (2013)など再現性や頑健性に疑問のある研究
- 擬陽性の生じやすさをシミュレーションで検討
 - 標本サイズ, 独立変数の数, 共変量の効果
- 「n増し」と擬陽性の関係



n増しと擬陽性の関係 (Simmons et al. 2011)

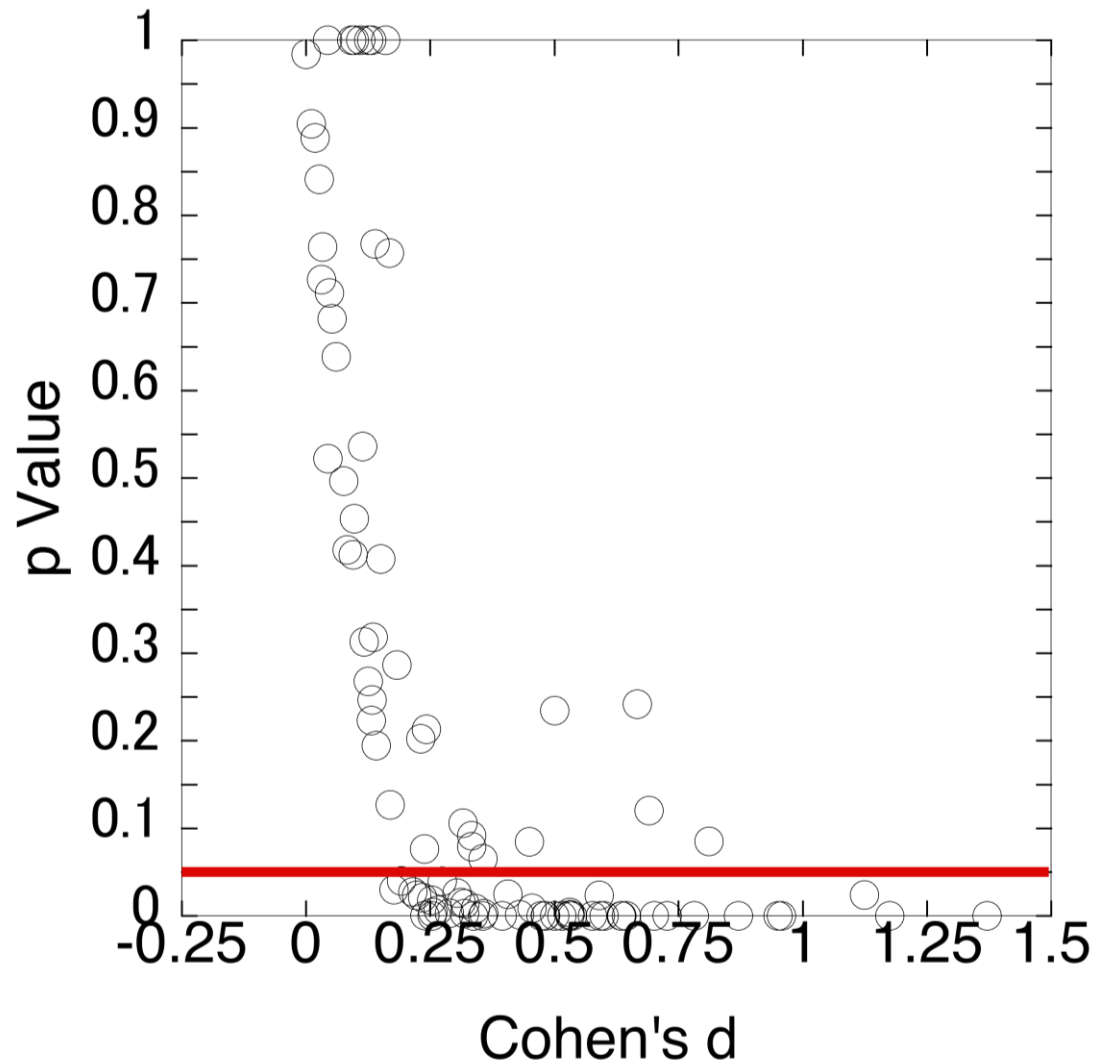


ただし、実際の研究場面では擬陽性の確率はもっと減る(Murayama et al., in press)

p値のみに頼るとどうなるか？

- 雑誌「パーソナリティ研究」
 - 論文記載のデータから効果量を算出
 - 2009年から2013年まで
- p値と効果量の関係を検討
 - 対応のないt検定を対象

独立な2群の差の検定 (N= 98)



p値と効果量の判断の非対応

- 同様の効果量でも, $p < .05$ と $p > .06$
 - 特に小さな効果 ($d < .2$)
 - 中程度の効果量でも, 有意差として検出されないことがある。

指標	小さな効果	中程度の効果	大きな効果
d	.2	.5	.8
η^2	.01	.06	.14

APA Publication Manual (p. 33)

However, complete reporting of all tested hypotheses and estimates of appropriate effect sizes and confidence intervals are the minimum expectations for all APA journals.²

- すべてのAPAの論文誌において、著者は仮説検定の結果および効果量と信頼区間を適切に報告してください。これらは最低限求められることです。



- 複数の指標から、統合的かつ整合的にデータを解釈すべき

Rosnow & Rosenthal (1989)

- 「神は $p < .05$ を $p < .06$ と等しく, そして同じくらい強く愛してくださる。」



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

伝えるための
心理統計



大久保圭
岡田謙介

Psychology
Psychological
Psychology
Psychology
Psychology
Psychology
Psychology
Psychology

心理学会