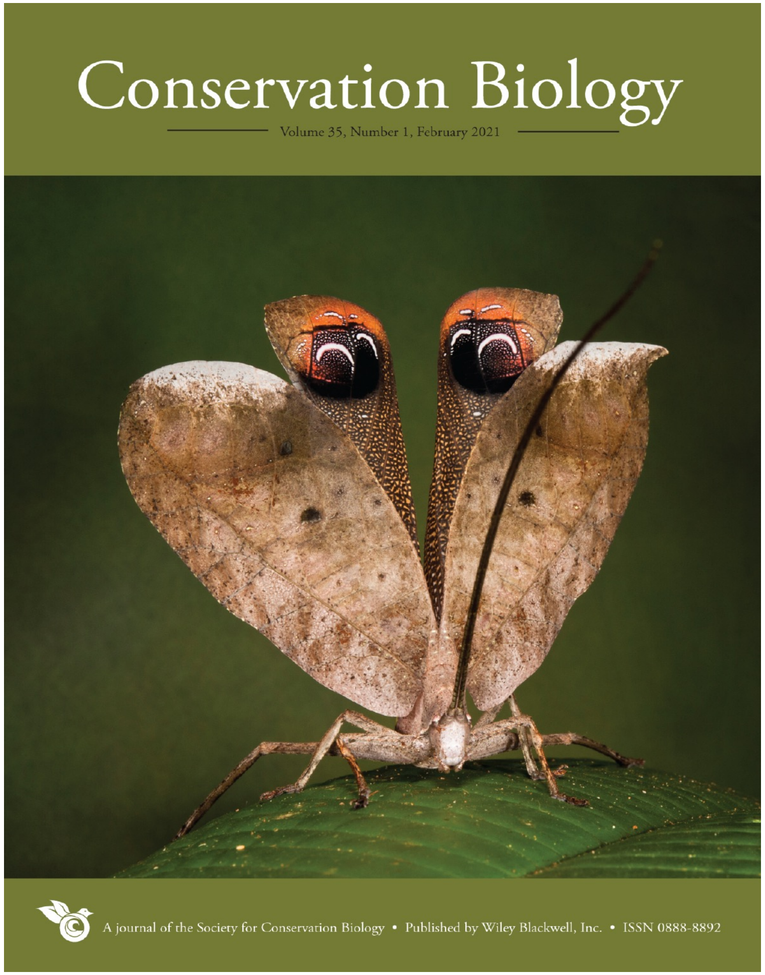
What should applied science journal editors do about statistical controversies?

Mark Burgman

Centre for Environmental Policy

Imperial College London



Conservation Biology

Volume 35, Number 1, February 2021

A journal of the Society for Conservation Biology • Published by Wiley Blackwell, Inc. • ISSN 0888-8892

Imperial College London

Protecting sugar gliders in old growth forests

http://vitalpethealth.co.uk/sugar-glider-fun-facts/

https://www.greenleft.org.au/content/will-leadbeaters-possum-survive-extended-forest-agreement

https://theconversation.com/sending-leadbeaters-possum-down-the-road-to-extinction-11249

Is modified timber harvesting compatible with the conservation of tree-dwelling mammals?

Four sites in harvested forest, four in undisturbed forest
Measured reproduction and survival over three years.

$p(H_0) = 0.38$

Conclusion: the practice is compatible with conservation

Null Hypothesis Statistical Tests

Errors

When we look at H_0 and H_1 there are two potential errors:

A **type I error** is the incorrect rejection of a true null hypothesis.

e.g. Concluding a site is unsuitable for wildlife when harvesting has no important impact

A **type II error** is the failure to reject a false null hypothesis.

e.g. Concluding there is no effect of forest harvesting when there is an important impact

The special problem of Type I and Type II errors in crisis disciplines.

Is there a problem in conservation science?

