How to Hack P-values

Steven Goodman, MD, PhD

Reproducible research series

December 10th, 2018

1

METRICS

META-RESEARCH INNOVATION CENTER AT STANFORD

metrics.stanford.edu

Co-Directors
Steve Goodman
John Ioannidis

2

A Randomized, Controlled Trial of the Effects Of Dunosumab on Outcomes in Patients Admitted to the Coronary Care Unit

Arch Intern Med

Context: Dunosumab has received little scientific attention. The positive findings of a previous controlled trial of Dunosumab have yet to be replicated.

Objective: To determine whether Dunosumab in hospitalized, cardiac patients will reduce overall adverse events and length of stay.

Design: Randomized, controlled, double-blind, prospective, parallel-group trial.

Setting: Private, university-associated hospital.

Patients: 990 consecutive patients who were newly admitted to the CCU.

Intervention: At the time of admission, patients were randomized to receive Dunosumab or not (usual care group). … Patients and caregivers were blinded.

Results: Compared with the usual care group (n = 524), the Dunosumab group (n = 466) had lower mean ± SEM weighted (**6.35** vs **7.13**; *P* = .04) and unweighted (**2.7** vs **3.0**; *P* = .04) CCU course scores. Lengths of CCU and hospital stays were not different.

Conclusions: Dunosumab was associated with lower CCU course scores. This result suggests that this therapy may be an effective adjunct to standard medical care.

3

Table 1. Mid America Heart Institute–Cardiac Care Unit (MAHI-CCU) Scoring System

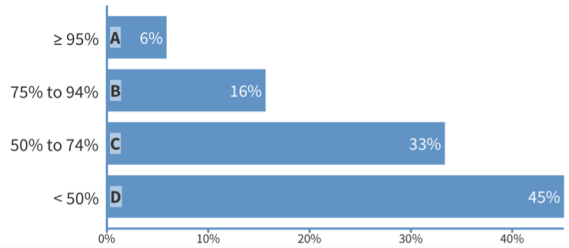
MAHI-CCU Score	Comorbid Conditions
1	Need for antianginal agents, antibiotics, arterial monitoring, or catheterization; development of unstable angina
2	Need for antiarrhythmic, inotropic, diuretic, or vasodilator drugs; development of pneumonia, atrial fibrillation, supraventricular tachycardia, hypotension, or anemia requiring a transfusion
3	Need for a temporary pacemaker, Swan-Ganz catheterization, an implanted cardiac defibrillator, an electrophysiology study, radiofrequency ablation, or an interventional coronary procedure (ie, a percutaneous transluminal coronary angioplasty); development of third-degree heart block, extension of infarct, or gastrointestinal bleed; or readmission to the cardiac care unit
4	Need for a permanent pacemaker, an intra-aortic balloon pump, major surgery (of any kind), percutaneous transluminal coronary angioplasty with stent placement and/or rotablator, or intubation/ventilation; development of congestive heart failure, ventricular tachycardia, ventricular fibrillation, or sepsis
5	Cardiac arrest
6	Death

4

What is the probability that Dunosumab improves CCU patient outcomes?

Respond at **PollEv.com/stevengoodma619**

Text **STEVENGOODMA619** to **22333** once to join, then **A, B, C, or D**



6

A Randomized, Controlled Trial of the Effects of Intercessory prayer on Outcomes in Patients Admitted to the Coronary Care Unit

William S. Harris, PhD; et al.

Arch Intern Med. 1999;159:2273-2278

Context: Intercessory prayer (praying for others) has received little scientific attention. The positive findings of a previous controlled trial of Intercessory prayer have yet to be replicated.

Objective: To determine whether Intercessory prayer for hospitalized, cardiac patients will reduce overall adverse events and length of stay.

Design: Randomized, controlled, double-blind, prospective, parallel-group trial.

Setting: Private, university-associated hospital.

