Deprecating Statistical Significance: Toward Better Science

Department of Speech-Language-Hearing Sciences: Pro-Sem
April 26, 2019

Andrew Zieffler

Educational Psychology

University of Minnesota
Driven to Discover℠

A repeated-measures analysis of variance was implemented using condition as a main factor and subject group as a random blocking factor. The effects of condition and group were **significant** ($p < .001$), as was the interaction between Condition × Group, $F(9, 276) = 169$, $p = .004$.

Goldsworthy & Markle (2019)
Journal of Speech, Language, and Hearing Research

Significant positive effects of bimodal stimulation on speech perception were found in most conditions. The average summation effect values were 13% ± 14% for the BB-NALNL2 setting, 15% ± 17% for the SB-NALNL2 setting, and 9% ± 15% for the BB-Audiogram+ setting. These effects were **significantly different from zero** (one-sample Wilcoxon signed-ranks test, $Z = 84$, $p = .008$; $Z = 93$, $p = .01$; and $Z = 65$, $p = .04$, respectively). The average squelch effects were **not significantly different from zero**, except for the SB-NALNL2 condition. For this condition, the squelch effect was 9% ± 15% (one-sample Wilcoxon signed-ranks test, $Z = 84$, $p = .048$). The average head shadow effects were **not significantly different from zero**, except for the BB-NALNL2 condition. For this condition, the head shadow decreased with 12% ± 15% (one-sample Wilcoxon signed-ranks test, $Z = 90$, $p = .02$). The median of the head shadow effect was 14%. However, **no significant differences** in bimodal benefit were found between the three HA fittings for any of the measures summation, squelch, or head shadow (Friedman test: $p < .05$).

Vroegop, Dingemanse, van der Schroeff, & Goedegebure (2019)
American Journal of Audiology

TABLE 3 Pearson correlations for the variables included in the regression analysis.

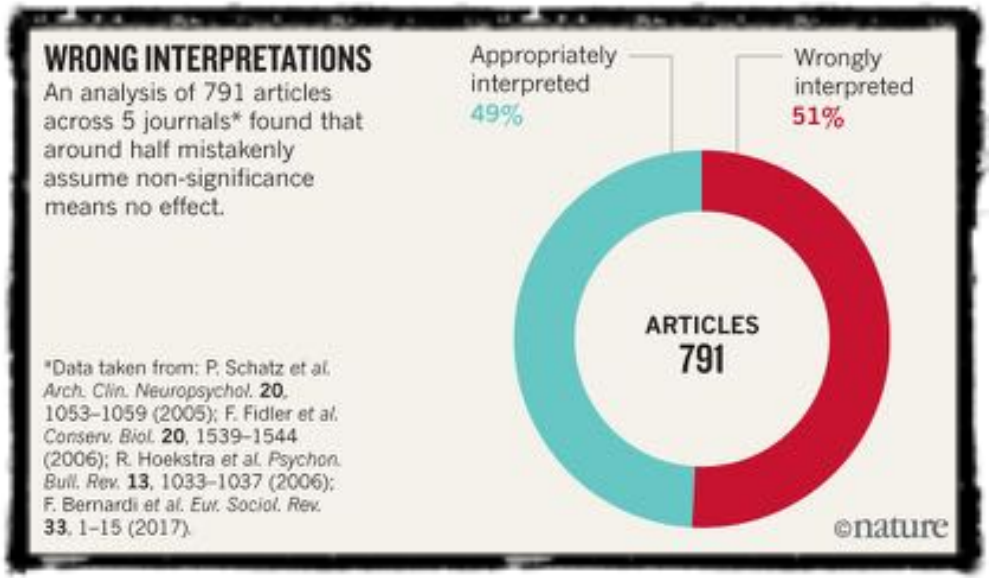
Variable	1	2	3	4	5	6	7	8
1. Communicative participation	—							
2. Problems thinking	−.536*;	—						
3. Slurred speech	−.474*;	.366*;	—					
4. Vision loss	−.336*;	.301*;	.347*;	—				
5. Pain	−.316*;	.346*;	.218*;	.235*;	—			
6. Mobility	−.300*;	.149*;	.255*;	.281*;	.193*;	—		
7. Depression	−.560*;	.521*;	.361*;	.326*;	.441*;	.246*;	—	
8. Fatigue	−.627*;	.685*;	.406*;	.369*;	.473*;	.395*;	.701*;	—
9. Social support	.281*;	−.198*;	−.183*;	−.126*;	−.221*;	−.104*;	−.373*;	−.249*;
10. Age	.019	−.041	−.045	.051	.104*;	.351*;	−.050	.085
11. Education	.045	−.139*;	−.006	−.106*;	−.077	−.146*;	−.118*;	−.116*;
12. Paid work	.254*;	−.215*;	−.133*;	−.219*;	−.133*;	−.490*;	−.134*;	−.314*;
13. Gender	.028	.056	−.103*;	−.068	.090	−.100	−.002	−.019
14. MS duration	−.020	−.046	−.003	.095	.025	.374*;	−.089	.035

*Significant at $p \leq .01$.