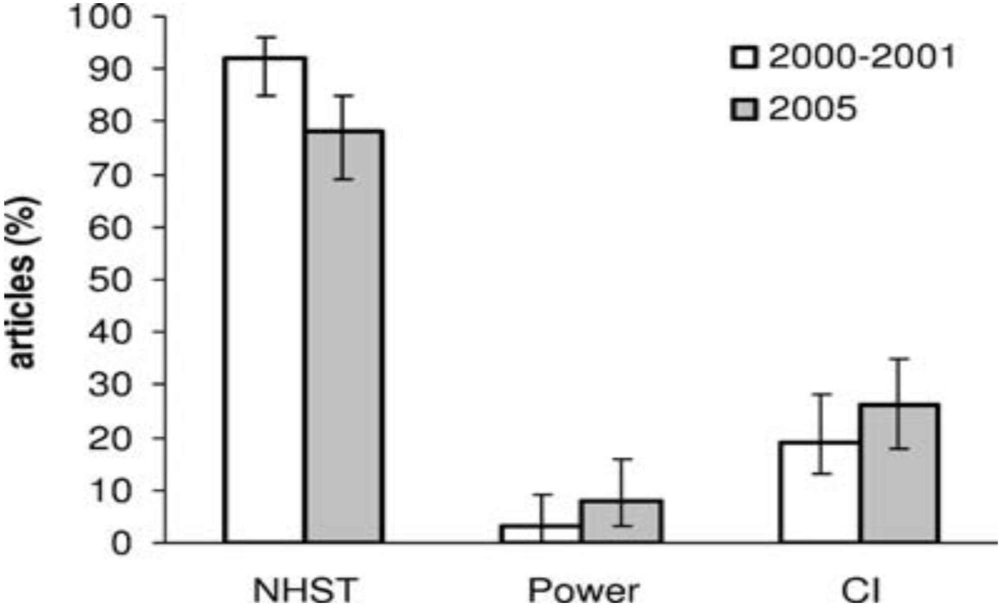
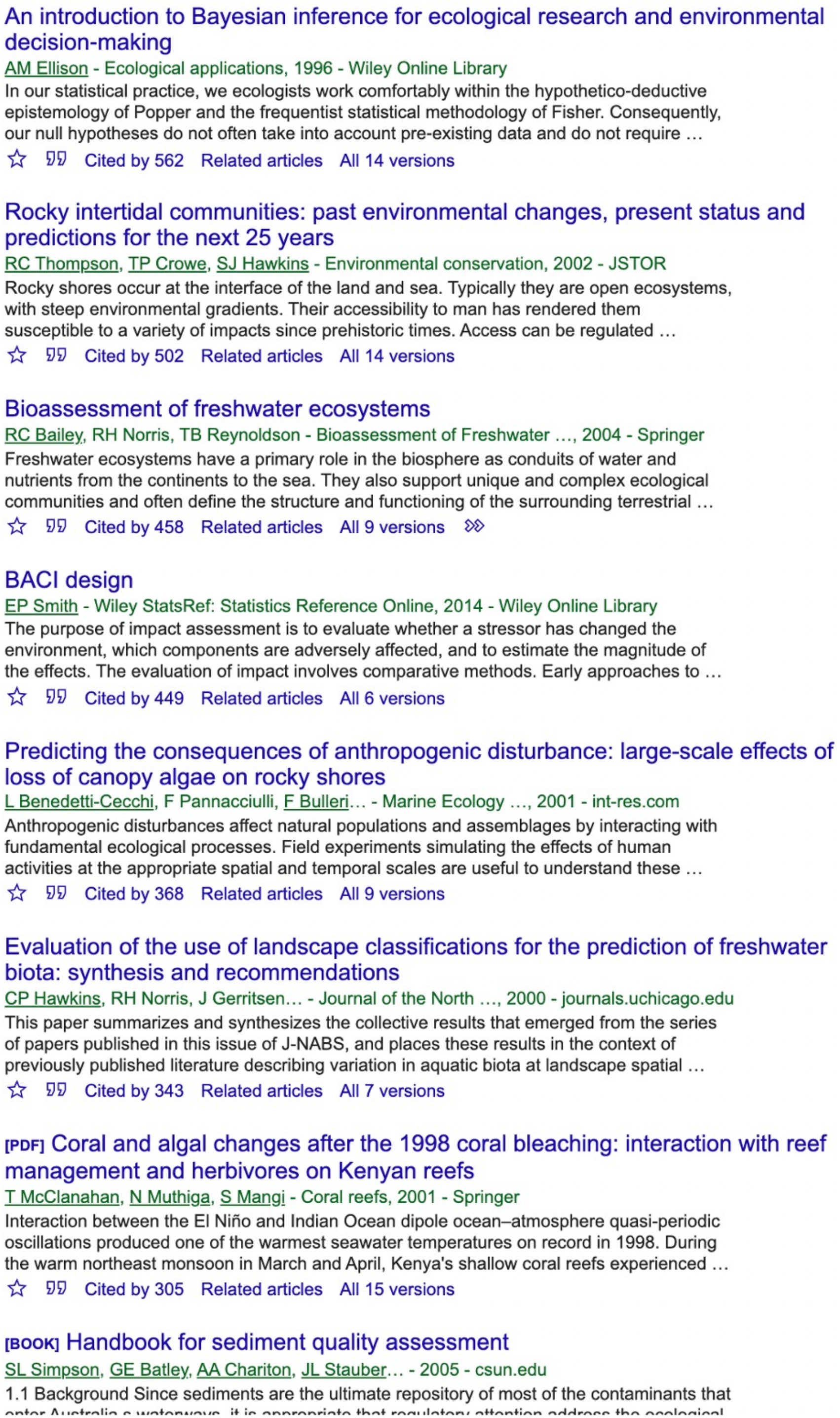