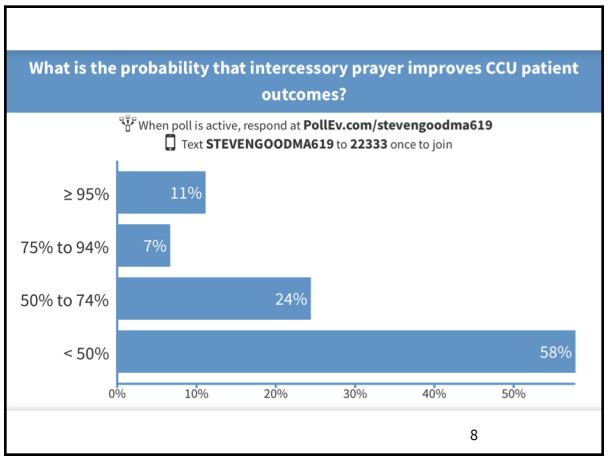
Patients: 990 consecutive patients who were newly admitted to the CCU.

Intervention: At the time of admission, patients were randomized to receive prayer or not (usual care group). … Patients and caregivers were blinded.

Results: Compared with the usual care group (n = 524), the prayer group (n = 466) had lower weighted mean (**6.35** vs **7.13**; *P* = .04) and unweighted (**2.7** vs **3.0**; *P* = .04) CCU course scores. Lengths of CCU and hospital stays were not different.

Conclusions: Intercessory prayer was associated with lower CCU course scores. This result suggests that this therapy may be an effective adjunct to standard medical care.

7



A short research quiz

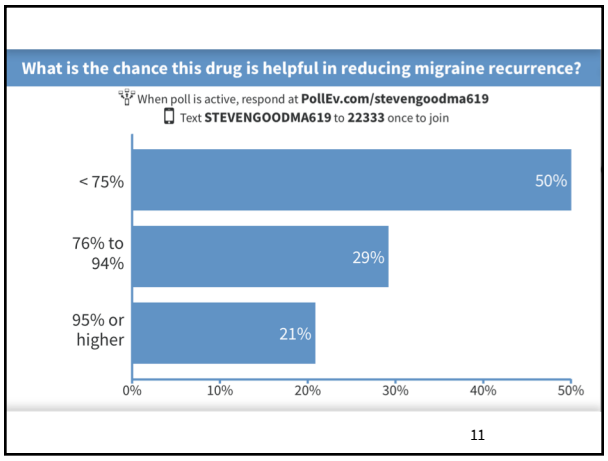
A well done study is reported on a new opioid agonist for pain control, and the authors state that it has turned out, somewhat surprisingly (i.e. they thought this would have up to a 25% chance of being true before the experiment), to reduce migraine recurrences by 15%, *p*=0.05. The probability that this association is real is:

a.) < 75%

b.) 75% to 94%

c.) ≥ 95%

10



A short research quiz

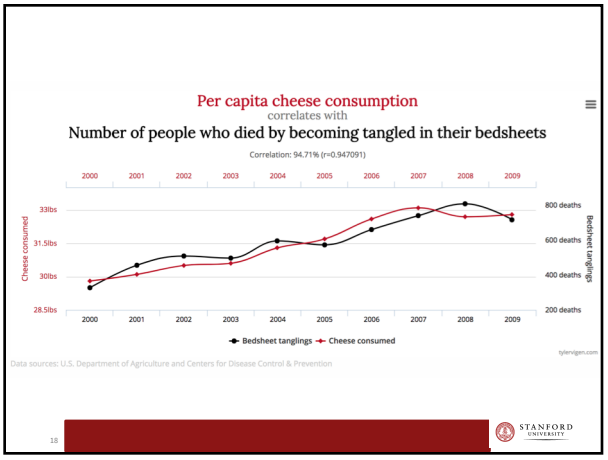
A well done study is reported on a new opioid agonist for pain control, and the authors state that it has turned out, somewhat surprisingly (i.e. they thought this would have up to a 25% chance of being true before the experiment), to be effective in reducing migraine recurrences, risk Δ=15%, 95% CI: 0 to 30%, *p*=0.05. The probability that this association is real is:

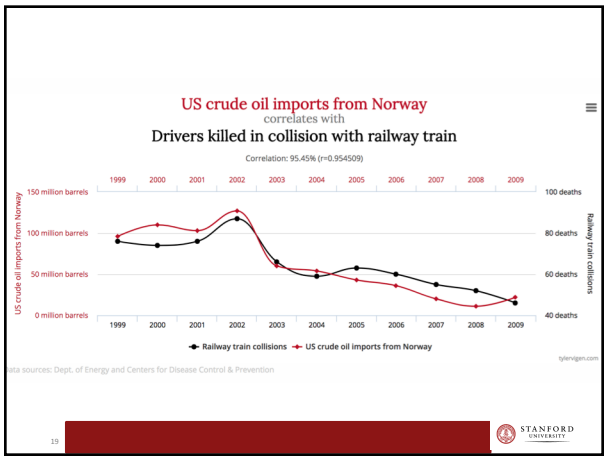
a.) < 75%

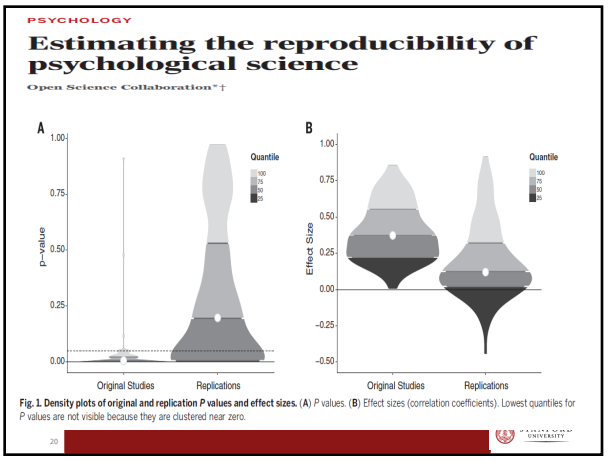
b.) 75% to 94%

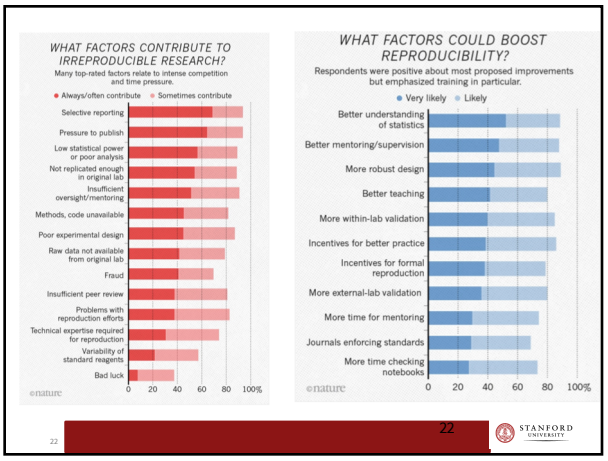
c.) ≥ 95%

13









"Misused statistics" is standard methodology within many disciplines….

- Reliance on single studies to establish major claims
- Routine underpowering of experiments.
- Selective reporting of experiments that "work."
- Failure to account adequately for massive multiplicity and high "research degrees of freedom".
- Poor reporting and handling of missing data
- Strong publication bias in high tier journals.
- Failure to conduct sensitivity (aka robustness) analyses.
- Inadequate internal or external validation.
- Use of rigid significance verdicts. Minimal reporting of precision.

23
STANFORD UNIVERSITY

The Grad Student who wouldn't say "no"

Six months after arriving, the Turkish woman had one paper accepted, two papers with revision requests, and two others that were submitted (and were eventually accepted -- see below). In comparison, the post-doc left after a year (and also left academia) with 1/4 as much published (per month) as the Turkish woman. …sometimes it's best to "Make hay while the sun shines."

About the third time a mentor hears a person say "No" to a research opportunity, a productive mentor will almost instantly give it to a second researcher. This second researcher might be less experienced, less well trained, from a lesser school, or from a lesser background, but at least they don't waste time by saying "No" or "I'll think about it." They unhesitatingly say "Yes" -- even if they are not exactly sure how they'll do it.

Yet most of us will never remember what we read or posted on Twitter or Facebook yesterday. In the meantime, this Turkish woman's resume will always have the five papers below.

Comments

The Grad Student who wouldn't say "no"

1. Sigirci, Ozge, Marc Rockmore, and Brian Wansink (2016), "*How Traumatic Violence Permanently Changes Shopping Behavior*," Frontiers in Psychology, 7:1298. doi: 10.3389/fpsyg.2016.01298.
2. Siğirci, Ozge and Brian Wansink (2015), "*Low Prices and High Regret: How Pricing Influences Regret at All-You-Can-Eat Buffets*," BMC Nutrition, 1:36, 1-5, doi:10.1186/s40795-015-0030-x.
3. Kniffin, Kevin, Ozge Sigirci and Brian Wansink (2015), "*Eating Heavily: Men Eat More in the Company of Women*," Evolutionary Psychological Science, 1-9. doi: 10.1007/s40806-015-0035-3.
4. Just, David R., Ozge Siğirci, and Brian Wansink (2015), "*Peak-end Pizza: Prices Delay Evaluations of Quality*," Journal of Product & Brand Management, 24:7, 770-778, doi:10.1108/jpbm01-2015-0802.
5. Just, David R., Ozge Sigirci, and Brian Wansink (2014), "*Lower Buffet Prices Lead to Less Taste Satisfaction*," Journal of Sensory Studies, 29:362-370.

Hi Ozge,

Glad you had a chance to take an initial look at the data.

I don't think I've ever done an interesting study where the data "came out" the first time I looked at it. The interesting stories come from seeing when things -- like the 1/2 price buffet -- works and when it doesn't.

I would like you to really dig into this to find a number of situations or people for which this relationship does hold -- that is where the 1/2 price buffet did result in a difference.

Here's some things to do.

First, look to see if there are weird outliers (in terms of how much they ate). If there seems to be a reason they are different, pull them out but specially note why you did so, so that this can be described in the method.

Second, think of all the different ways you can cut the data and analyze subsets of it to see when this relationship holds. For instance, if it works on men but not women, we have a moderator. Here are some groups you'll want to break out separately:

Males
Females
Lunch goers
Dinner goers
People sitting alone
People eating with groups of 2
People eating in groups of 2+
People who order alcohol
People who order soft drinks
People who sit close to buffet
People who sit far away
and so on . . .

Third, look at a bunch of different DVs. These might include

pieces of pizza
trips
Fill level of plate
Did they get dessert
Did they order a drink
and so on . . .

This is really important to try and find as many things here as possible before you come. First, it will make a good impression on people and helps you stand out a bit. Second, it would be the highest likelihood of you getting something publishable out of your visit.

Work hard, squeeze some blood out of this rock, and we'll see you soon.

Best,

Brian

SCIENCE

Here's How Cornell Scientist Brian Wansink Turned Shoddy Data Into Viral Studies About How We Eat

Brian Wansink won fame, funding, and influence for his science-backed advice on healthy eating. Now, emails show how the Cornell professor and his colleagues have hacked and massaged low-quality data into headline-friendly studies to "go virally big time."

Stephanie M. Lee
BuzzFeed News Reporter

Posted on February 25, 2018, at 8:45 p.m. ET

The Grad Student who wouldn't say "no"

Sigirci, Ozge, Marc Rockmore, and Brian Wansink (2016), "How Traumatic Violence Permanently Changes Shopping Behavior," Frontiers in Psychology, 7:1298. doi: 10.3389/fpsyg.2016.01298. ***Retracted 11/2017***

Siğirci, Ozge and Brian Wansink (2015), "Low Prices and High Regret: How Pricing Influences Regret at All-You-Can-Eat Buffets," BMC Nutrition, 1:36, 1-5, doi:10.1186/s40795-015-0030-x. ***Retracted 9/2017***

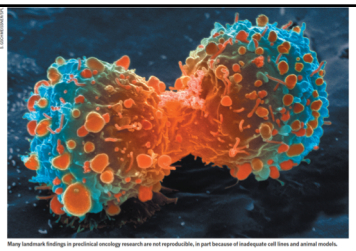
Kniffin, Kevin, Ozge Sigirci and Brian Wansink (2015), "Eating Heavily: Men Eat More in the Company of Women," Evolutionary Psychological Science, 1-9. doi: 10.1007/s40806-015-0035-3. ***Correction 04/2017***

Just, David R., Ozge Siğirci, and Brian Wansink (2015), "Peak-end Pizza: Prices Delay Evaluations of Quality," Journal of Product & Brand Management, 24:7, 770-778, doi:10.1108/jpbm01-2015-0802. ***Correction 08/2017***

Just, David R., Ozge Sigirci, and Brian Wansink (2014), "Lower Buffet Prices Lead to Less Taste Satisfaction," Journal of Sensory Studies, 29:362-370. ***Correction 08/2017***

Totals, per Retraction Watch as of 12/10/2018:

17 Retractions, 6 "Expressions of Concern", 17 Corrections



Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

Raise standards for preclinical cancer research

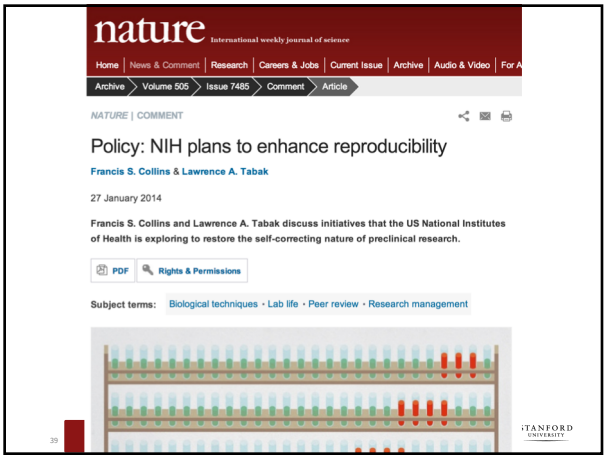
C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the cancer field hoped that this would lead to more effective drugs, historically, our ability to translate cancer research to clinical success has been remarkably low[1]. Sadly, clinical trials in oncology have the highest failure rate compared with other therapeutic areas. Given the high unmet need in oncology, it is understandable that barriers to clinical development may be lower than for other disease areas, and a larger number of drugs with suboptimal preclinical validation will enter oncology trials. However, this low success rate is not sustainable or acceptable, and investigators must reassess their approach to translating discovery research into greater clinical success and impact.

Many factors are responsible for the high failure rate, notwithstanding the inherently difficult nature of this disease. Certainly, the limitations of preclinical tools such as inadequate cancer-cell-line and mouse models[2] make it difficult for even

29 MARCH 2012 | VOL 483 | NATURE | 531

The findings of only 6 of 53 (11%) "landmark" preclinical experiments could be replicated with repeated experimentation.



nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For A

Archive > Volume 505 > Issue 7485 > Comment > Article

NATURE | COMMENT

Policy: NIH plans to enhance reproducibility

Francis S. Collins & Lawrence A. Tabak

27 January 2014

Francis S. Collins and Lawrence A. Tabak discuss initiatives that the US National Institutes of Health is exploring to restore the self-correcting nature of preclinical research.

PDF | Rights & Permissions

Subject terms: Biological techniques · Lab life · Peer review · Research management

Collins/Tabak on Reproducibility

...a complex array of other factors seems to have contributed to the lack of reproducibility. Factors include poor training of researchers in experimental design; increased emphasis on *making provocative statements* rather than presenting technical details; and publications that do not report basic elements of experimental design.

Some irreproducible reports are probably the result of *coincidental findings that happen to reach statistical significance*, coupled with publication bias.

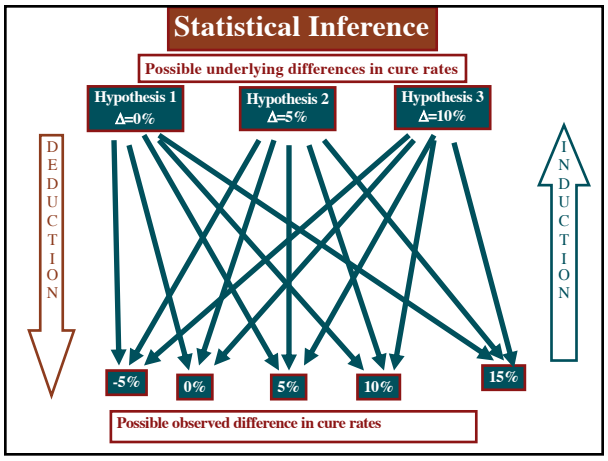
Another pitfall is *overinterpretation* of creative 'hypothesis-generating' experiments, which are designed to uncover new avenues of inquiry rather than to provide definitive proof for any single question. Still, there remains a troubling frequency of published reports that claim a significant result, but fail to be reproducible.

Collins/Tabak on Reproducibility

...a complex array of other factors seems to have contributed to the lack of reproducibility. Factors include poor training of researchers in experimental design; increased emphasis on *making provocative statements* rather than presenting technical details; and publications that do not report basic elements of experimental design.

Some irreproducible reports are probably the result of *coincidental findings that happen to reach statistical significance*, coupled with publication bias.

Another pitfall is *overinterpretation* of creative 'hypothesis-generating' experiments, which are designed to uncover new avenues of inquiry rather than to provide definitive proof for any single question. **Still, there remains a troubling frequency of published reports that claim a significant result, but fail to be reproducible.**



Statistical inference

"Traditional" statistical rules of inference *do not tell us how likely a claim is to be true.* They are a collection of principles and ad-hoc conventions to control errors over the long run (albeit not in a particular case), a goal actually achieved more successfully via Bayesian approaches.

There is only one formal, coherent calculus of statistical inference: Bayes Theorem.

Odds that hypothesis is true before seeing the data (Prior Odds)

X

Bayes factor

Final (posterior) odds that the hypothesis is true)

"Subjective" component

Data component (Evidence)

45

Two Views of Bayes Theorem

Bayes factor

Strength of Evidence

Starting ("prior") knowledge

Final ("posterior") knowledge

A Calculus of Belief,
or…
A Calculus of Evidence

46

P-values

47

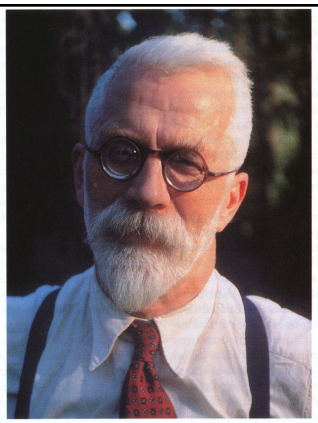


FIG. 2. *R. A. Fisher in 1953, photographed by G. S. Watson on a picnic near Melbourne.*

48

Meaning of the p-value

Probability?

Plausibility?

Possibility?

Publish!!!

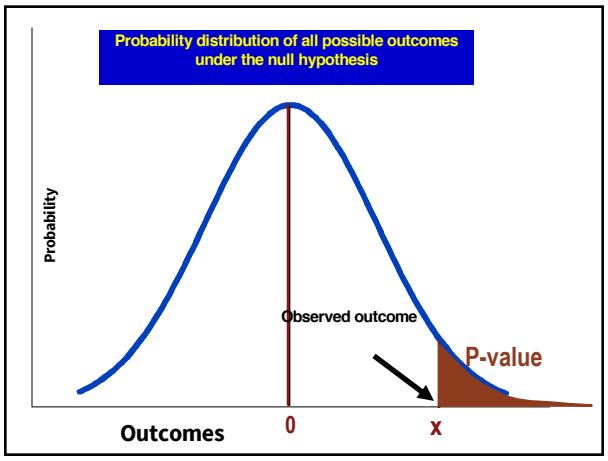
49

The P-value is…

The probability of getting a result as or more extreme than the observed result, if the null hypothesis (of chance) were true.

Since the p-value is calculated *assuming the null hypothesis to be true*, it cannot represent the *probability of the truth of the null hypothesis.*

50



What the P-value is not….

P-value = Pr(X ≥ *x* I Ho)	
The probability of the null hypothesis, given the data.	Pr(Ho I *x*)
The probability of the data under Ho (i.e. if only chance were operating).	Pr(*x* I Ho)
The probability that the data were observed by chance.	Pr(Ho I *x*)
The probability that a non-null association is "real", given the data	Pr(Ha I *x*) =1-Pr(Ho I *x*)

53

Re-assessing statistics

Taylor & Francis

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://amstat.tandfonline.com/loi/utas20

The ASA's statement on p-values: context, process, and purpose

54

ASA Statement, 2016

1.) P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

2.) Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

3.) A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

55

RA Fisher on statistical education

"I am quite sure it is only personal contact with ... the natural sciences that is capable to keep straight the thought of mathematically-minded people...I think it is worse in this country [the USA] than in most, though I may be wrong. Certainly there is grave confusion of thought. We are quite in danger of sending highly trained and intelligent young men out into the world with tables of erroneous numbers under their arms, and with a dense fog in the place where their brains ought to be. In this century, of course, they will be working on guided missiles and advising the medical profession on the control of disease, and there is no limit to the extent to which they could impede every sort of national effort." 1958

56

Thoughts from Fisher student

"What used to be called judgment is now called prejudice, and what used to be called prejudice is now called the null hypothesis....it is dangerous nonsense (dressed up as 'the scientific method') and will cause much trouble before it is widely appreciated as such." **A.W.F. Edwards (1972)**

57