# Should we abolish “statistical significance”?

## Replication problems in scientific research and the role of statistics

Presentation at the Methods Hub Seminar Series,

Faculty of Social Sciences, The University of Hong Kong

19 April 2024

Dr Peter Martin
Associate Professor in Biostatistics & Psychological Methods
Department of Primary Care and Population Health, University College London

# Overview

**Part 1: The first statistical hypothesis test** (What is a p-value and what is it not?)

**Part 2: Lies, damned lies and p-values** (Why are p-values under attack?)

**Part 3: A crisis of replication?** (Is science broken?)

**Part 4: The importance of replication** (Three cautionary tales)

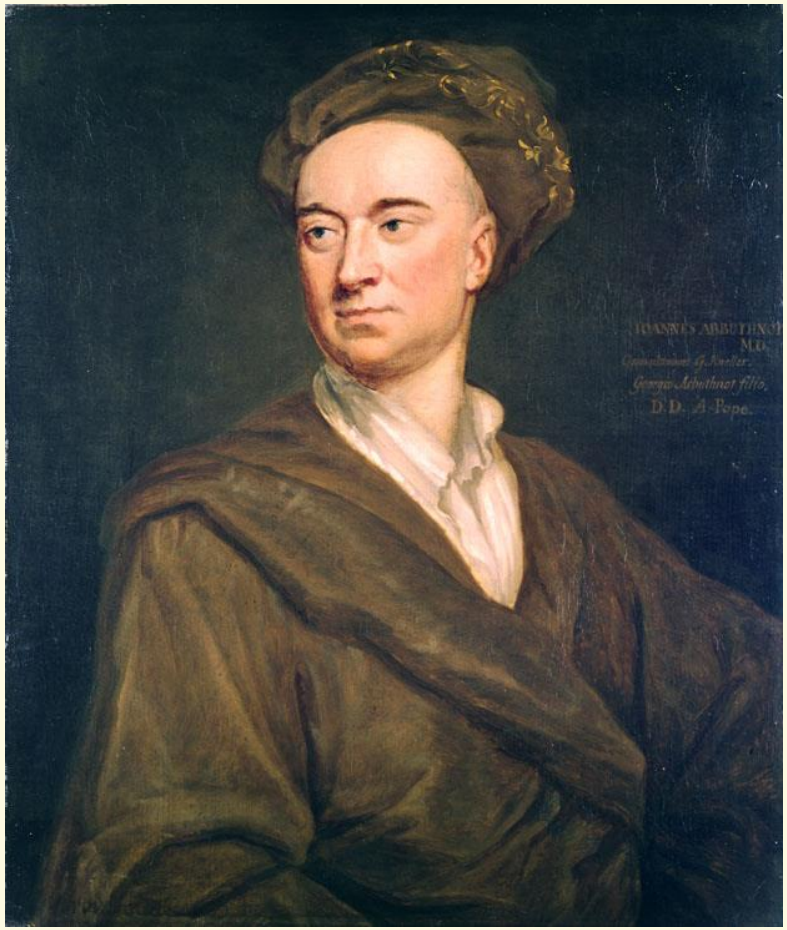
**Part 5: Does "statistical significance" cause false findings?** (Yes, sometimes.)

**Part 6: So what can we do?** (Proposals on how to do things better.)

# Part 1:

# The first statistical hypothesis test

## John Arbuthnot (1667 – 1735), Scottish physician and polymath



Arbuthnot conducted what is probably the first ever statistical hypothesis test (Gigerenzer & Marewski, 2014; Stigler, 2016).

# The first null hypothesis: the sex ratio at birth

Arbuthnot analysed 82 years of London birth statistics. In each of these 82 years, more boys than girls were born. He calculated the probability of this happening, if the male-female birth ratio were 1:1. This probability is:

$$p = \left(\frac{1}{2}\right)^{82} = 0.000000000000000000000002068.$$

“From whence it follows that it is Art, not Chance, that governs” (Arbuthnott, 1710, p. 189).

## Arbuthnot's conclusion

As some researchers today might say: "$p < 0.05$, so the result is statistically significant".

Arbuthnot interpreted his finding as evidence for a divine plan. He observed that boys and young men were more likely to die before they could marry than girls and young women. He theorized that divine intervention compensates by adjusting the sex ratio at birth.

"This Equality of Males and Females is not the Effect of Chance but Divine Providence, working for a good End […]" (Arbuthnott, 1710, p. 186).

