

# p-value in 2020

Its values and pitfalls

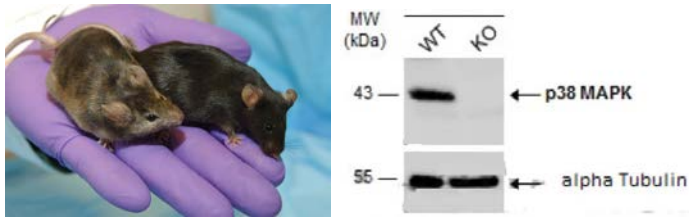
Technical Journal Club

September 2019

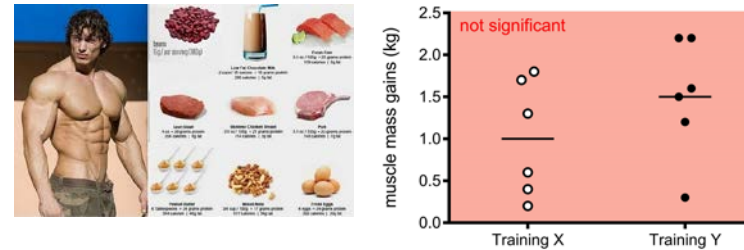
Marc Emmenegger (Aguzzi lab)

# Experiments measure differences

Wildtype and knockout



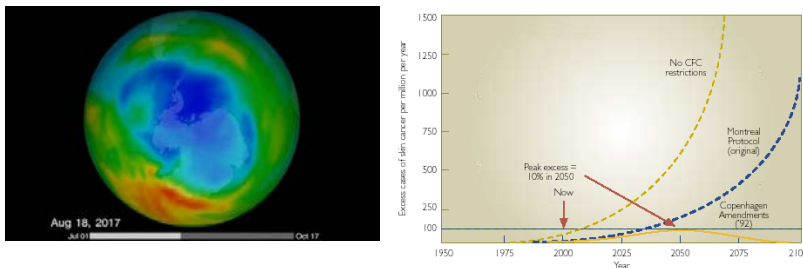
Training and muscle mass gain



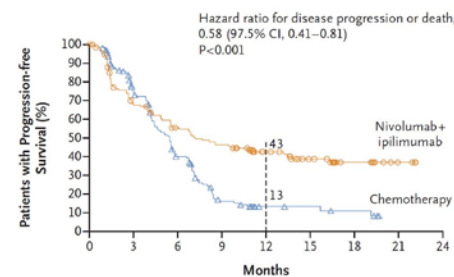
# trees on earth



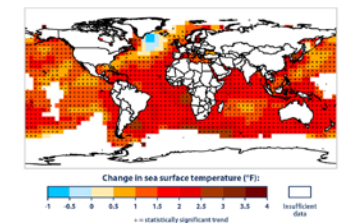
Ozone depletion and skin cancer



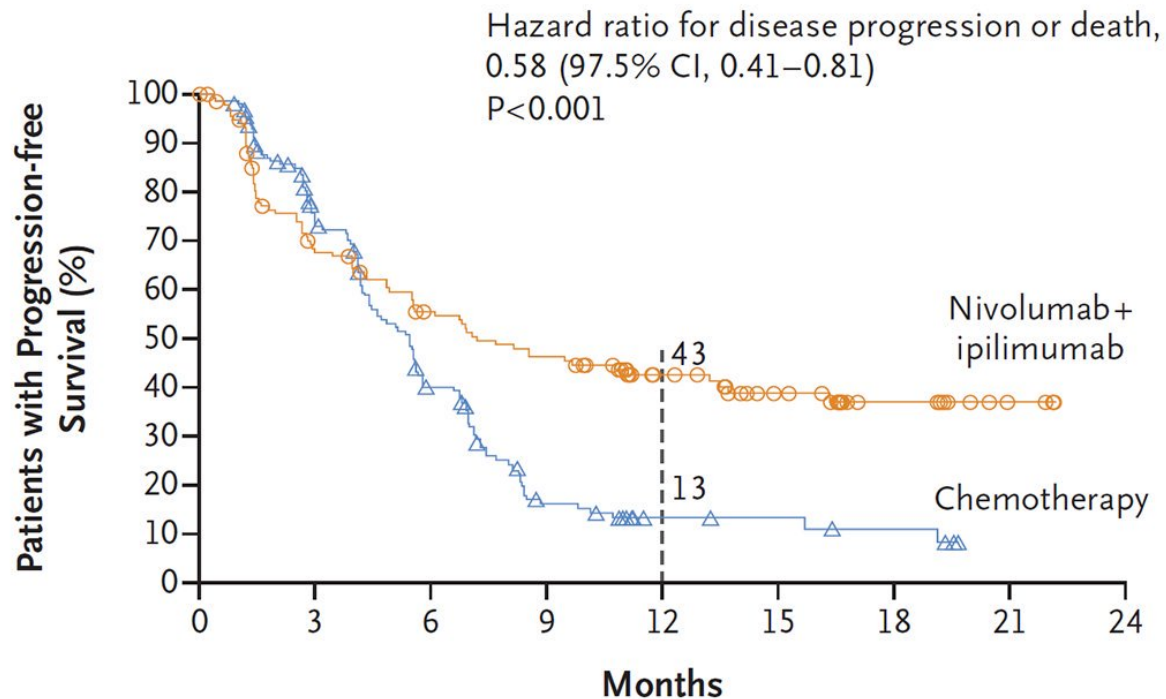
Effect of treatment modality on PFS



Global warming - sea temp



# Evaluating differences



- Difference due to chance?
- Difference due to random variation?
- Use statistics → Is the difference we observe larger than if it was just by luck?

# Agenda for today's JC

- Biological data is often noisy – what are true differences and what are random differences?
- The p-value in its historical perspective.
- The problem of statistical significance – reproducibility crisis, p-value hacking, truth inflation
- Recommendations for statistical testing.

# Agenda for today's JC

- Biological data is often noisy – what are true differences and what are random differences?
- The p-value in its historical perspective.
- The problem of statistical significance – reproducibility crisis, p-value hacking, truth inflation
- Recommendations for statistical testing.

# Study results versus reality




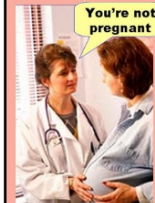
Test  
↙

Yes/No  
→



Yes/No  
↓

In reality

		Significant result	Non-significant result
Study conclusion	Significant result	Correct conclusion	 False positive (type 1 error) <b>p-value</b>
	Non-significant result	 False negative (type 2 error) <b>Statistical power</b>	Correct conclusion

# Drawing conclusions

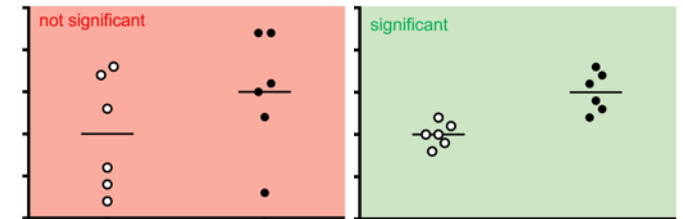
## Question:

- Does training modality influence muscle mass gain?
- Does depletion of protein X influence infectivity?
- Am I infected with HIV and pose a serious risk to others?



## Measurement:

- Sample size
- Random variation
- Robustness
- Methodology
- Effect size

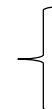


## Interpretation:

- There are always differences
- Statistical testing to interpret difference
- Power of analysis
- False-positive discovery rate



parameters



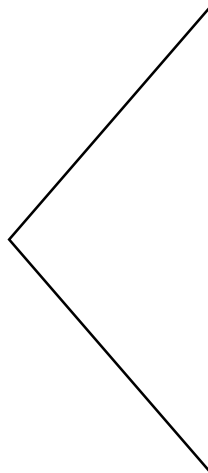
## Conclusions:

- «Training X may be inferior to training Y in terms of muscle mass gain.»
- «Our results show that it is possible that protein X positively influences infectivity.»
- «There is a 75% chance that you are infected with HIV according to our tests.»