Baylor, Yorkston, Banner, Britton, & Amtmann (2010)
American Journal of Speech-Language Pathology

The ASA *Statement on P-Values and Statistical Significance* stopped just short of recommending that declarations of "statistical significance" be abandoned. We take that step here. We conclude, based on our review of the articles in this special issue and the broader literature, that **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.**

Wasserstein, Schirm, & Lazar (2019)
The American Statistician



We agree, and call for the entire concept of statistical significance to be abandoned. We are far from alone. When we invited others to read a draft of this comment and sign their names if they concurred with our message, 250 did so within the first 24 hours. A week later, we had **more than 800 signatories** — all checked for an academic affiliation or other indication of present or past work in a field that depends on statistical modelling.

Amrhein, Greenland, & McShane (2019)
Nature

A Brief History of Significance and Hypothesis Testing

Karl Pearson
(*1857–1936*)

"When Pearson began his statistical career, the problem of assessing whether or not standard distributions provided acceptable fits to sets of data was well known. Simple tests using normal approximations to binomial distributions were introduced by Laplace, and **comparisons of statistics with their expected values had taken over from subjective inspections of a set of frequencies**."

Plackett (1983)
International Statistical Review

- Pearson extended this work with his 1900 paper in which he developed chi-squared (χ^2) test for goodness-of-fit by comparing observed to predicted values
- Used Walter F. R. Weldon's empirical data to generalized the distribution theory and derive an appropriate test statistics (χ^2)
  - Weldon tossed 12 dice over and over again (actually 26,306 times!)
  - If a die landed as a 5 or 6 it was counted as a "success". He recorded the number of success for each toss of the 12 dice.
- Refuted much of the earlier work on the laws of errors using Weldon's empirical data

No. of Dice in Cast with 5 or 6 Points.	Observed Frequency, n'.	Theoretical Frequency, n.	Deviation, ϵ.
0	185	203	− 18
1	1149	1217	− 68
2	3265	3345	− 80
3	5475	5576	−101
4	6114	6273	−159
5	5194	5018	+176
6	3067	2927	+140
7	1331	1254	+ 77
8	403	392	+ 11
9	105	87	+ 18
10	14	13	+ 1
11	4	1	+ 3
12	0	0	0
	26306	26306	

$\chi^2 = 43.87$

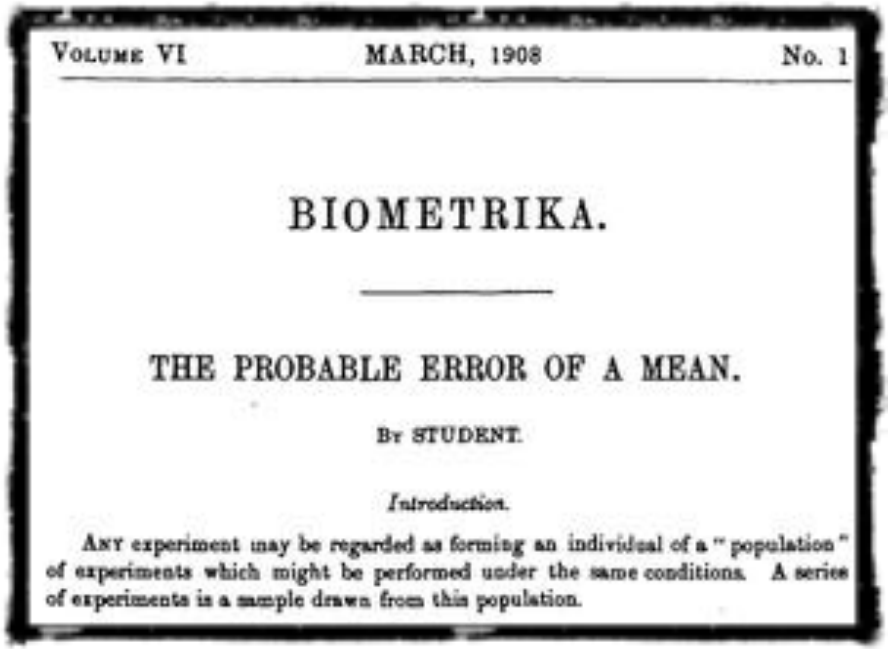
$P = .000016$

"Or the odds are 62,499 to 1 against such a system of deviations on a random selection. With such odds it would be reasonable to conclude that dice exhibit bias toward the higher points."

Pearson (1900)
Philosophical Magazine and Journal of Science

William S. Gosset
(*1877–1937*)

William Sealy Gosset was the Head Experimental Brewer at Guinness Brewing Company. While Pearson as a biometrician typically had hundreds of observations, Gosset's work in the brewery dealt primarily with small samples.



VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

- In 1908, Gosset published *The Probable Error of the Mean*, a paper in which he developed the *t*-distribution
- Extended Pearson's work on parameter estimation for small samples
  - His major finding was that some of the parameter values of a distribution could be estimated from others
- Also presented *t*-table for several sample sizes
- Some of his mathematics was corrected by R.A. Fisher, most notably dividing by *n–1* rather than *n*

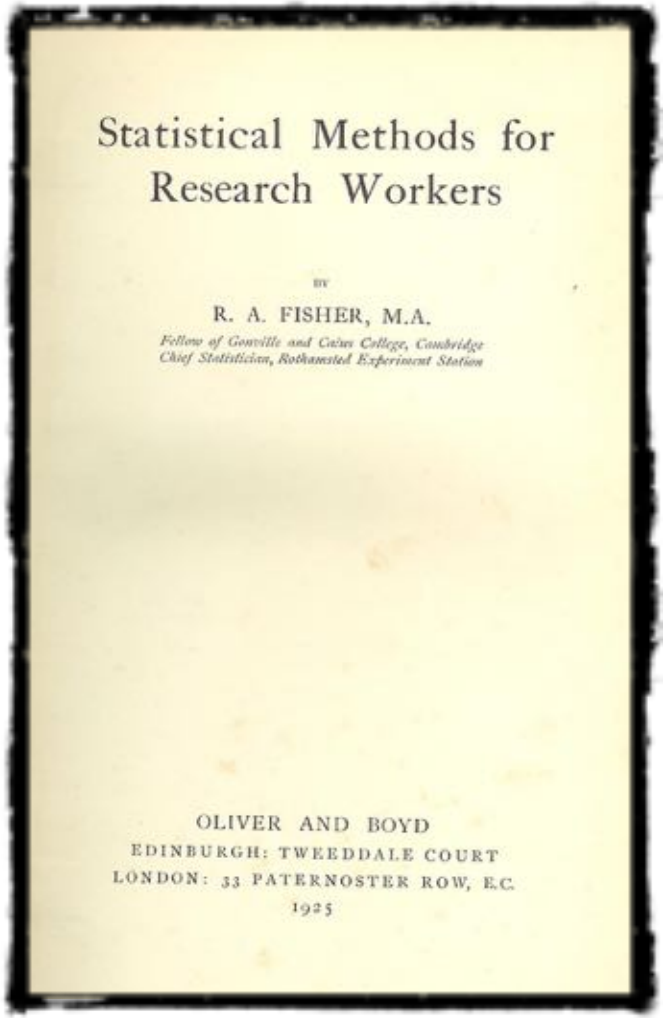
To prevent disclosure of confidential information, the Guinness Board of Directors allowed its scientists to publish research on condition that they do not mention "1) beer, 2) Guinness, or 3) their own surname". Gosset published under the pseudonym "Student", and thus his most noteworthy achievement is now called Student's t-distribution.

Wikipedia

Fisher was the godfather of modern-day significance testing. He viewed a set of data as a random sample of all possible measurements, and consequently viewed the test statistic generated from that sample as random. (He coined the term "statistic" to differentiate this from the underlying parameter.) Because of the random nature of the statistic, Fisher proposed that measurements need to be evaluated relative to the probability distribution of the test statistic.

Salsburg (2001)
The Lady Tasting Tea

R. A. Fisher
(*1890–1962*)


Statistical Methods for Research Workers

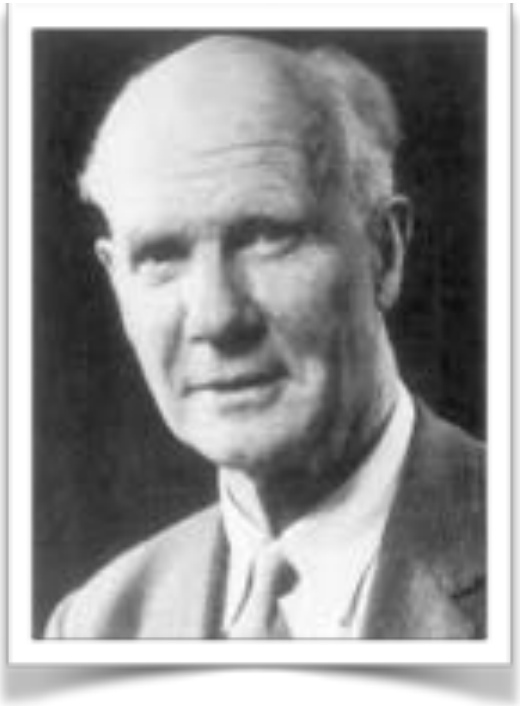
BY

R. A. FISHER, M.A.

Fellow of Gonville and Caius College, Cambridge
Chief Statistician, Rothamsted Experiment Station

OLIVER AND BOYD
EDINBURGH: TWEEDDALE COURT
LONDON: 33 PATERNOSTER ROW, E.C.
1925

- In his 1925 book (left), Fisher popularized the idea of the *p*-value as a measure of.the discrepancy between data and a null hypothesis
- To help researchers evaluate the *p*-value, he also introduced idea of *significance level* and proposes the criterion of 0.05 as "a convenient cutoff" for results that exceed chance
  - Applying this to the normal distribution led him to the 1.96 SD rule



Egon Pearson
(*1895–1980*)

Jerzy Neyman
(*1891–1981*)

"After Fisher's work was introduced, Jerzy Neyman and Egon Pearson tackled some unanswered questions. For example,…you can choose to compare [many different test statistics] so long as you can derive a *p*-value for the comparison. But how do you know which is best? What does 'best' even mean for hypothesis testing?"

Reinhart (2015)
Statistics Done Wrong

- In their 1933 paper, Neyman and Pearson introduced the formulation of Neyman-Pearson hypothesis testing
  - Major discovery was that significance testing made no sense unless an *alternative hypothesis* was included
  - The choice of the alternative hypothesis dictates the way the significance test is carried out
  - Probability of detecting that alternative hypothesis they referred to as "power" of the test
- Two major conclusions stemmed from this work
  - Power of the test was a measure of how good the test was; thus when choosing between any two tests, choose the more powerful
  - The set of alternative hypotheses needs to be well specified (and small) for the test to be meaningful.

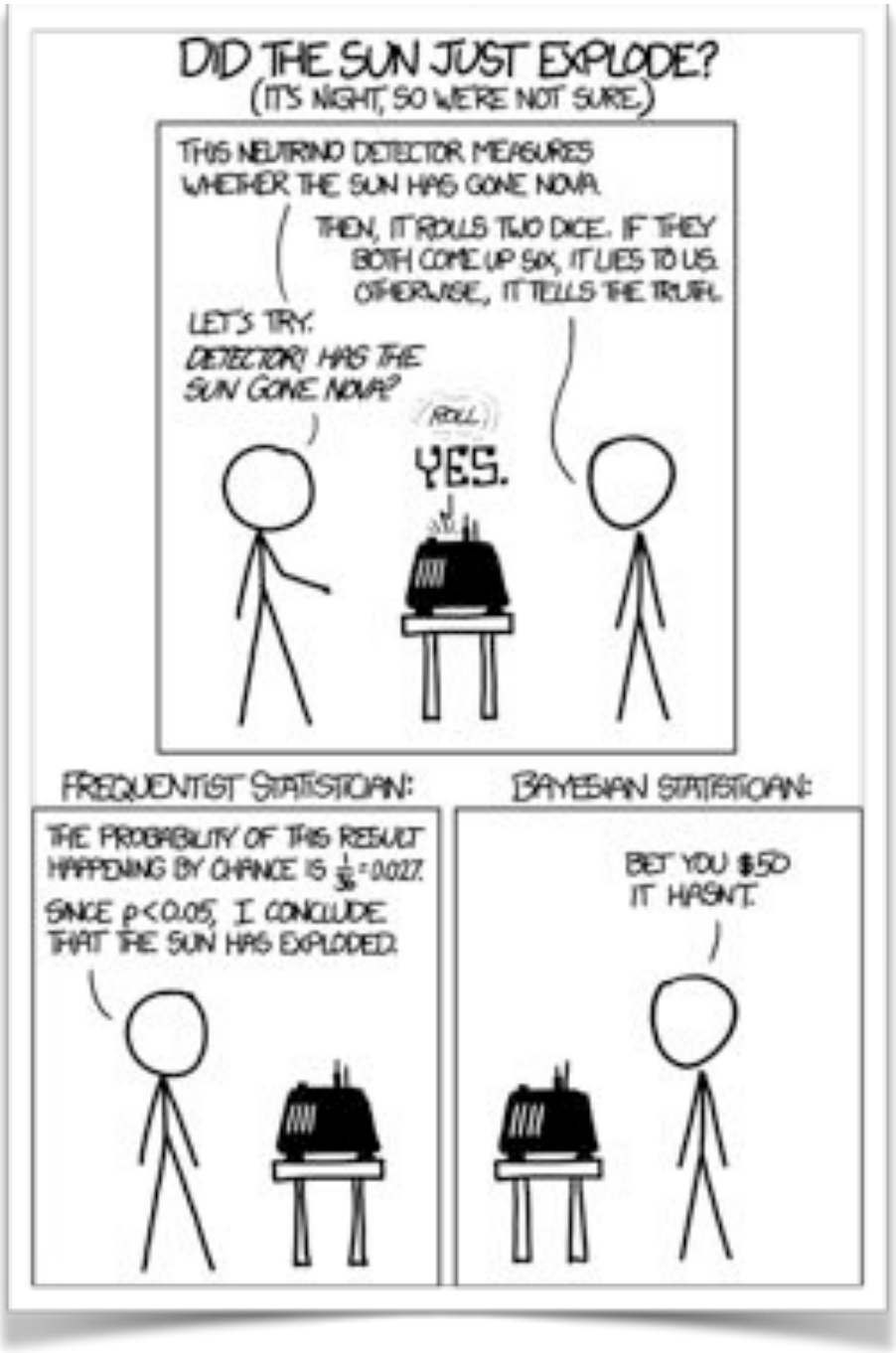
What's the Problem?

- When Neyman formalized the mathematics of his and E. Pearson's hypothesis testing framework, he **relied on John Venn's formulation of mathematical probability** to justify the use of *p*-value
  - Venn claimed the probability associated with an event is the long-run proportion of times that event occurs (i.e., Frequentist notion of probability)
  - John Maynard Keynes (1921) showed this definition has major inconsistencies that make this notion of probability untenable

- In the N-P framework a scientist sets a significance level (.05) and rejects the null hypothesis when the *p*-value is less than this level
  - This is predicated on the Frequentist notion of probability

- The N-P framework has been **severely criticized** from its inception
  - Fisher was their most ardent attacker; he didn't think it was compatible with the process of scientific discovery.
  - W. Edwards Deming also criticized hypothesis tests, calling them nonsensical
  - Neyman himself raised doubts about whether optimum tests could be found and in his later work rarely used hypothesis tests

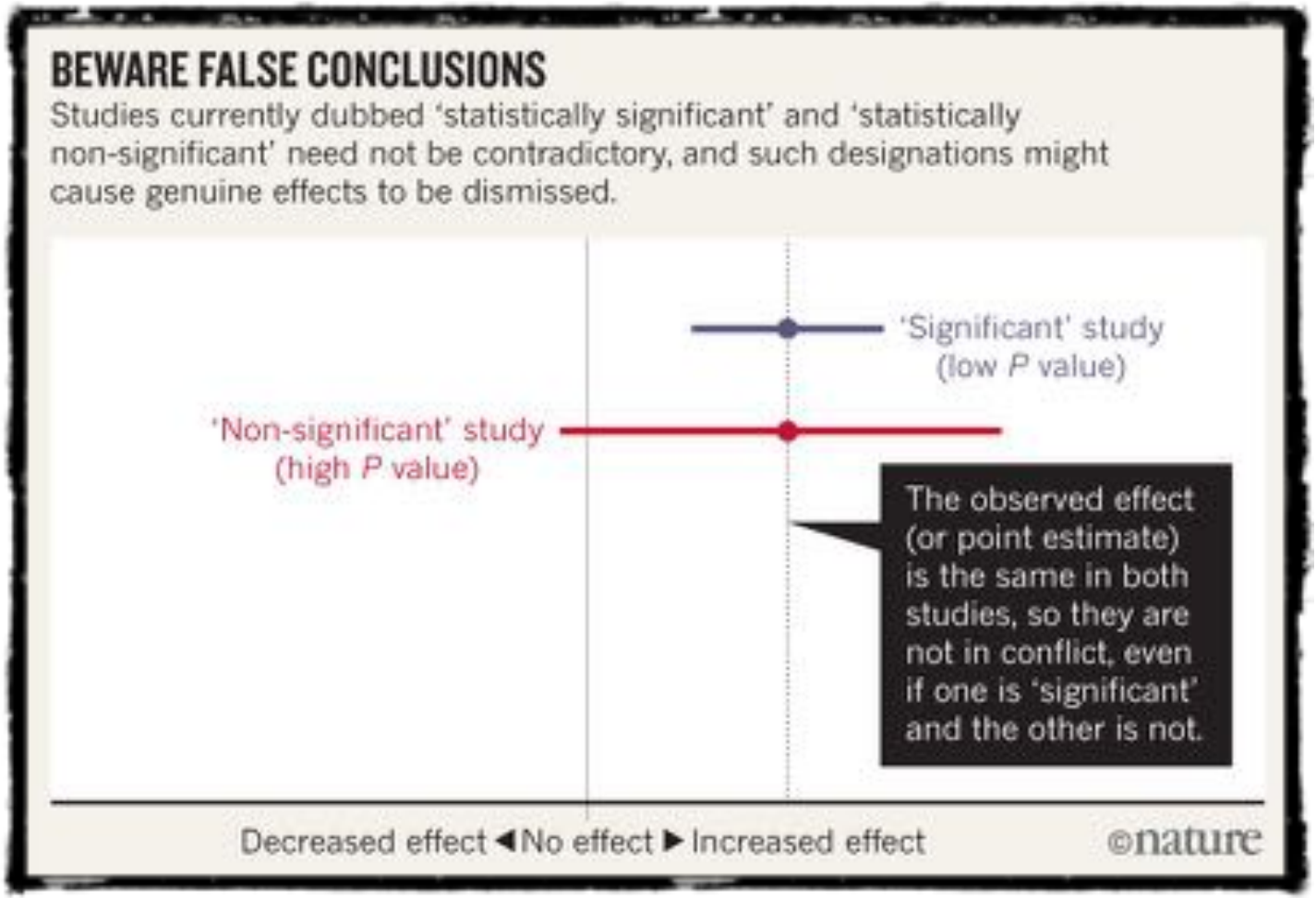
- Despite all this, the N-P framework has a **prominent foothold in scientific culture**
  - Erich Lehmann (definitive text on hypothesis testing) and Abraham Wald (statistical decision theory) expanded the N-P framework
  - Statistics textbooks (and courses) present a conglomeration of the Fisherian and Neyman-Pearson frameworks (mea culpa)
  - For many years many journal editors required hypothesis tests
  - Drug regulatory agencies and courts of law use hypothesis tests



Where does that leave us?

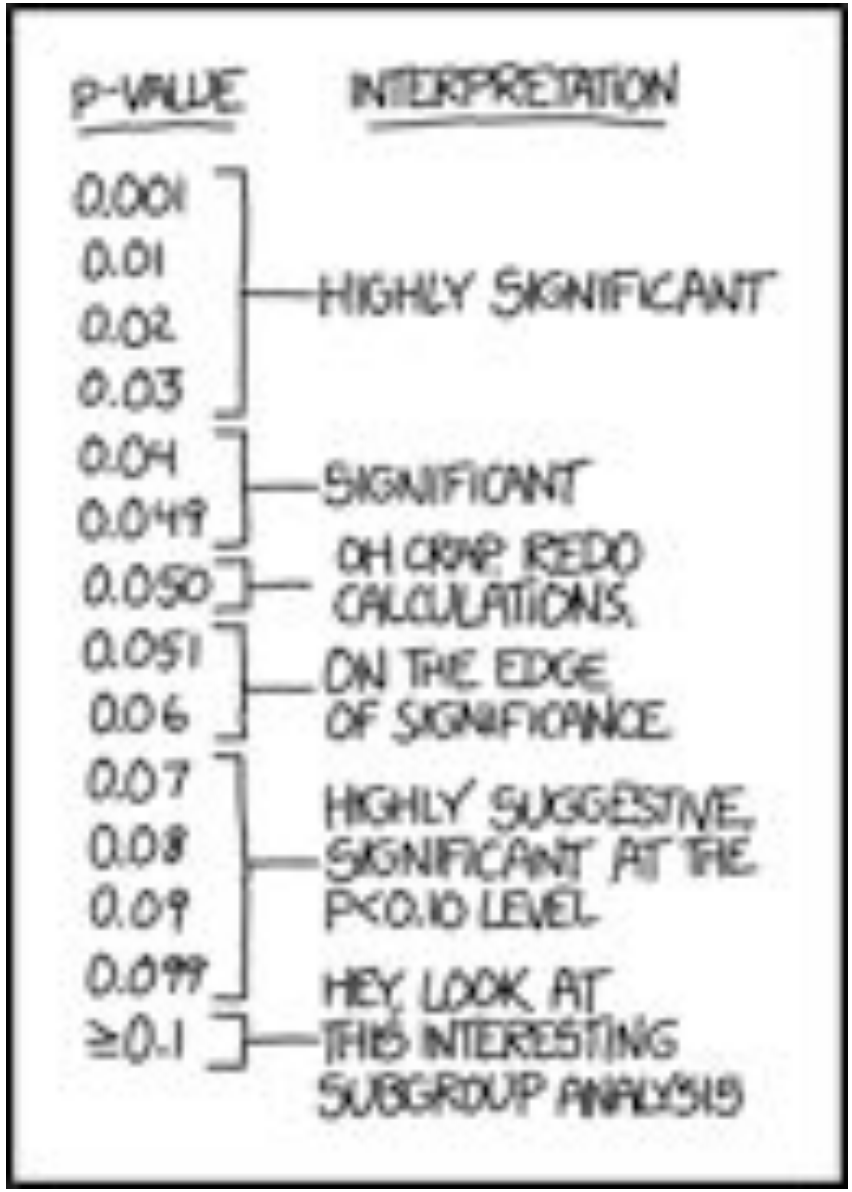
The trouble is human and cognitive more than it is statistical: **bucketing results into 'statistically significant' and 'statistically non-significant' makes people think that the items assigned in that way are categorically different**. The same problems are likely to arise under any proposed statistical alterna- tive that involves dichotomization, whether frequentist, Bayesian or otherwise.

Amrhein, Greenland, & McShane (2019)
Nature



- Scientists / researchers **privilege statistically significant results**
  - Why? Because of a false belief that such results are "real"
  - Encourages researchers to choose data and methods that yield statistical significance

- This **biases the results** reported in the literature
  - Statistically significant estimates are biased upwards in magnitude
  - Statistically non-significant estimates are biased downwards in magnitude



Solution #1: Accept Uncertainty

> Uncertainty is a part of the research process, and making a binary statement about a hypothesis does not change that. Researchers should accept this uncertainty in all their statistical conclusions and seek ways to quantify, visualize, and interpret the potential for error.
>
> Calin-Jagerman & Cumming (2019)
> *The American Statistician*

- Provide **estimates of the uncertainty** along with any point estimate
  - This has been advocated by APA for quite some time.

- Interpret interval estimates in a way that **avoids overconfidence**
  - Describe the practical implications for ALL values inside the interval, especially the observed effect and the limits of the interval
  - The interval gives the MOST compatible values given the data. Just because a value is outside the interval does not make it incompatible, just less compatible.
  - Not all values in the interval are equally compatible with the data
  - Using a 95% CI is just as arbitrary as choosing .05. The level should be justified based on the application/use of the interval.
  - The intervals hinge on the statistical assumptions

Solution #2: Be Thoughtful Researchers

> "[M]ost scientific research is exploratory in nature…[t]he design, conduct, and analysis of a study are necessarily flexible, and must be open to the discovery of unexpected patterns that prompt new questions and hypotheses. In this context, statistical modeling can be exceedingly useful for elucidating patterns in the data, and researcher degrees of free- dom can be helpful and even essential, though they still carry the risk of overfitting. **The price of allowing this flexibility is that the validity of any resulting statistical inferences is undermined.**"
>
> Tong (2019)
> *The American Statistician*

- Think about the broader science rather than any one individual study
  - One study does not a science make
- Ask pertinent questions prior to carrying out the research
  - What is a meaningful effect size?
  - What does the related prior evidence say about mechanism? Data collection? Costs? Benefits?
- Consider multiple approaches to the research and analysis
  - Employ an entire toolbox of statistical techniques
- Be clear in your communication of results
  - Acknowledge the uncertainty
  - Acknowledge researcher degrees of freedom, assumption violations, etc.

Solution #3: Pre-Registration of Studies

P**re-register studies** and commit to publishing ALL results, statistically significant or not. However pre-registration is not a panacea. Many important results are only found after looking at the data, which is problematic in fields where obtaining data for follow-up studies may be difficult.

Gelman & Loken (2014)
American Scientist

Support for study pre-registration is increasing; websites such as the *Open Science Framework* (http://osf.io/) and *AsPredicted* (http://AsPredicted. org/) offer services to pre-register studies, the *Preregistration Challenge* offers education and incentives to conduct pre-registered research (http://cos.io/prereg), and journals are adopting the Registered Reports publishing format to encourage pre-registration and add results-blind peer review

Munafo et al. (2017)
Nature

Solution #4: Change Editorial/Academic Practices

> "Institutional reform is necessary for moving beyond statistical significance in any context—whether journals, education, academic incentive systems, or others."
>
> Tong (2019)
> *The American Statistician*

- Consider how you evaluate/review scholarship
  - Evaluation of manuscripts for publication should be "results-blind." Instead, assess them for suitability for publication based on the **substantive importance of the research** without regard to their reported results. (Kmetz, 2019; Locascio, 2017; 2019)

- Ask authors to remove any language/form of "statistically significant"
  - Instead have them focus on the effect and whether that is meaningful
  - Journal editors and board members can change their policies and author guidelines to discourage this type of binary thinking

- Reduce incentives that "reward" unreliable work
  - Replication studies are just as necessary as "new" research, yet are typically not published or pursued in a publish-or-perish environment
  - Good science takes time and may result in fewer publications over a given period of time. Quality is not necessarily associated with quantity and many of the metrics we use for academic evaluation reward quantity.

Thank You!

Email: zief0002@umn.edu

Slides: http://www.datadreaming.org/files/2019-slhs-prosem.pdf

References

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature, 567*, 305–307. https://www.nature.com/articles/d41586-019-00857-9#ref-CR9

Baylor, C., Yorkston, K., Banner, A., Britton, D., & Amtmann, D. (2010). Variables associated with communicative participation in people with multiple sclerosis: A regression analysis. *American Journal of Speech-Language Pathology, 19*(2), 143–153. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2873072/

Calin-Jagerman, R., & Cumming, G. (2019). The new statistics for better science: Ask how much, how uncertain, and what else Is known. *The American Statistician, 73*, 271–280.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh: Oliver and Boyd.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460. https://www.americanscientist.org/article/the-statistical-crisis-in-science

Goldsworthy, R. L., & Markle, K. L. (2019). Pediatric hearing loss and speech recognition in quiet and in different types of background noise. *Journal of Speech, Language, and Hearing Research, 62*(3), 758–767.

Keynes, J. M. (1921). *Treatise on probability.* London: Macmillan & Co.

Kmetz, J. L. (2019). Correcting corrupt research: Recommendations for the profession to stop misuse of p-values. *The American Statistician, 73,* 36–45.

Locascio, J. L. (2017). Results blind science publishing. *Basic and Applied Social Psychology, 39*, 239–246.

Locascio, J. L. (2019). The impact of results blind science publishing on statistical consultation and collaboration. *The American Statistician, 73,* 346–351.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Sert, du, N. P., et al. (2017). A manifesto for reproducible science. *Nature Human Behavior, 1*, 1–9. doi: 10.1038/s41562-016-0021

References (cntd.)

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London (Series A), 231*, 289–337.

Pearson, K. P. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine and Journal of Science, L*, 157–175.

Plackett, R. L. (1983). Karl Pearson and the chi-squared test. *International Statistical Review, 51*(1), 59–72. Dos: 10.2307/1402731

Reinhart, A. (2015). *Statistics done wrong: The woefully complete guide*. San Francisco, CA: No Starch Press

Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. New York: W. H. Freeman.

Student. (1908). The probable error of the mean. *Biometrika, 6*(1), 1–25.

Tong, C. (2019). Statistical inference enables bad science; Statistical thinking enables good science. *The American Statistician, 73*, 246–261.

Vroegop, J. L., Dingemanse, J. G., van der Schroeff, M. P., & Goedegebure, A. (2019). Comparing the effect of different hearing aid fitting methods in bimodal cochlear implant users. *American Journal of Audiology, 28*(1), 1–10.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "$p < 0.05$." *The American Statistician, 73*(sup1), 1–19. doi: 10.1080/00031305.2019.1583913

William Sealy Gossett. (n.d.). In *Wikipedia*. https://en.wikipedia.org/wiki/William_Sealy_Gosset