# Limitations of hypothesis tests

Arbuthnot's test is useful and does address an interesting question, but it fails to answer many other interesting questions:

- He did not use his theory to predict specifically how many more boys than girls he expects to be born.
- He did not consider alternative explanations for his findings.
- He did not use his data to estimate the male-female birth ratio.

The use of significance tests today sometimes has the same shortcomings.

# Part 2:

# Lies, damned lies, and p-values?


The American Statistician ›
Volume 73, 2019 - Issue sup1: Statistical Inference in the 21st Century: A World Beyond p < 0.05
Listen
Articles
Abandon Statistical Significance
Blakeley B. McShane, David Gal, Andrew Gelman, Christian Robert & Jennifer L. Tackett
Pages 235-245 | Received 30 Oct 2017, Accepted 06 Sep 2018, Published online: 20 Mar 2019
Download citation
https://doi.org/10.1080/00031305.2018.1527253
Check for updates

nature International weekly journal of science



*NATURE* | RESEARCH HIGHLIGHTS: SOCIAL SELECTION

# Psychology journal bans *P* values

**Test for reliability of results 'too easy to pass', say editors.**

**Chris Woolston**

26 February 2015 | Clarified: 09 March 2015



A controversial statistical test has finally met its end, at least in one journal. Earlier this month, the editors of *Basic and Applied Social Psychology* (*BASP*) announced that the journal would no longer publish papers containing *P* values because the statistics were too often used to support lower-quality research [1].

Nature slips up …

*Nature* **519,** 9 (05 March 2015) | doi:10.1038/519009f

## - Clarifications

**Clarified:** This story originally asserted that "The closer to zero the *P* value gets, the greater the chance the null hypothesis is false." *P* values do not give the probability that a null hypothesis is false, they give the probability of obtaining data at least as extreme as those observed, if the null hypothesis was true. It is by convention that smaller *P* values are interpreted as stronger evidence that the null hypothesis is false. The text has been changed to reflect this.

… just to check whether we were paying attention?

# Common misunderstandings of statistical hypothesis tests

Many researchers misunderstand statistical hypothesis tests (Gigerenzer, 2004; Greenland et al., 2016)**:**

- confusion between statistical significance and scientific/clinical importance
- $p < 0.05$ is mistakenly interpreted as “proof” of an effect
- $p > 0.05$ is mistakenly interpreted as “proof” of absence of an effect
- It’s mistakenly assumed that the null hypothesis always has to specify “no effect” or “no difference”
- Failure to recognize conditions under which p-values are valid indicators of the strength of statistical evidence

# Part 3:

# A crisis of replication?


"… a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance." (Ioannidis, 2005b, p. e124)

# A crisis of replication?

- "Replication projects" have been conducted in several fields and have found low replication rates, eg:
  - Psychology (Open Science Collaboration, 2015)
  - Social science (Camerer et al., 2018)
  - Preclinical cancer research (Errington et al 2021)
  - Epidemiology and public health (Ioannidis, 2005a):
    - antioxidant vitamins (Lawlor et al 2004)
    - Workplace wellbeing programmes (Jones et al 2020)
- But overall, evidence on how much published research is false is still scarce (Lash, Collin, & Van Dyke, 2018)
- Replication, and failure of replication, is part of scientific progress (Lash et al., 2018)

# Part 4:

# The importance of replication,

# in three cautionary tales

# Cautionary tale 1: Power poses

Photo: Eric (HASH) Hersman, CC BY 2.0 license
( https://creativecommons.org/licenses/by/2.0/deed.en )