## Reporting:

YES or NO

But «yes» or «no» choice depend on our **parameters** (and on other factors)



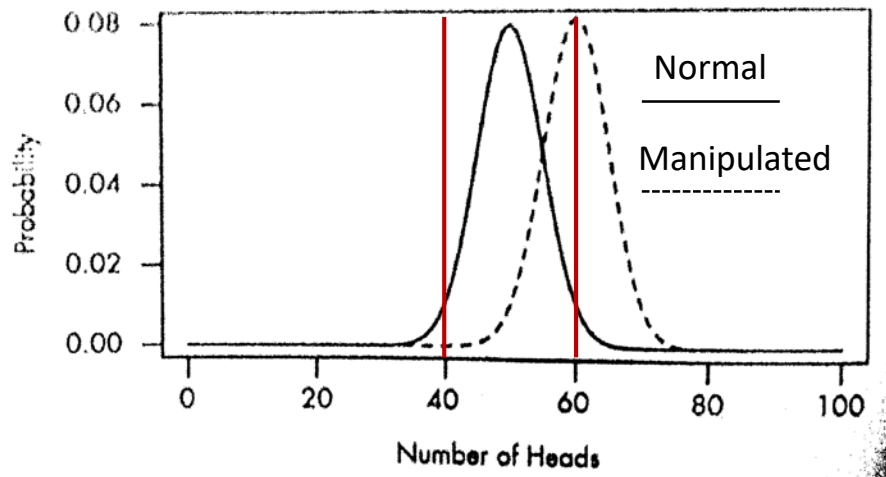
# Concept: statistical power



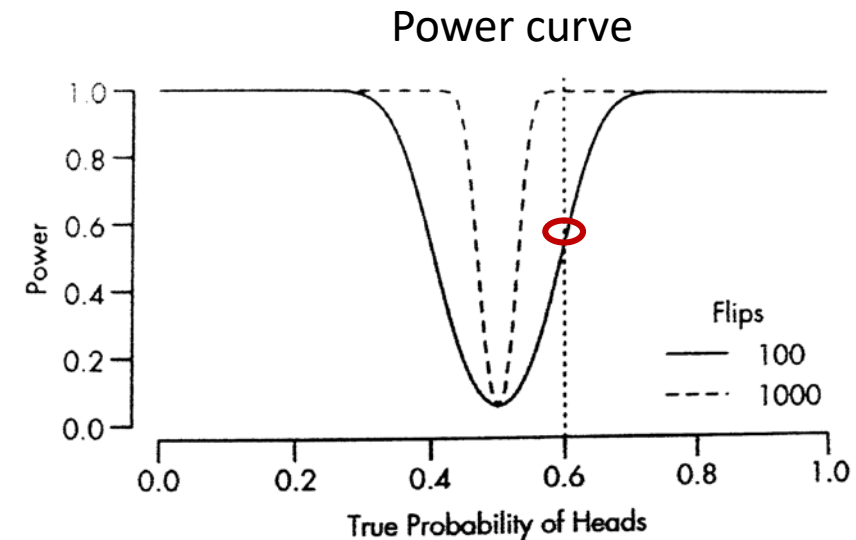
Flip coin 100 times

Result: 58x head, 42x tail

Is the coin rigged?



- Probability for 50x head < below 10%
- 95% 40-60 head
- Is the coin fair or not?
- Let us say that if  $p < 0.05$ , the coin is unfair



- If we flip 100x, and the true probability of the coin is 60% head (because it is unfair), the power is 0.5.
- The probability that I conclude it is rigged is 50%.
- Only if head > 80/100 flips, sample size large enough to detect bias.
- Increase sample size: Sensitivity is increased.




# Concept: statistical significance


- Our lab developed a new medicine to treat flu – mouse experiments suggest that the average duration of the flu is reduced.
- Double-blind study: 50 patients + new medicine, 50 patients + placebo.
- Not all flu take same time, biological variability. Need for statistical hypothesis testing.
- P-value: Probability (assumption  $H_0$ : no true effect) that data are equal/more extrem than actually observed results → The lower the p-value, the higher the **degree of surprise** = the lower the evidence for the null hypothesis (no true effect).
- But does not tell you whether your alternative hypothesis is true.
- Additionally: Statistical significance  $\neq$  practical (biological/medical/...) significance.

# Significance of power and statistical significance

- In Science and Nature, less than 3% of articles calculate statistical power before starting their study (Tressoldi, 2013, PlosOne).
- Conclusion: «no statistical significance» while not noticing that insufficient data to detect any but the most enormous differences (Tsang, 2009, J of Clin Epid).
- Clinical trials in cancer research: Only around ½ published studies with negative results had enough statistical power to detect even huge difference in primary outcome variable (Bedard, 2007, J. Clin Oncol). Similar in other fields (Brown, 1987, Ann of Emer Med; Chung, 1998, J of Hand Surg).

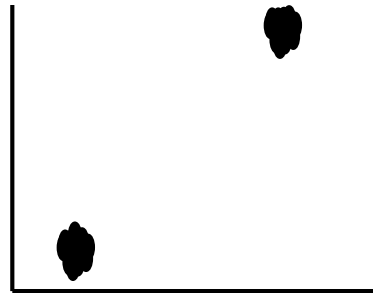
## Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson & Marcus R. Munafò 

*Nature Reviews Neuroscience* **14**, 365–376 (2013) | [Download Citation](#) 

# Why do we get false results?

Wishful thinking

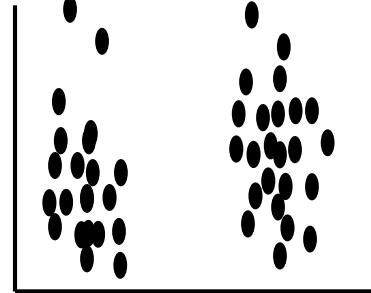


A

B



Experimental reality



A

B

- Crystall clear difference
  - Has probably already been found
  - Interpretation straightforward
- Small difference (may nevertheless be ultra important)
  - In reality, effect may be much bigger (but measurement is imperfect)
  - Often the case in biology, e.g.
  - **Significant or not = real or not?**
  - Interpretation challenging

- Imperfect methods may blur our measurements.
- Natural variability due to hard-to-control parameters.
- If it is all too obvious, we may not need to design extremely sophisticated experiments.
- The effect size of what we look at is often small (but we hope that the consequences can be big).

# Agenda for today's JC


- Biological data is often noisy – what are true differences and what are random differences?
- The p-value in its historical perspective.
- The problem of statistical significance – reproducibility crisis, p-value hacking, truth inflation
- Recommendations for statistical testing.

# P-value: Historical perspective

Special Report | Published: 11 July 2016


## What is the P-value anyway?

R P Gale  & M-J Zhang

*Bone Marrow Transplantation* 51, 1439–1440 (2016) | [Download Citation](#) 

## What is the (p-) value of the P-value?

R P Gale , A Hochhaus & M-J Zhang

*Leukemia* 30, 1965–1967 (2016) | [Download Citation](#) 

*One should try everything in life except incest, folk dancing and calculating a P-value.*

Inferential processes

| *After Sir Thomas Beecham, 2nd Baronet, CH*

Transplant study with new drug that decreases likelihood of GvHD compared with placebo. You do the test and get a  $p = 0.055$  ( $\alpha < 0.05$ ). What can you say?

- A. The new drug is ineffective.
- B. The results can be accounted for by chance.
- C. The null hypothesis is true (= no difference between new drug and placebo).
- D. All of the above are true.
- E. **None of the above!**

# P-value: Historical perspective

- **R. A. Fisher** introduced p-value into scientific research as measure of inference.
- «Probability of the observed result, plus more extreme results if the null hypothesis were true.»
- P-value as a component of a complex process (causal inference), not the only component:
  - No relationship between causal factors.
  - No systematic errors.
  - Use of appropriate statistical test.
- In addition: Highly contingent on assumptions (otherwise, incorrect use, and not suggested by Fisher).
- **Neyman and Pearson:** In contrast to Fisher, postulated a mutually exclusive alternative hypothesis to the null hypothesis.
  - Type I error (false-positive)
  - Type II error (false-negative)

# Agenda for today's JC

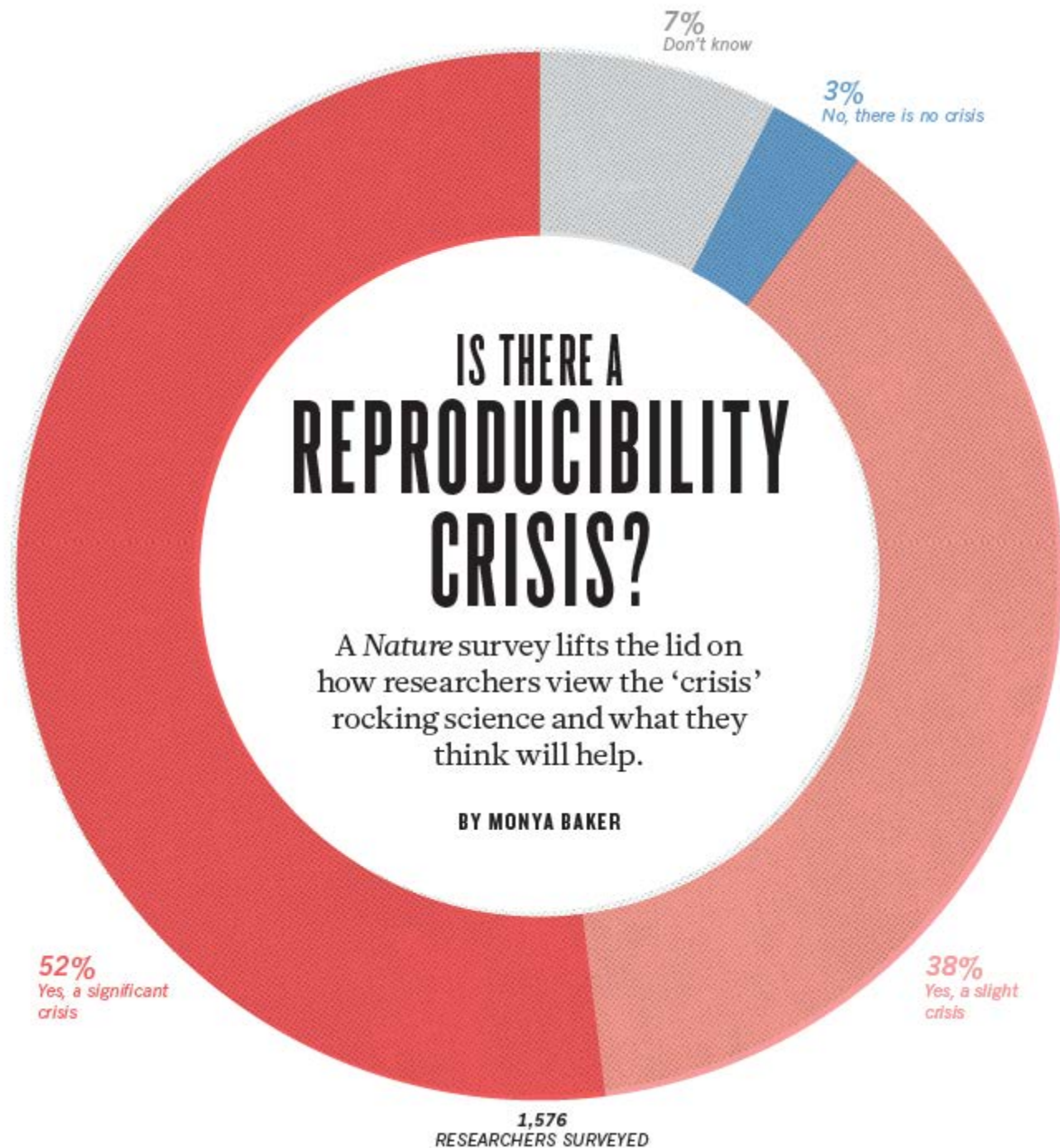
- Biological data is often noisy – what are true differences and what are random differences?
- The p-value in its historical perspective.
- The problem of statistical significance – reproducibility crisis, p-value hacking, truth inflation
- Recommendations for statistical testing.



# FOOLING OURSELVES

HUMANS ARE REMARKABLY GOOD AT SELF-DECEPTION.  
BUT GROWING CONCERN ABOUT REPRODUCIBILITY IS DRIVING MANY  
RESEARCHERS TO SEEK WAYS TO FIGHT THEIR OWN WORST INSTINCTS.



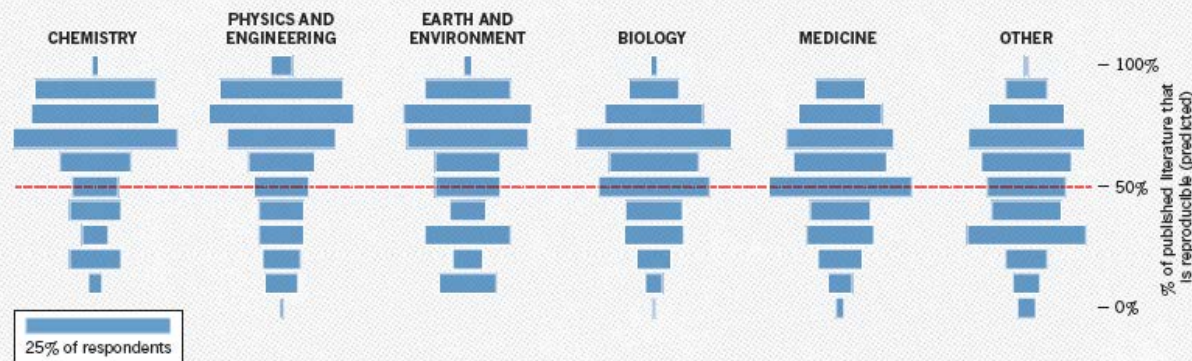


# A 'CRISIS' IN NUMBERS

Nature surveyed 1,576 scientists online to get their thoughts on reproducibility in their field and in science in general. See [go.nature.com/2vj4r4y](https://go.nature.com/2vj4r4y) for more charts and access to the full data.

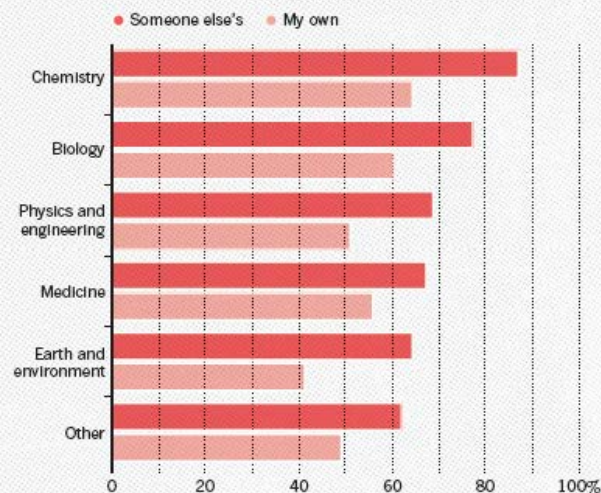
## HOW MUCH PUBLISHED WORK IN YOUR FIELD IS REPRODUCIBLE?

Physicists and chemists were most confident in the literature.



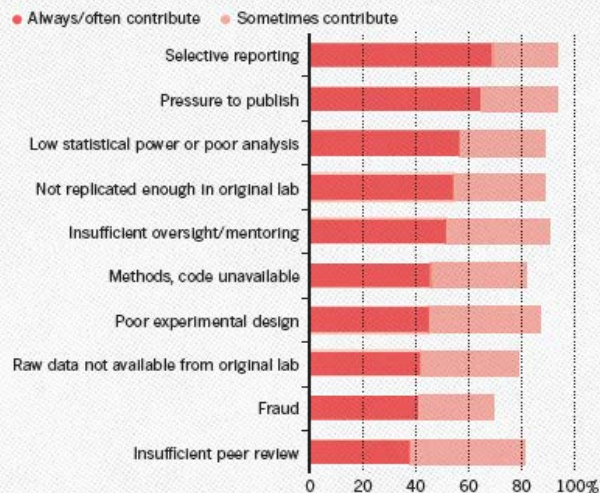
## HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?

Most scientists have experienced failure to reproduce results.



## WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

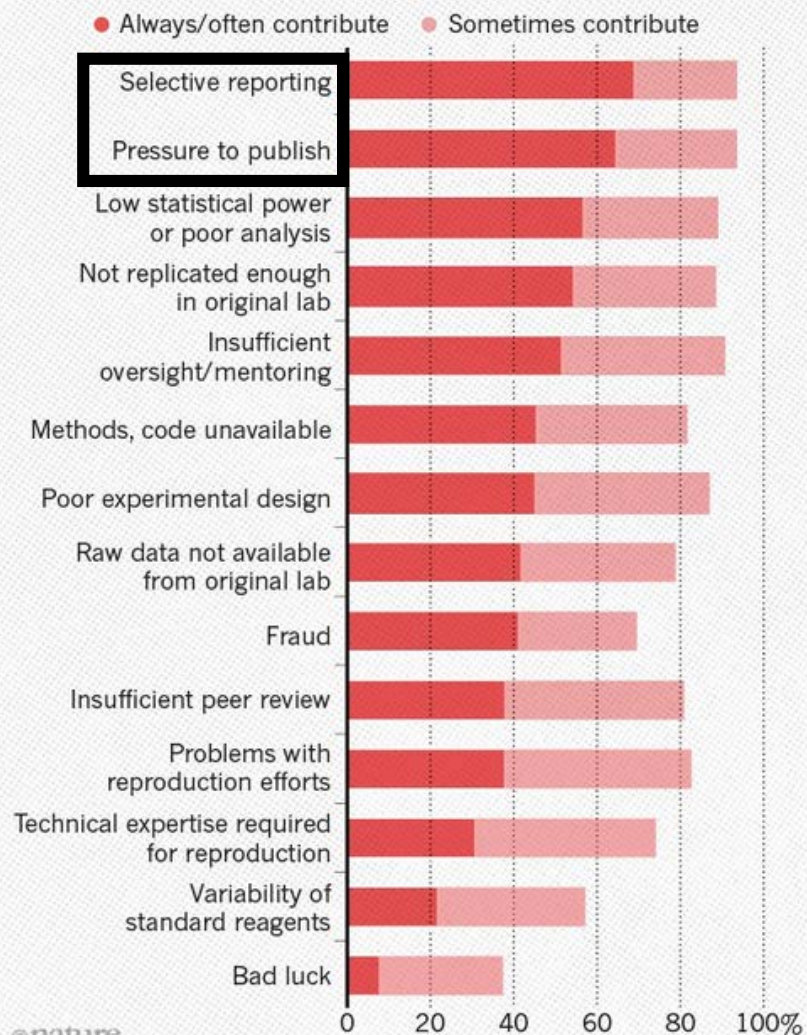
Many top-rated factors relate to intense competition and time pressure.





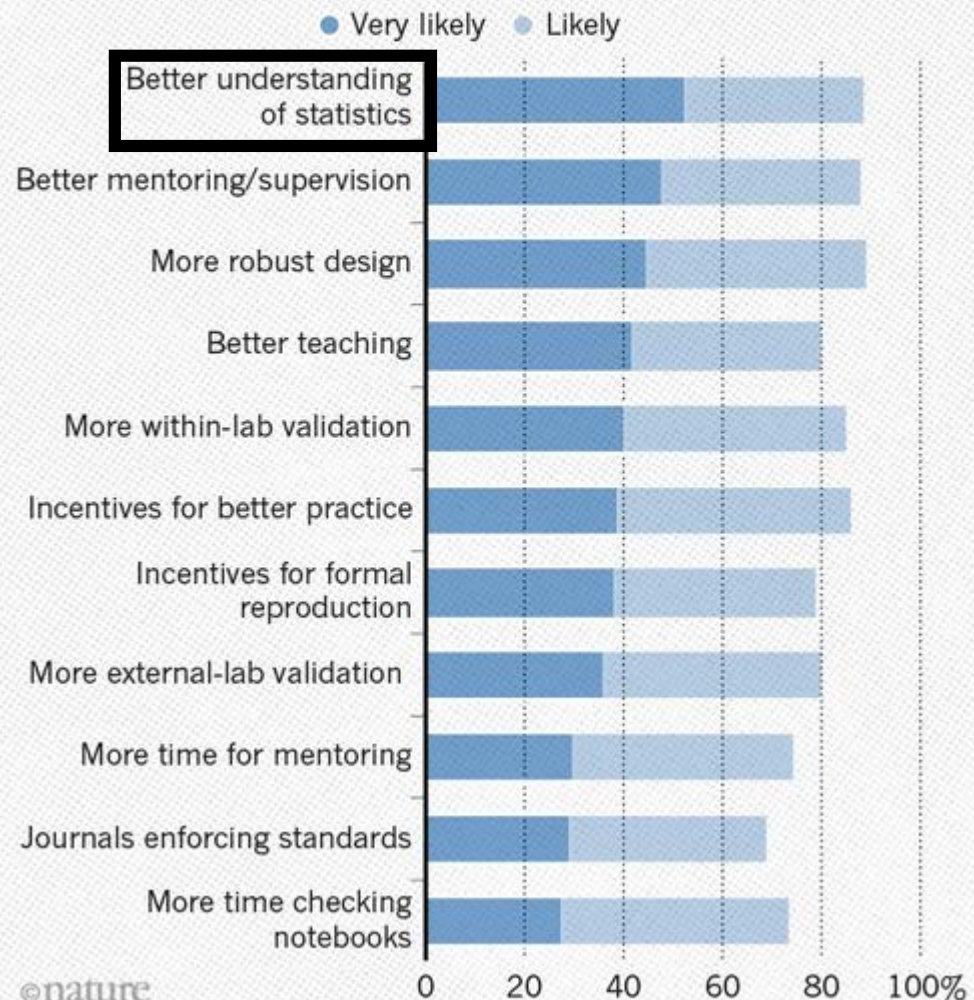
## WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

Many top-rated factors relate to intense competition and time pressure.



## WHAT FACTORS COULD BOOST REPRODUCIBILITY?

Respondents were positive about most proposed improvements but emphasized training in particular.





# STATISTICAL ERRORS

P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

BY REGINA NUZZO

# The ASA Statement on $p$ -Values: Context, Process, and Purpose

Ronald L. Wasserstein & Nicole A. Lazar

**To cite this article:** Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA Statement on  $p$ -Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: [10.1080/00031305.2016.1154108](https://doi.org/10.1080/00031305.2016.1154108)

**To link to this article:** <https://doi.org/10.1080/00031305.2016.1154108>

IN FOCUS NEWS

REPRODUCIBILITY

## Statisticians issue warning on $P$ values

*Statement aims to halt missteps in the quest for certainty.*



# **False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant**

**Joseph P. Simmons<sup>1</sup>, Leif D. Nelson<sup>2</sup>, and Uri Simonsohn<sup>1</sup>**

<sup>1</sup>The Wharton School, University of Pennsylvania, and <sup>2</sup>Haas School of Business, University of California, Berkeley

- Our job as scientists is to discover truths about the world.
- Perhaps, the most costly error is false-positives – they are particularly persistent.
- It is uncommon for prestigious journals to publish null finding – failures to replicate previous findings are never conclusive.
- Current practice makes false-positives much more likely than maximum false-positive rate of 5% suggests.

# False-positive psychology

- Why? Researcher's degree of freedom:
  - Should more data be collected?
  - Should some observations be excluded?
  - Which conditions should be combined?
  - Which should be compared?
  - Which controls should be considered?
  - Transformation of specific measures?
- If all this is done before even starting data collecting: unproblematic.
- But usually, this process starts post-hoc – this is accepted practice.
- Finally, it is published what worked, i.e. what delivered significant (= meaningful) results.

# False-positive psychology – Study 1

## Musical contrast and subjective age

- We investigated whether listening to children's song induces an age contrast, making people feel older.
- In exchange for payment, 30 Upenn undergraduates were randomly assigned to listen to control song (Kalimba) or children's song (Hot Potato).
- After listening to the song, participants completed an ostensibly unrelated survey: How old do you feel right now (very young, young, neither young nor old, ...).
- They also reported their father's age.
- Analysis of covariance revealed the predicted effect: People felt older after listening to Hot Potato (**p = 0.033**).
- Study 2 – extension of study 1.



# False-positive psychology – Study 2

Musical contrast and chronological rejuvenation: Does listening to a song about old age make people **actually** younger?

- Same method as study 1.
- 20 Upenn undergraduates listen to either «When I am sixty-four» or «Kalimba».
- Indication of birth date and father's age.
- Analysis of covariance revealed the predicted effect: After listening to «When I am sixty-four», participants were nearly a year-and-a-half younger rather than «Kalimba» (**p = 0.04**).
- **This study supports the idea that listening to songs about old age rejuvenates people.**
- Problems that should be explicitly stated (but most often not recognised in scientific community):
  - Rule for data collection termination.
  - Enough observations per condition.
  - List all variables in study.
  - All experimental conditions.
  - Report eliminations of observations.
  - If analysis contains covariate, must report statistical result without covariate.

# False-positive psychology – Discussion

- Authors now implement their «guidelines» and re-report study 2 where they have shown that people become younger after listening to Beatles song:

**Table 3.** Study 2: Original Report (in Bolded Text) and the Requirement-Compliant Report (With Addition of Gray Text)

---

Using the same method as in Study 1, we asked 20 34 University of Pennsylvania undergraduates to listen only to either “**When I’m Sixty-Four**” by The Beatles or “Kalimba” or “Hot Potato” by the Wiggles. We conducted our analyses after every session of approximately 10 participants; we did not decide in advance when to terminate data collection. **Then, in an ostensibly unrelated task, they indicated only their birth date (mm/dd/yyyy) and** how old they felt, how much they would enjoy eating at a diner, the square root of 100, their agreement with “computers are complicated machines,” **their father’s age**, their mother’s age, whether they would take advantage of an early-bird special, their political orientation, which of four Canadian quarterbacks they believed won an award, how often they refer to the past as “the good old days,” and their gender. **We used father’s age to control for variation in baseline age across participants.**

An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to “**When I’m Sixty-Four**” (adjusted  $M = 20.1$  years) rather than to “Kalimba” (adjusted  $M = 21.5$  years),  $F(1, 17) = 4.92, p = .040$ . Without controlling for father’s age, the age difference was smaller and did not reach significance ( $M_s = 20.3$  and  $21.2$ , respectively),  $F(1, 18) = 1.01, p = .33$ .

---

- Authors conclusion: «The redacted version of the study we reported in this article fully adheres to currently acceptable reporting standards and is, not coincidentally, deceptively persuasive.»
- «The requirement-compliant version [...] would be – appropriately – all but impossible to publish.»

## ***P* values and the search for significance**

Little *P* value

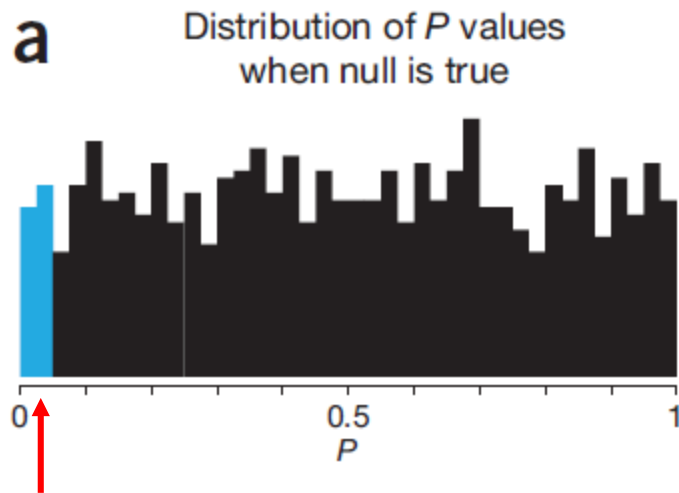
What are you trying to say  
Of significance?

—Steve Ziliak

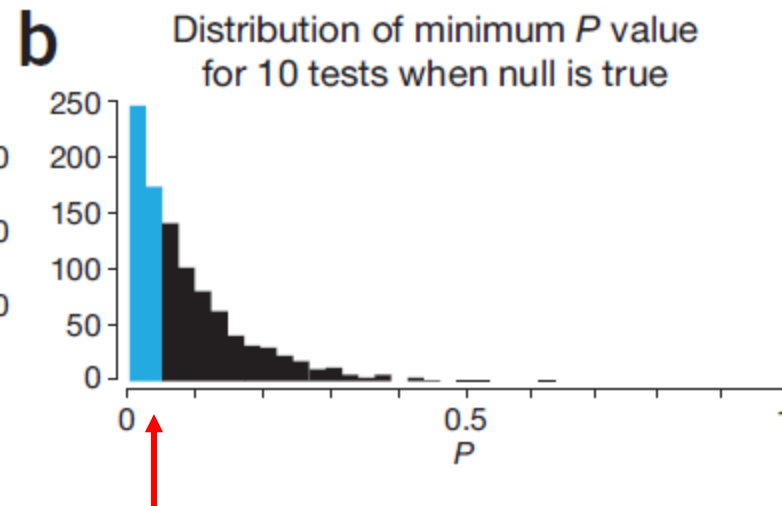
- Significance of experimental results often assessed using p-values and estimates of size effect.
- However, interpretation of assessment tools can be invalidated by selection bias when testing multiple hypotheses, fitting multiple models or informally selecting results that seem interesting after observing data (as seen). → p-value hacking, data dredging.
- Statistically significant results may not translate into biologically meaningful conclusions.
- Here, what is the smallest p value we can expect if the null hypothesis is true but we have done many tests, either explicitly or implicitly?

# P values – search for significance

- Study: 10 physiological variables are measured in 100 individuals.
- Determine whether any variable is predictive of systolic blood pressure.
- Assumptions: None of the variables are predictive in the population, all variables are independent.



One variable as predictor:  
5% of samples with  $p < 0.05$



Each variable as predictor:  
40% probability to find at least on sample with  $p < 0.05$ .

# P values – search for significance

- Selection bias and truth inflation → **of course, should correct for multiple comparisons.**
- However, what if we plot each predictor against SBP and feel that 1 predictor might have a quadratic relationship with SBP?
- Adjust for 10 comparisons (10 plots), 20 comparisons (linear or quadratic effects?), or more (account for nonlinear relationships)?
- **The more models we consider, the greater the danger of overfitting data and producing false-positives.**

# P values – search for significance

- During statistical analysis, we must carefully distinguish between using data to confirm inferences and using data to generate hypotheses.
- In confirmatory use, p-values and confidence intervals can be computed and interpreted as taught in basic statistic courses.
- In exploratory use, only if adjusted correctly for multiple testing (but how?) or selection → no simple and well-accepted ways of doing this.

# Interpreting $P$ values

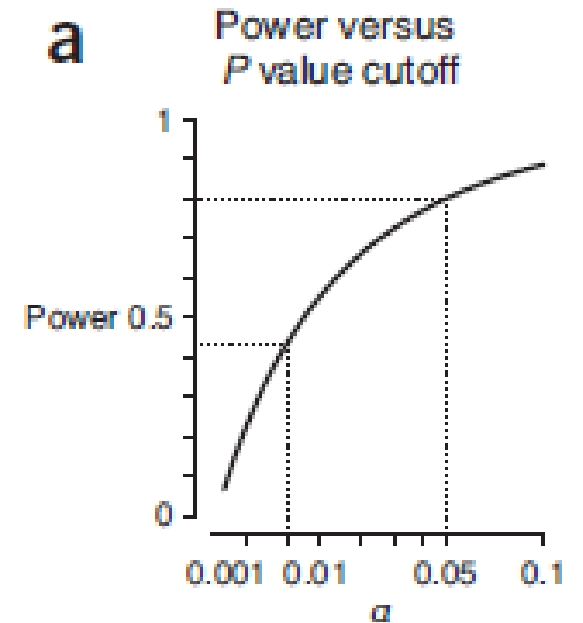
A  $P$  value measures a sample's compatibility with a hypothesis, not the truth of the hypothesis.

- $P$ -values are convenient.
  - Very easy to interpret (go/no go).
  - Can be misleading when taken as only metric.
  - Need to be supplemented with other information to avoid misinterpretation.
1. Use of more stringent  $p$ -value cutoffs supported by Bayesian analysis.
  2. Use of the observed  $p$ -value to estimate the false discovery rate (FDR).
  3. Combination of  $p$ -values and effect sizes to create more informative confidence intervals.
- } How strongly does data support null versus alternative hypotheses

} How strongly does data support the parameter values in the confidence interval

# Interpreting P values – cutoff + Bayesian

- A p-value is probability statement about the observed sample in the context of the hypothesis.
- E.g. does disease affect the level of a biomarker?
- P-value of comparison of the mean biomarker levels in healthy versus diseased would be the probability that a difference in means at least as large as the one observed can be generated from random samples if the disease does not affect the mean biomarker level.
- $H_0$ : Disease does not affect the mean biomarker level.
- Samples,  $n = 10$  individuals, randomly chosen from healthy and diseased populations under the assumption of normal distribution with  $\sigma^2 = 1$ .
- At this sample size, two-tailed t-test has 80% power to reject null at significance  $\alpha = 0.05$  when the effect size is 1.32.
- Result: We observe a difference in sample means of 1.2 with pooled standard deviations of 1.1. → **p-value 0.025**.

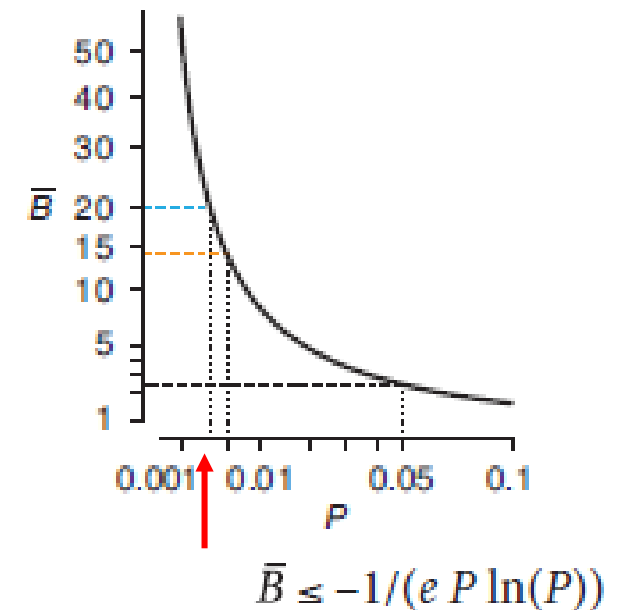




# Interpreting P values – cutoff + Bayesian

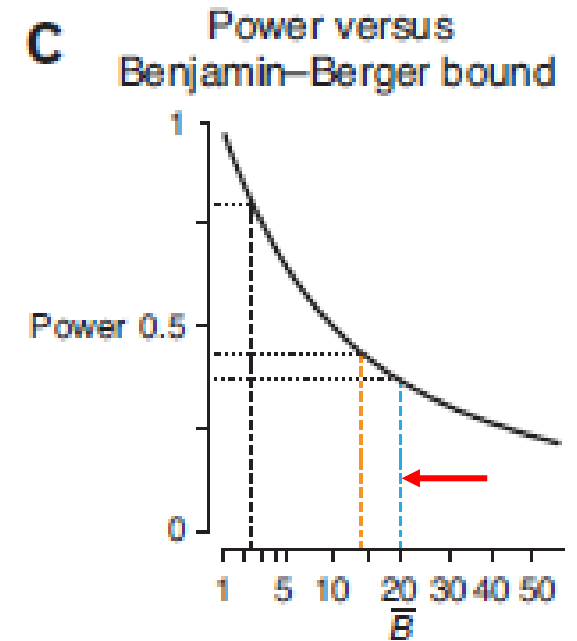
- Once p-value is computed, useful to assess the strength of evidence of the truth or falsehood of the null hypothesis. → Bayesian analysis.
- Decisions about statistical significance can be based on **Bayes factor, B**, which is ratio of average likelihoods under the alternative and null hypotheses.
- Usage of Bayesian analysis adds element of subjectivity because it requires the specification of a prior distribution for the model parameters under both hypotheses.
- The upper bound for the Bayes factor,  $\bar{B}$ , can be calculated using the p-value, and the prior does not need to be specified.  $\bar{B} = 10$  means that the alternative hypothesis is at most ten times more likely to be true than the null.
- Bayes factor of 20 or more is commonly considered to be strong evidence for the alternative hypothesis.
- For our example,  $p = 0.025$ , i.e. alternative hypothesis is at most  $\bar{B} \leq 3.9$  time more likely than null → Considered very weak evidence.
- For  $\bar{B} \geq 20$  (minimum),  **$p \leq 0.0032$** .

**b** Benjamin–Berger bound calibrates the *P* value



# Interpreting P values – cutoff + Bayesian

- Let us repeat our experiment using  $\alpha = 0.005$ .
- With same effect size of 1.32, power would only be 43%.
- To reach again a power of 80%,  $n = 18$  instead of  $n = 10$ .

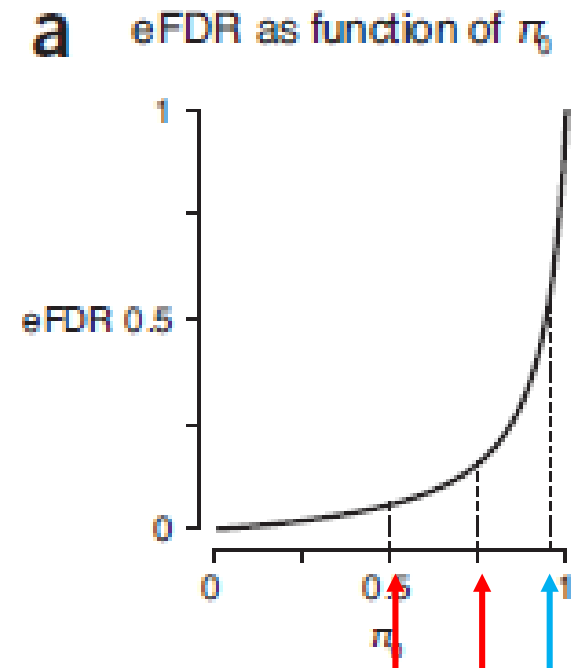


# Interpreting P values – FDR



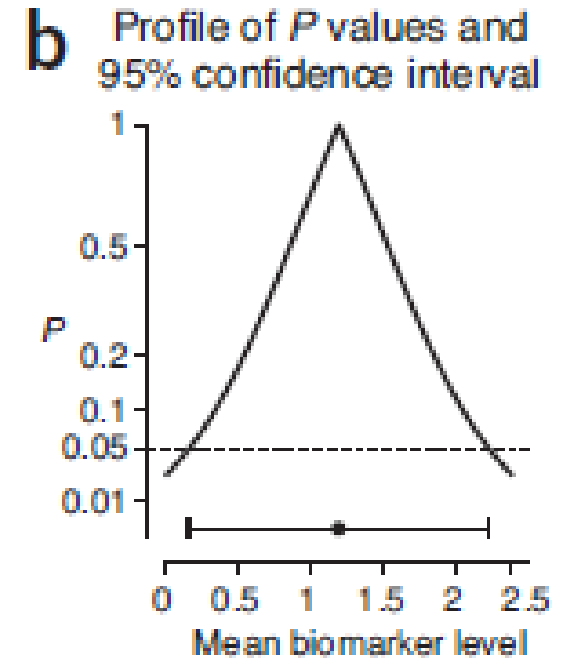
# Interpreting P values – FDR

- *Back to the biomarker example, which is low-throughput.*
- Suggestion to use the heuristic that  $\pi_0$  is probability that null hypothesis is true based on prior evidence, with  $\pi_0$  between **0.5** and **0.75** for primary hypotheses and at **0.95** for hypotheses formulated after exploration of the data (post-hoc test). → While proportion of tests that are truly null can be estimated from data in the high-throughput approach, low-throughput requires prior odds that investigator would be willing to put on the truth of the null hypothesis.
- **eFDR =  $p \pi_0 / (p \pi_0 + \beta(1 - \pi_0))$**
- $p = 0.025$  and  $\beta = 80\%$  → eFDR = **0.03**, **0.09**, **0.38**.
- For primary hypothesis, **3%** of tests where we reject the null hypothesis at this level of  $p$  are false discoveries (**i.e. 3% false-positives**)
- If we test only after exploring data, we would expect **38%** of discoveries to be false.



# Interpreting P values – confidence intervals

- Many investigators and journals advocate supplementing p-values with confidence intervals, which provide a range of effect sizes compatible with the observations.
- Mean biomarker levels that would not lead to significant results at  $\alpha = 0.05$  when observed level is 1.2.



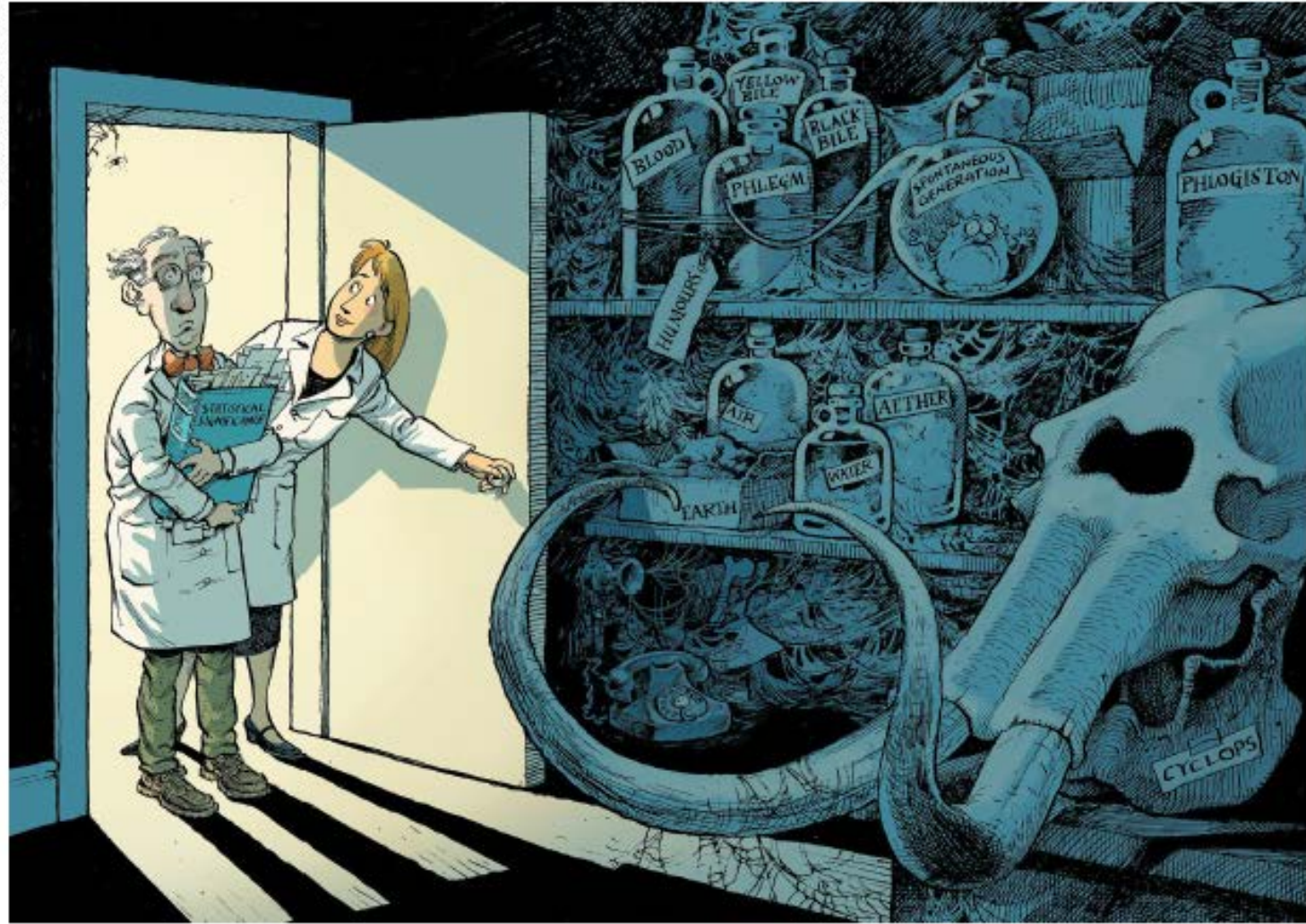
# Interpreting P values – summary

- P-values can provide a useful assessment of whether data observed in an experiment are compatible with the null hypothesis.
- Interpretation can be greatly assisted by accompanying heuristics, based on Bayes factor or FDR.
- Variability of the p-value from different samples points to the need to provide multiple sources of evidence before drawing scientific conclusions.

# Agenda for today's JC

- Biological data is often noisy – what are true differences and what are random differences?
- The p-value in its historical perspective.
- The problem of statistical significance – reproducibility crisis, p-value hacking, truth inflation
- **Recommendations for statistical testing.**





# 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.



ROBERT TAYLOR



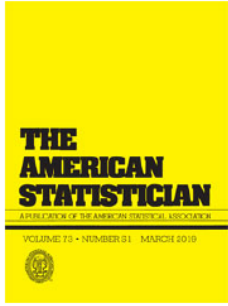
## Rein in the four horsemen of irreproducibility

*Dorothy Bishop describes how threats to reproducibility, recognized but unaddressed for decades, might finally be brought under control.*

More than four decades into my scientific career, I find myself an outlier among academics of similar age and seniority: I strongly identify with the movement to make the practice of science more robust. It's not that my contemporaries are unconcerned about doing science well; it's just that many of them don't seem to recognize that there are serious problems with current practices. By contrast, I think that, in two decades, we will look back on the past 60 years — particularly in biomedical science — and marvel at how much time and money has been wasted on flawed research.

How can that be? We know how to formulate and test hypotheses in controlled experiments. We can account for unwanted variation with statistical techniques. We appreciate the need to replicate observations.

Yet many researchers persist in working in a way almost guaranteed not to deliver meaningful results. They ride with what I refer to as the four horsemen of the reproducibility apocalypse: publication bias, low statistical power, *P*-value hacking and HARKing (hypothesizing after results are known). My generation and the one before us have done little to rein these in.



The American Statistician



ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: <https://www.tandfonline.com/loi/utas20>

## Moving to a World Beyond " $p < 0.05$ "

Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar

Review articles

## The reign of the $p$ -value is over: what alternative analyses could we employ to fill the power vacuum?

Lewis G. Halsey

Published: 22 May 2019 | <https://doi.org/10.1098/rsbl.2019.0174>

# Recommendations by the American Statistical Association

- Don't:
  - Base your conclusion solely on whether association or effect was found to be statistically significant.
  - Believe that an association or effect exists just because it was statistically significant.
  - Believe that " is absent.
  - Believe that p-value gives probability that chance alone produced the observed association or effect or probability that test hypothesis is true.
  - Conclude anything about scientific or practical importance based on statistical significance (or lack thereof).
- Do:
  - **A**ccept uncertainty
  - Be **t**houghtful
  - Be **o**pen
  - Be **m**odest
- = ATOM

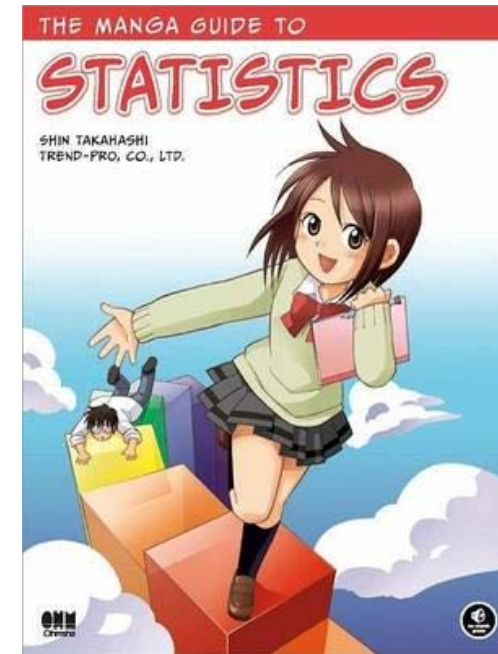
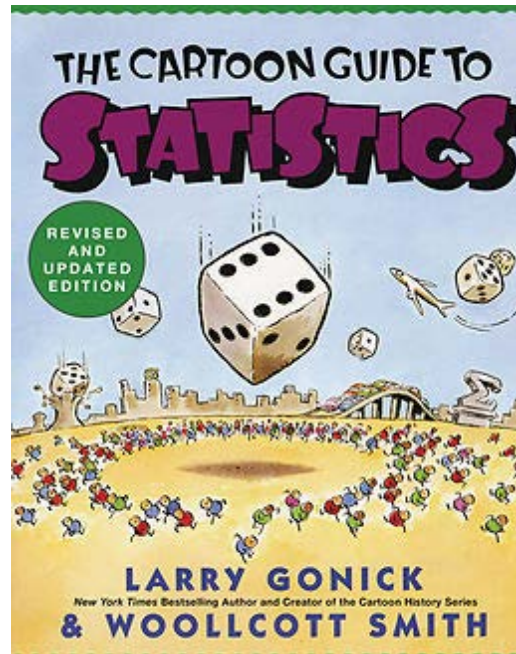
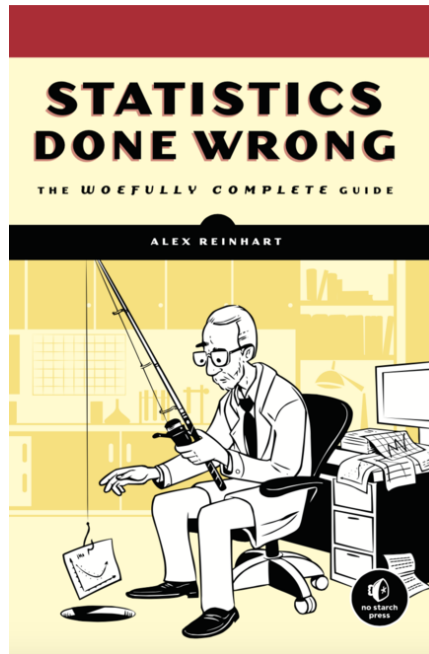
# Personal opinions

- Apart from often being misinterpreted, the p-value alone is not an adequate proxy to guide yes/no-decision making process.
- The interpretation of results has been often quite easy in the past, at the cost of scientific rigour.
- This ease of interpreting is most likely illusory.
- If we want to foster reproducible science, we will probably have to start with our own experiments (which probably means that we will have to publish more studies that openly endorse uncertainty, not considered sexy).
- I will try to implement some of the recommendations mentioned here (Bayesian approaches as a complement to p-values, e.g.).
- Our and future generations of scientists probably need much more (and in-depth) expertise in statistics, which will not just be a tool but an integral part of biomedical science.

# Sources

- Altman, N. (2017). P values and the search for significance.
- Gale, R. P. (2016). What is the (p-) value of the P-value?
- Gale, R. P. (2016). What is the P-value anyway?
- Baker, M. (2016). Statisticians issue warning on P values.
- Wasserstein, R. L. (2016). The ASA statement on p-values: Context, process, and purpose.
- Greenland, S. (2016). Statistical tests, p-values, confidence intervals, and power: A guide to misinterpretations.
- Wasserstein, R. L. (2019). Moving to a world beyond « $p < 0.05$ »
- Bishop, D. (2019). Rein in the four horsemen of irreproducibility.
- Altman, N. (2017). Interpreting P values.
- Leek, J. T. (2015). P values are just the tip of the iceberg.
- Benjamin, D. J. (2017). Redefine statistical significance.
- Amrhein, V. (2017). Remove, rather than redefine, statistical significance.
- Head, M. L. (2015). The extent and consequences of p-hacking in science.
- Amrhein, V. (2019). Retire statistical significance.
- Amrhein, V. (2019). Inferential statistics as decriptive statistics: There is no replication crisis if we don't expect replication.
- Various others (2017). Five ways to fix statistics.
- Halsey, L. G. (2019). The reign of the p-value is over: what alternative analyses could we employ to fill the vacuum?
- Simmons, J. P. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant.

# Concepts



- Reinhart, A. (2015). Statistics Done Wrong.
- Takahashi, S. (2009). The Manga Guide to Statistics.
- Gonick, L. (1993). Cartoon Guide to Statistics.

The test concludes:  
**not pregnant**