Figure 2. Percentage of Conservation Biology *and* Biological Conservation *in 2000–2001 versus the percentage of articles in 2005 that reported null-hypothesis testing (NHST), statistical power (power), and confidence intervals (CI). Error bars are 95% CIs.*

Table 1. Percentage of articles (and 95% CIs) in *Conservation Biology* and *Biological Conservation* with statistical significance tests, confidence intervals, and figures.

		Conservation Biology *and* Biological Conservation *(2001 and 2002)*		Conservation Biology *and* Biological Conservation *(2005)*	
	Good practice[a]*?*	*articles,% (n)*	*95% CI (%)*	*articles,% (n)*	*95% CI (%)*
Any statistical significance test		92 (92/100)	85–96	78 (78/100)	69–85
nil null hypothesis[b]	×	79 (73/92)	70–86	97 (76/78)	91–99
ambiguous use of *significant*[c]	×	68 (63/92)	58–77	63 (49/78)	52–73
exact *p* value[d]	√	62 (56/92)	51–70	69 (54/78)	58–78
p value asterisks (i.e., *, **, ***)[e]	×	25 (23/92)	17–35	22 (17/78)	14–32
nonsignificant result		80 (74/92)	71–87	86 (67/78)	77–92
statistical power	√√	3 (2/74)	0–9	8 (5/67)	3–16
indirect reference to power[f]	√	30 (22/74)	21–41	30 (20/67)	20–42
interpret as "no effect"	×	47 (35/74)	35–57	63 (42/67)	51–73
Any confidence interval	√√	19 (19/100)	13–28	26 (26/100)	18–35
interpreted confidence interval[g]	√√√	26 (5/19)	12–49	31 (8/26)	17–50
Any figure with data	√√	77 (77/100)	68–84	69 (69-100)	59–77
error bars on figure[b]	√√√	40 (31/77)	30–51	51 (35-69)	39–62



An introduction to Bayesian inference for ecological research and environmental decision-making
AM Ellison - Ecological applications, 1996 - Wiley Online Library
In our statistical practice, we ecologists work comfortably within the hypothetico-deductive epistemology of Popper and the frequentist statistical methodology of Fisher. Consequently, our null hypotheses do not often take into account pre-existing data and do not require …
☆ ❞ Cited by 562 Related articles All 14 versions

Rocky intertidal communities: past environmental changes, present status and predictions for the next 25 years
RC Thompson, TP Crowe, SJ Hawkins - Environmental conservation, 2002 - JSTOR
Rocky shores occur at the interface of the land and sea. Typically they are open ecosystems, with steep environmental gradients. Their accessibility to man has rendered them susceptible to a variety of impacts since prehistoric times. Access can be regulated …
☆ ❞ Cited by 502 Related articles All 14 versions

Bioassessment of freshwater ecosystems
RC Bailey, RH Norris, TB Reynoldson - Bioassessment of Freshwater …, 2004 - Springer
Freshwater ecosystems have a primary role in the biosphere as conduits of water and nutrients from the continents to the sea. They also support unique and complex ecological communities and often define the structure and functioning of the surrounding terrestrial …
☆ ❞ Cited by 458 Related articles All 9 versions ≫

BACI design
EP Smith - Wiley StatsRef: Statistics Reference Online, 2014 - Wiley Online Library
The purpose of impact assessment is to evaluate whether a stressor has changed the environment, which components are adversely affected, and to estimate the magnitude of the effects. The evaluation of impact involves comparative methods. Early approaches to …
☆ ❞ Cited by 449 Related articles All 6 versions

Predicting the consequences of anthropogenic disturbance: large-scale effects of loss of canopy algae on rocky shores
L Benedetti-Cecchi, F Pannacciulli, F Bulleri… - Marine Ecology …, 2001 - int-res.com
Anthropogenic disturbances affect natural populations and assemblages by interacting with fundamental ecological processes. Field experiments simulating the effects of human activities at the appropriate spatial and temporal scales are useful to understand these …
☆ ❞ Cited by 368 Related articles All 9 versions

Evaluation of the use of landscape classifications for the prediction of freshwater biota: synthesis and recommendations
CP Hawkins, RH Norris, J Gerritsen… - Journal of the North …, 2000 - journals.uchicago.edu
This paper summarizes and synthesizes the collective results that emerged from the series of papers published in this issue of J-NABS, and places these results in the context of previously published literature describing variation in aquatic biota at landscape spatial …
☆ ❞ Cited by 343 Related articles All 7 versions

[PDF] Coral and algal changes after the 1998 coral bleaching: interaction with reef management and herbivores on Kenyan reefs
T McClanahan, N Muthiga, S Mangi - Coral reefs, 2001 - Springer
Interaction between the El Niño and Indian Ocean dipole ocean–atmosphere quasi-periodic oscillations produced one of the warmest seawater temperatures on record in 1998. During the warm northeast monsoon in March and April, Kenya's shallow coral reefs experienced …
☆ ❞ Cited by 305 Related articles All 15 versions

[BOOK] Handbook for sediment quality assessment
SL Simpson, GE Batley, AA Chariton, JL Stauber… - 2005 - csun.edu
1.1 Background Since sediments are the ultimate repository of most of the contaminants that enter Australia's waterways, it is appropriate that regulatory attention address the ecological

Claims of some disadvantages of NHST for environmental science

- It implies it is important to avoid declaring an impact when there is none, but less important to avoid declaring an activity is benign when there is an impact (it implies that Type 1 errors matter more than Type 2 errors)
- Thresholds for statistical significance usually are unrelated to biologically important thresholds
- Poor survey, monitoring and testing procedures reduce apparent impacts.

Challenges: Intuitions about *p* values

- You repeat a study 15 times, and calculate *p* values for each.

- Should you conclude there is no effect?

A. Yes

B. No

C. There's not enough data

D. Don't know

.315
.042*
.557
.059
.308
.000***
.001***
.923
.000***
.039*
.139
.230
.016*
.321
.002**

Some history in quotes

1960s

- Traditional null-hypothesis significance-testing is "... no longer a sound or fruitful basis for statistical investigation" Clark (1963)

1970s

- "I'm not ...nit-picking... I am saying that the whole business [NHST] is so radically defective as to be scientifically almost pointless" (Meehl, 1978).
- "... the reason students have problems understanding [NHST] is that they may be trying to think." Deming (1975)
- "... significance testing should be eliminated; ...it is ... harmful..." Carver (1978)

1980s

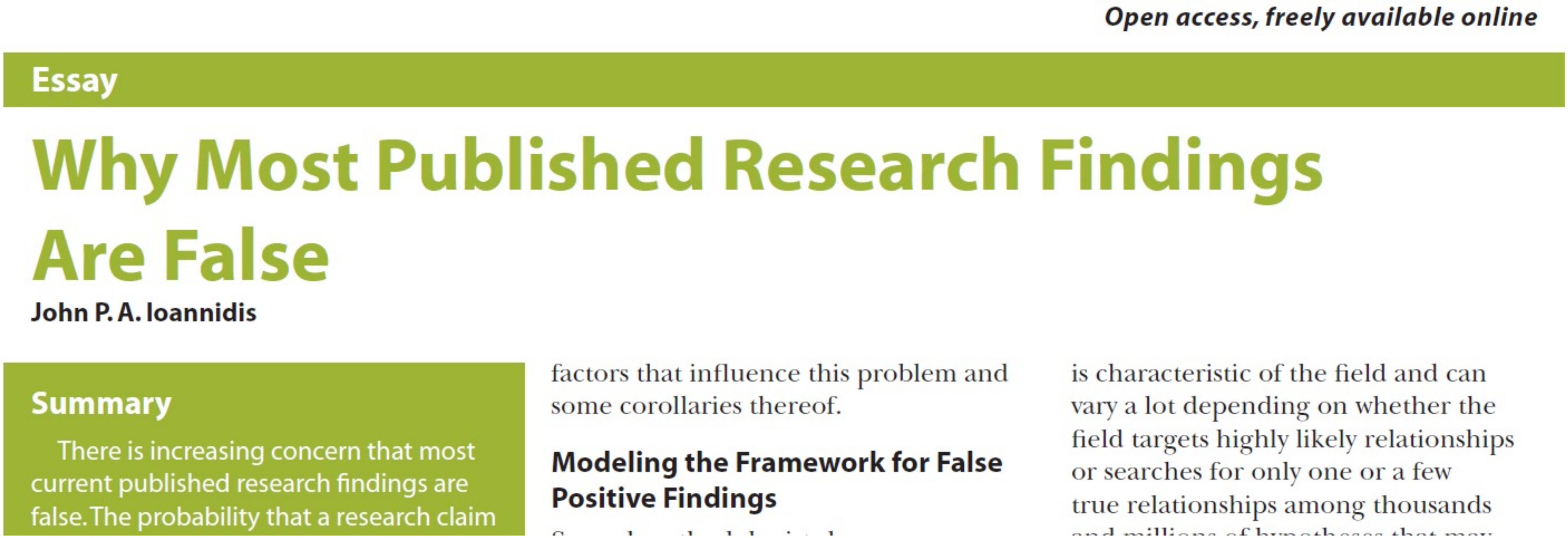
- "despite two decades of … attacks, the mystifying doctrine of null hypothesis testing is still today the Bible" (Gigerenzer & Murray, 1987).

1990s

- "... hypothesis testing does not tell us what we want to know … out of desperation, we nevertheless believe that it does." Cohen (1994)

2000s

- "...null hypothesis testing can actually impede scientific progress." (Kirk, 2003).



Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands

- The replication crisis (the 'reproducibility revolution')

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine 2*: e124. tiny.cc/mostfalse

The Ioannidis argument

The imperative to achieve statistical significance ($p < .05$) explains:

1. selective publication—file drawer
2. data selection, tweaking, and *p*-hacking until *p* is sufficiently small
3. why we think any finding that once meets the criterion of statistical significance is true and doesn't require replication

Not only are many false positives published

BUT many, perhaps even most, published findings are false

Data selection, tweaking, *p*-hacking
'Questionable research practices' (QRPs)

p hacking—it's very easy to:

- test a few extra participants
- drop or add dependent variables
- select which comparisons to analyze
- drop some results as aberrant
- try a few different statistical analysis strategies
- then finally choose which of all the above to report

RDFs
Researcher degrees of freedom

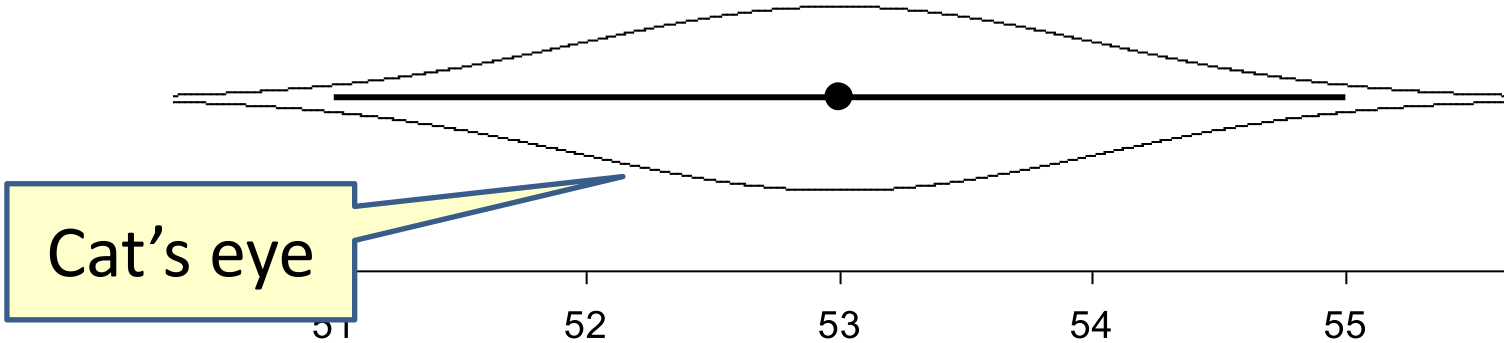
What should an editor recommend?

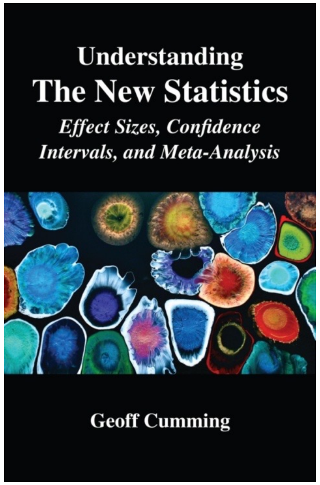
- Plot the data
- Stipulate a biologically important effect size *a priori*
- Calculate the power of any NHS tests *a priori*
- Calculate and display the size of the effect and its confidence interval (i.e. visually, and focus on estimation)
- Don't ask redundant or self-evident questions
- Pre-register all studies
- Don't use p-values?

The 'new' statistics: inference by eye

Research question: "How large is the effect of … on …?"

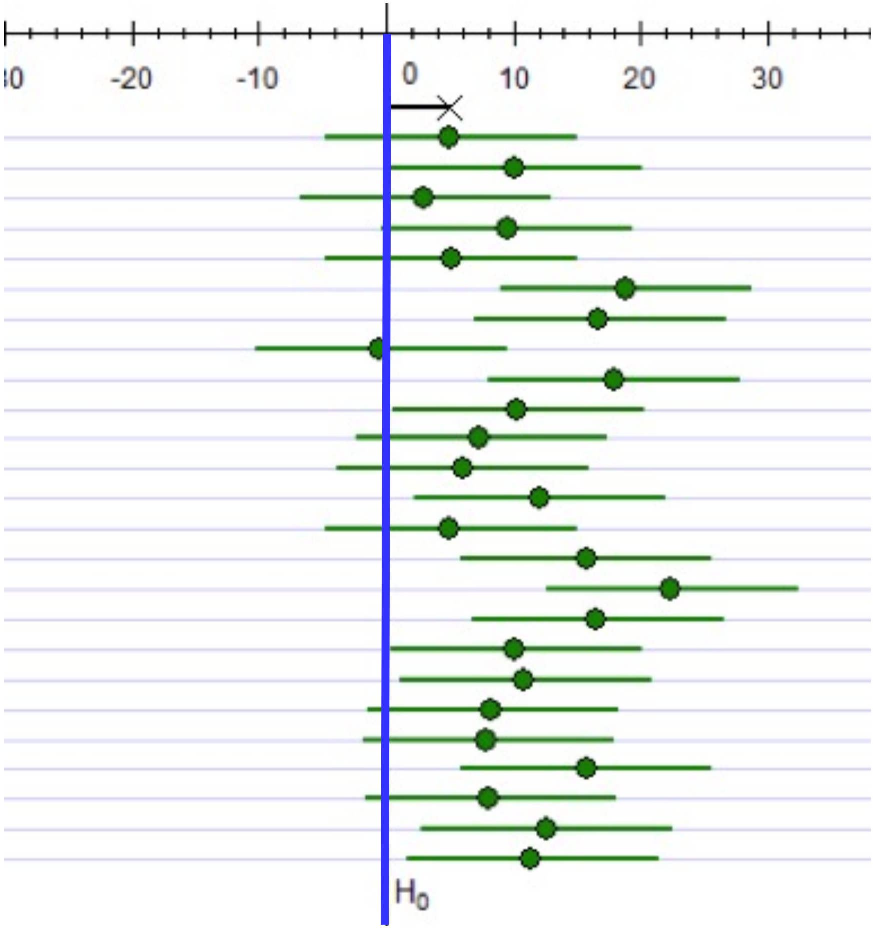
The best answer: A 95% confidence interval





Imperial College London

.315
.042*
.557
.059
.308
.000***
.001***
.923
.000***
.039*
.139
.230
.016*
.321
.002**



- Should you conclude there is no effect?

Individual judgements

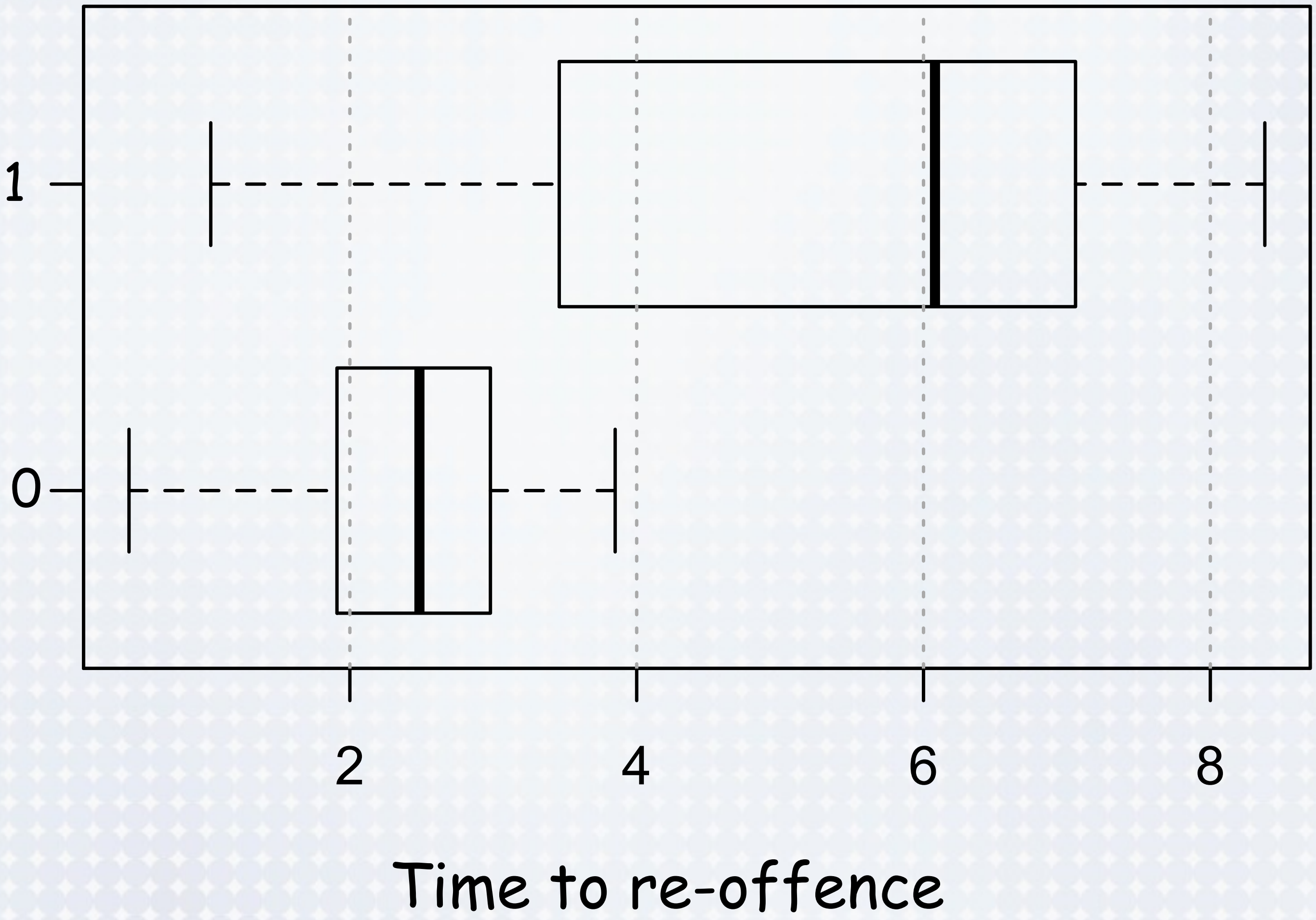
A. Yes

B. No

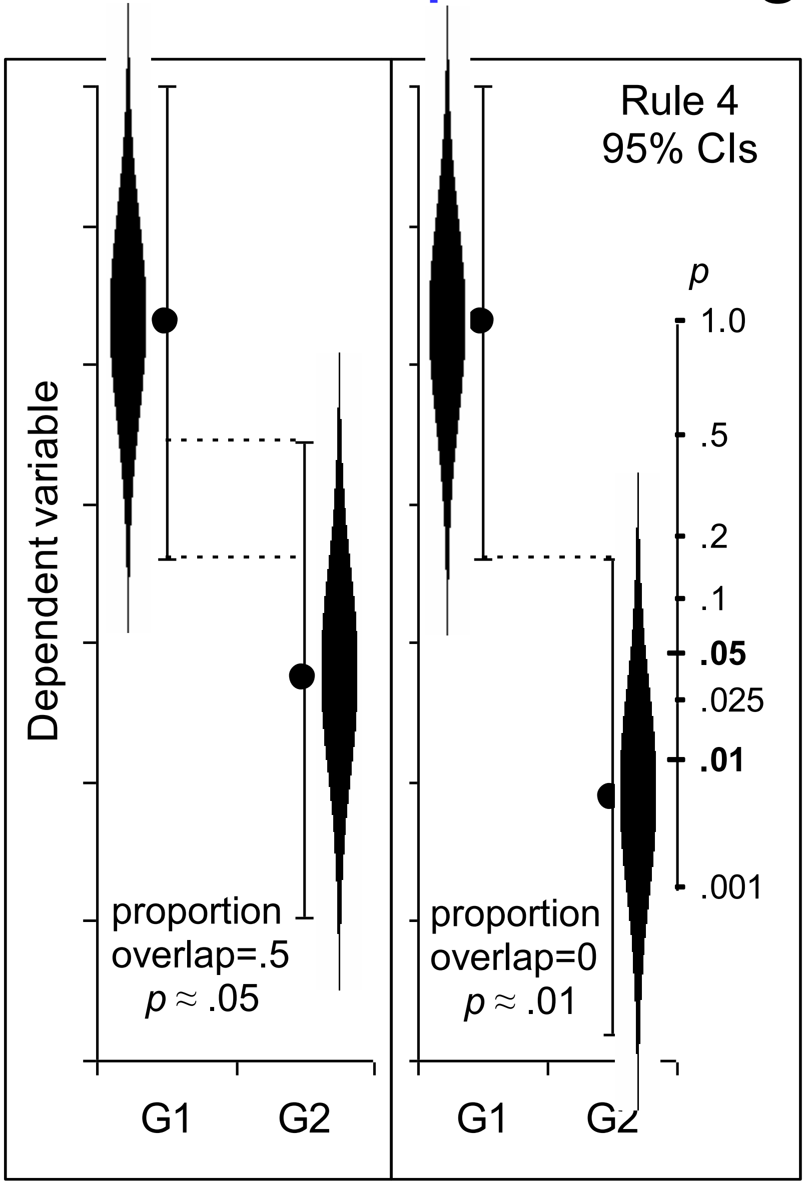
C. Not enough data

D. Don’t know

Inference on the difference of means



Two independent groups: Rules of Eye



- Two 95% CIs just touching (zero overlap) indicates moderate evidence of a population difference (approx p = .01)

- Moderate overlap (about half average MoE) is some evidence of a difference (approx p = .05)

- *When both samples sizes are at least 10, and the two MoEs do not differ by more than a factor of 2.*

- Use the rule without reference to p

Cumming, G. (2009). Inference by eye: Reading the overlap of independent confidence intervals. *Statistics in Medicine, 28*, 205-220.

Imperial College London

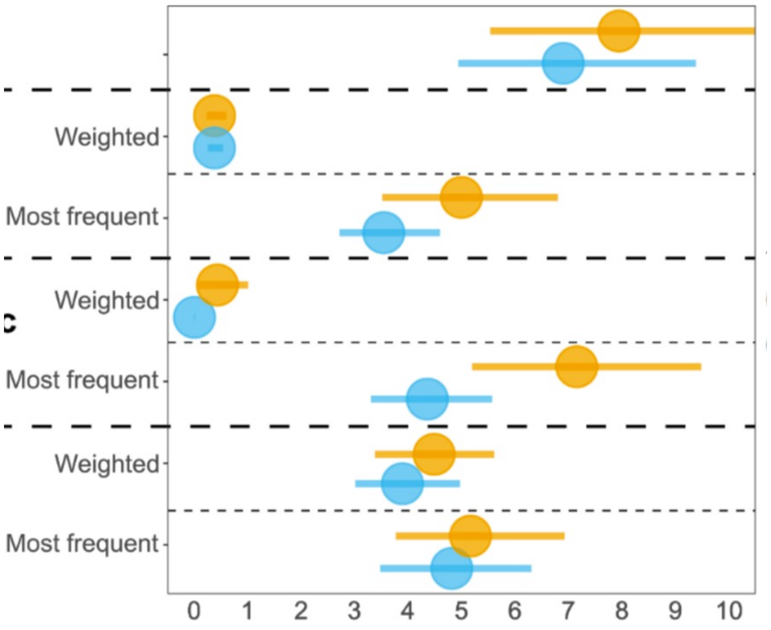
Paper 1.

using multinomial log-linear models, … XXX actions addressed fewer components than XXX actions (2.7 ± 0.3 vs. 4.5 ± 0.2, respectively)

with longer standing XXX … tending to address more components (mean ±SE: IR = 5.0 ± 0.3, BC = 4.3 ± 0.3, GD = 3.5 ± 0.5, RW = 2.9 ± 0.8, AM = 1.1 ± 0.3).

Paper 2.

Mean number of XXX, based on whether they were the most frequently studied or from a weighted list.



Imperial College London

ECOLOGICAL APPLICATIONS
ECOLOGICAL SOCIETY OF AMERICA

Article

Scalable Decision Rules for Environmental Impact Studies: Effect Size, Type I, and Type II Errors

Bruce D. Mapstone

First published: 01 May 1995 | https://doi.org/10.2307/1942031 | Citations: 128

"I propose a procedure by which the critical Effect Size is given primacy.

Statistical decision criteria are then selected according to the relative weighting of the perceived consequences of Type I or Type II errors."

Our solution

Articles reporting a NHST should be pre-registered.

Articles reporting p values for NHSTs should also report confidence intervals (CIs) for each estimate, in the text and in figures.

In all cases, authors must be explicit about what error intervals represent.