# Power poses: original study

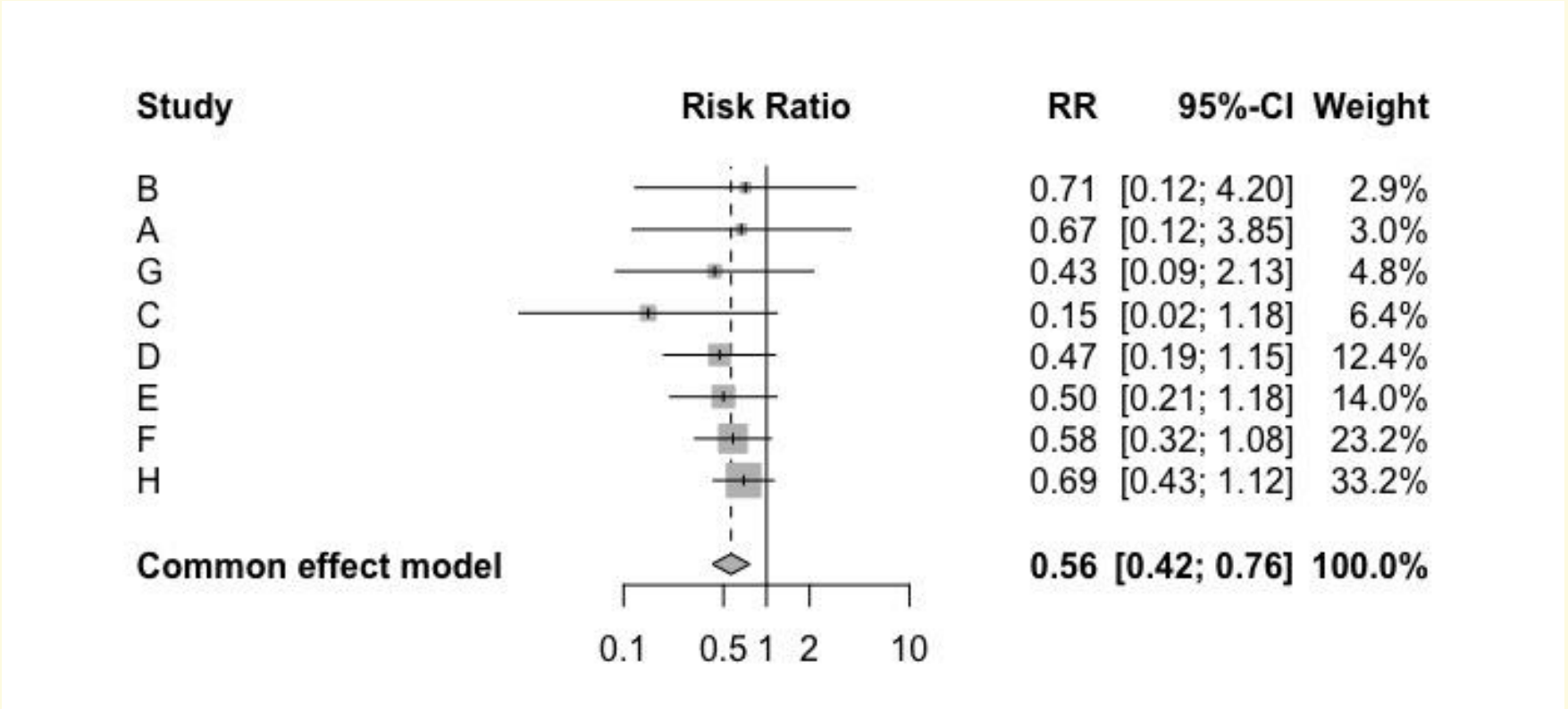
Carney et al (2010) published results from a small experiment (n = 42), which found evidence that holding a “high-power pose” for two minutes results in higher testosterone (p = 0.045) (among other things). Power posing was subsequently promoted as a way to boost performance in a popular TED talk by one of the co-authors, Amy Cuddy (pictured on previous slide).

Replications summarized by Cesario et al (2017) conclude that there is no evidence for an effect of power posing on testosterone, but some evidence of a small effect on self-reported feelings of power.

The initial study was methodologically well conducted (but see: Carney n.d.). It was a mistake, however, to take its results as established facts without replication.

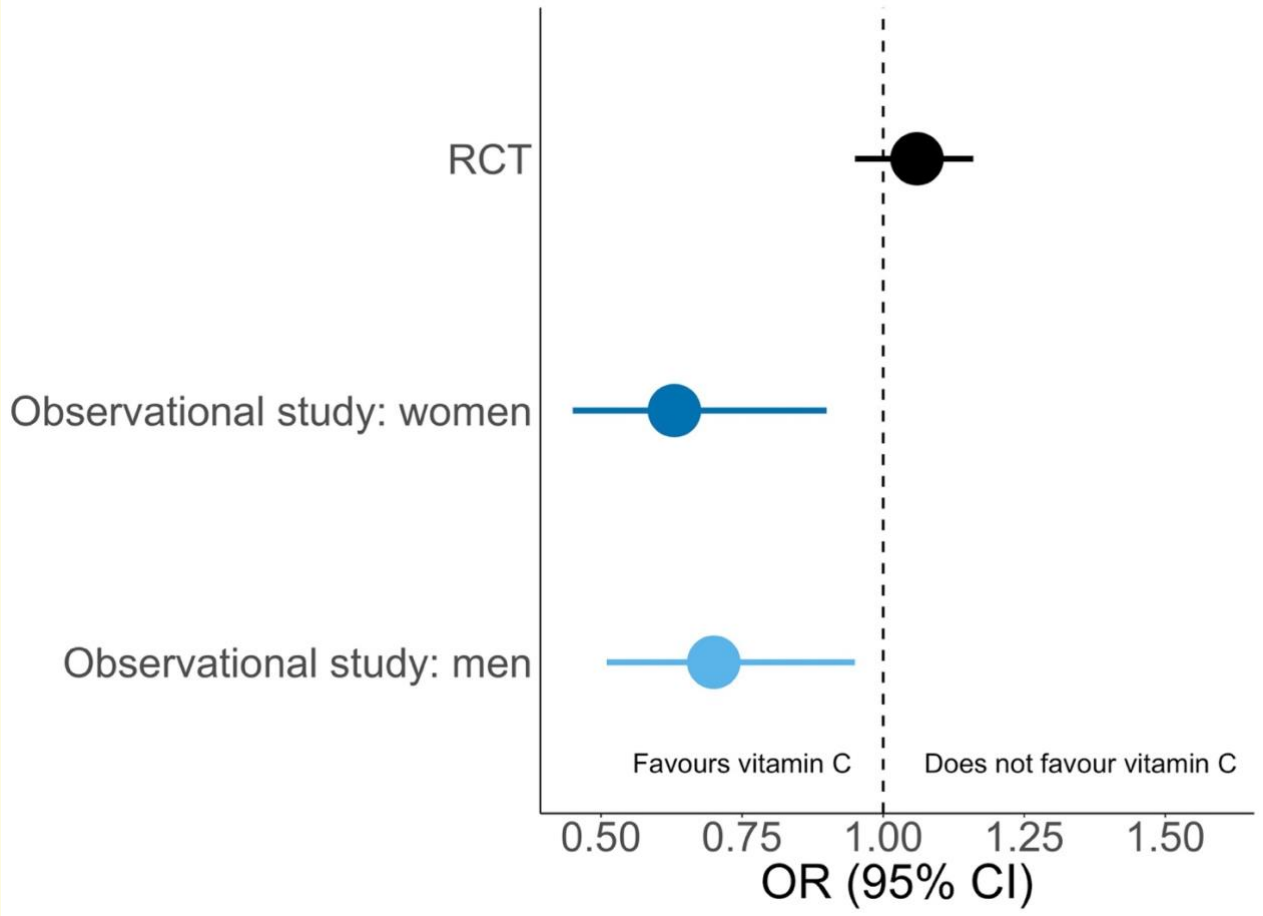
# Cautionary Tale 2: “Null findings”

(effect of exercise-based rehabilitation after coronary heart disease on death from cardiovascular disease – 3 year follow-up or longer)




(Dibben et al., 2021) *(adapted plot from Analysis 1.2.3; p. 221.)*

# Cautionary tale 3: Observational studies




The graph shows odds ratio estimates for the effect of vitamin C intake on the risk of coronary heart disease. Adapted from Lawlor et al (2004)

# Part 5: Does current use of “statistical significance” in research cause the publication of false results?

# “Significant findings”

Focus on p-values as the main (only?) measure of the strength of evidence for a finding risks ignoring many other factors that should be considered: related prior evidence, plausibility of mechanism, study design, data quality, …

All too often these are treated as “subordinate factors” in judging strength of evidence (McShane, Gal, Gelman, Robert, & Tackett, 2019).

In the minds of many authors and readers, “significance” trumps all.

# Mis-use of statistical hypothesis tests

The current research culture around statistical hypothesis tests can contribute to a distortion of evidence and may inhibit scientific progress. There are are least four ways in which this can happen.

1. Publication bias
2. Multiple testing
3. Fishing for statistical significance
4. Researcher degrees of freedom

# Publication bias

Most academic journals prefer to publish articles that present novel and surprising findings. Usually findings that are based on a "statistically significant" result seem more interesting than "null findings".

So studies with "statistically significant" results are more likely to get published. