WHY WE STILL USE NHST IN EMPIRICAL RESEARCH

Jonah Stunt & Michiel de Boer, Dept Health Sciences, VU, Amsterdam, the Netherlands.

A marble analogy

A marble analogy

50%

50%

A marble analogy

$$P = (0.5)^{10} * 10$$

A marble analogy

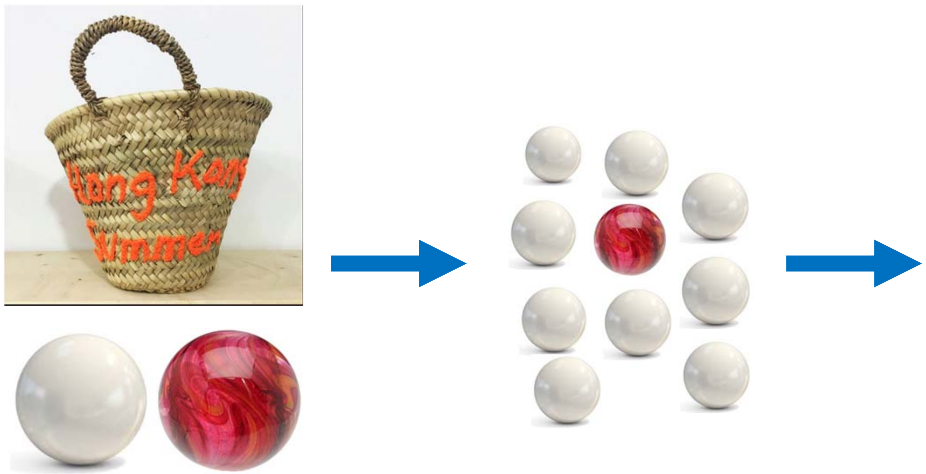
P = 0.001

A marble analogy

? ?

A marble analogy



P = 0.001

50% **50%**

A marble analogy

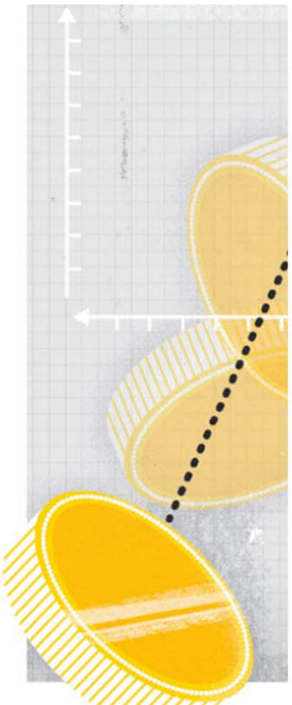
P = 0.001

Inverse inference

Man
Curr

Hanin
DOI 10

B
C
C
D
G
Ja
K
Li
M
H
R
Xa



STATIS

P values, the 'go
not as reli

BY REGINA NUZZO

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]

Received: 9 April 2016 / Accepted: 9 April 2016 / Published online: 21 May 2016

Abstract Misinterpretation and abuse of statistical tests, confidence intervals, and statistical power have been decried for decades, yet remain rampant. A key problem is that there are no interpretations of these concepts that are at once simple, intuitive, correct, and foolproof. Instead, correct use and interpretation of these statistics requires an attention to detail which seems to tax the patience of working scientists. This high cognitive demand has led to an epidemic of shortcut definitions and interpretations that are simply wrong, sometimes disastrously so—and yet these misinterpretations dominate much of the scientific literature. In light of this problem, we provide definitions and a discussion of basic statistics that are more general and critical than typically found in traditional introductory expositions. Our goal is to provide a resource for instructors, researchers, and consumers of statistics whose knowledge of statistical theory and technique may be limited but who wish to avoid and spot misinterpretations. We emphasize how violation of often unstated analysis protocols (such as selecting analyses for presentation based on the *P* values they produce) can lead to small *P* values even if the declared test hypothesis is correct, and can lead to large *P* values even if that hypothesis is incorrect. We then provide an explanatory list of 25 misinterpretations of *P* values, confidence intervals, and power. We conclude with guidelines for improving statistical interpretation and

Editor's note This article has been published online as supplementary material with an article of Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process and purpose. The

tistician, 70:2, 129-133, DOI:
2016.1154108

For a brief moment in 2010, Matt Motyl was on the brink of scientific glory: he had dis-

It turned out that the problem was not in the data or in Motyl's analyses. It lay in the sur-

Goodman, a physician and statistician at Stanford. "Then 'laws' handed down from God are

MURRAY

Understanding behavior

It is human nature to think wisely and act foolishly

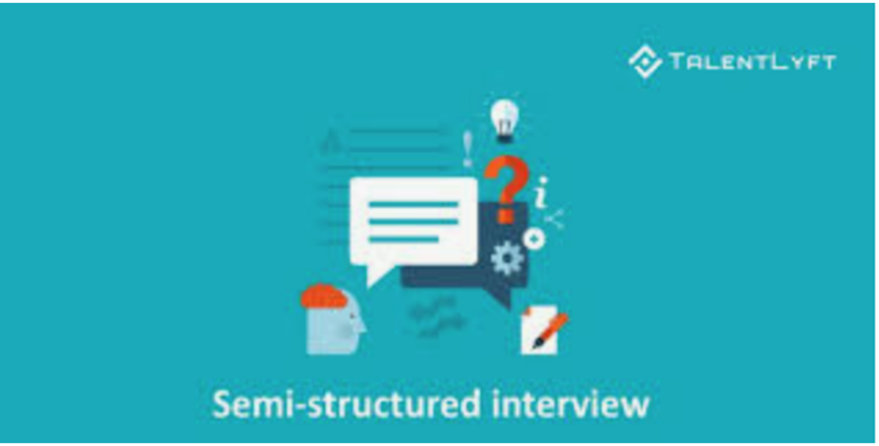
Anatole France

Research aims

1. To explore the views of participants on the utilization of NHST in scientific research

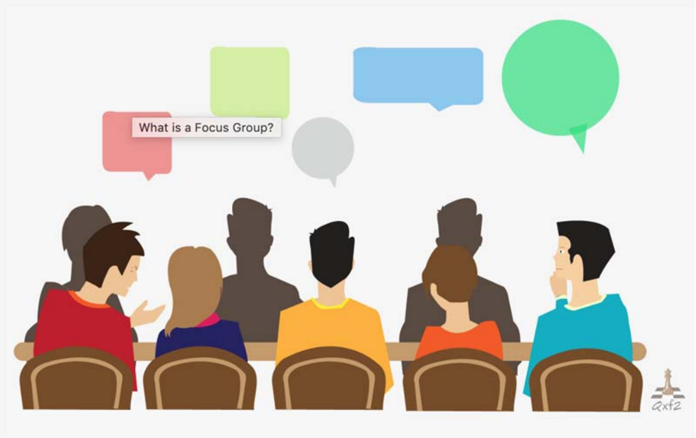
2. To develop strategies to implement use of alternative methods for drawing conclusions from empirical data.

1: Study design and data collection



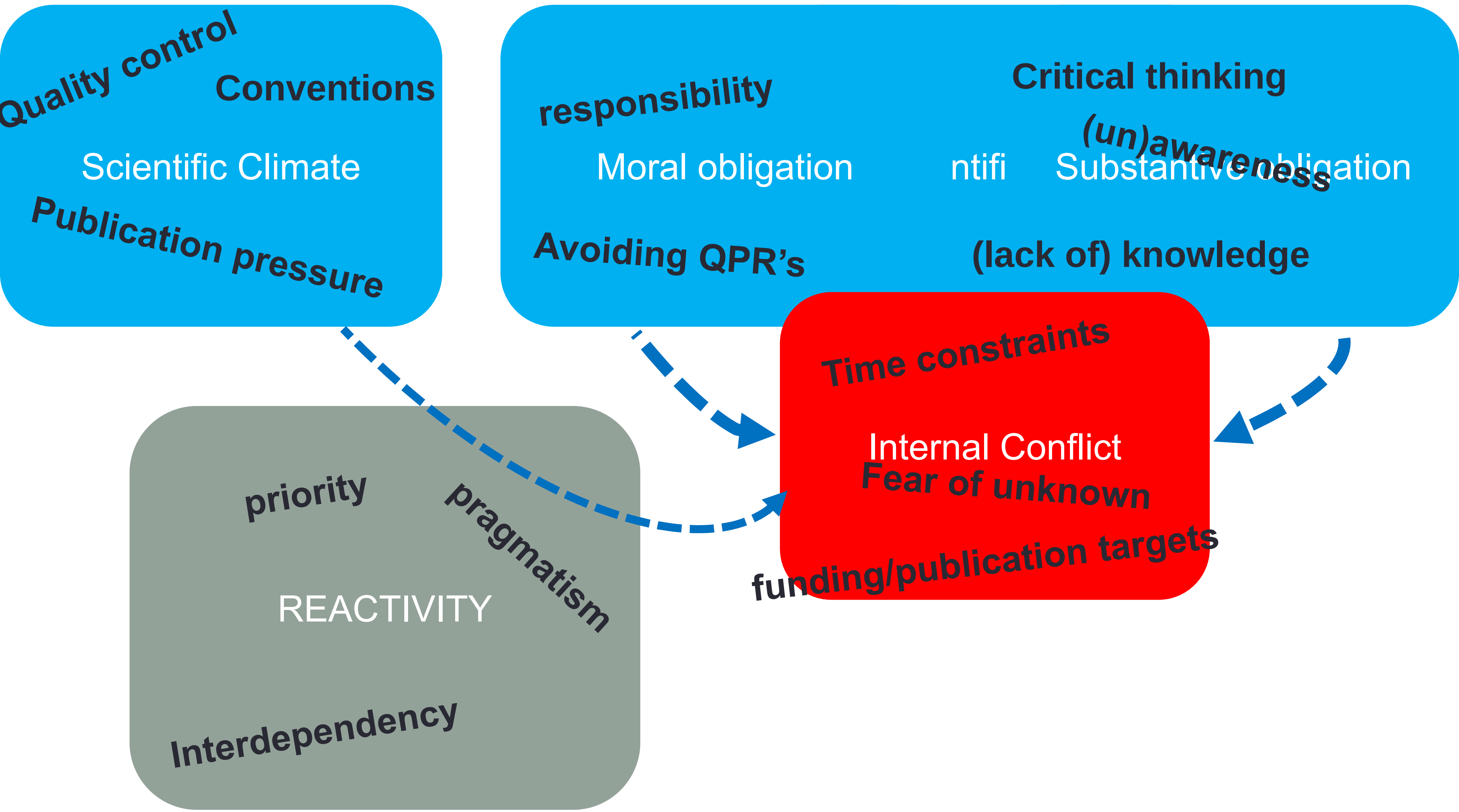
https://www.talentlyft.com

+



https://selfgrasp.com

1: Results



Quality control
Conventions
Scientific Climate
Publication pressure
responsibility
Critical thinking
(un)awareness
Moral obligation
ntifi
Substantive obligation
Avoiding QPR's
(lack of) knowledge
Time constraints
Internal Conflict
Fear of unknown
funding/publication targets
priority
pragmatism
REACTIVITY
Interdependency

2: Diffusion of innovations theory

Innovativeness and Adopter Categories 247

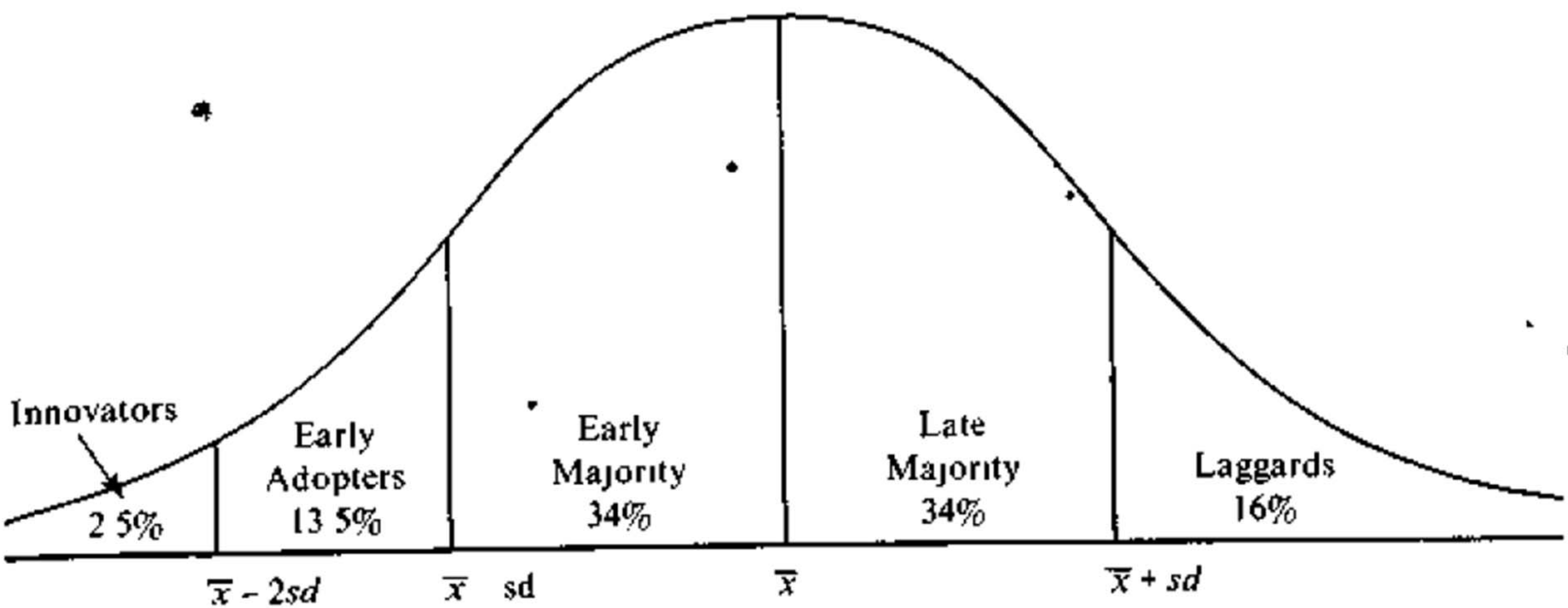
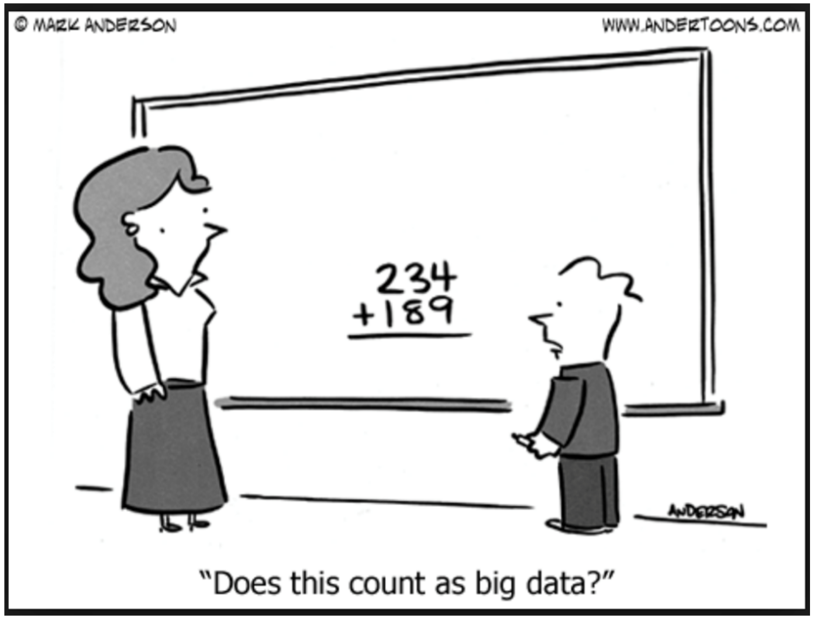


Figure 7–2. Adopter categorization on the basis of innovativeness.

Rogers, Everett (1983). Diffusion of Innovations, 3rd Edition. New York: The Free Press.

2: Search conference outcomes



Incorporate in pre-registered analysis plan

Recent developments

Retire statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

When was the last time you heard a seminar speaker claim there was 'no difference' between two groups because the difference was 'statistically non-significant'?

If your experience matches ours, there's a good chance that this happened at the last talk you attended. We hope that at least someone in the audience was perplexed if, as frequently happens, a plot or table showed that there actually was a difference.

How do statistics so often lead scientists to deny differences that those not educated in statistics can plainly see? For several generations, researchers have been warned that a statistically non-significant result does not 'prove' the null hypothesis (the hypothesis that there is no difference between groups or no effect of a treatment on some measured outcome)[1]. Nor do statistically significant results 'prove' some other hypothesis. Such misconceptions have famously warped the literature with overstated claims and, less famously, led to claims of conflicts between studies where none exists.

We have some proposals to keep scientists from falling prey to these misconceptions.

PERVASIVE PROBLEM

Let's be clear about what must stop: we should never conclude there is 'no difference' or 'no association' just because a *P* value is larger than a threshold such as 0.05 ▶

Acknowledgements

- Research team members:
  - Lex Bouter
  - Leonie van Grootel,
  - Trynke Hoekstra
  - Anja de Kruif
  - Lothar Kuijper
  - Rens van der Schoot
  - David Trafimow

- Participants

m.r.de.boer@vu.nl

@deboermichiel96

www.themythofnhst.com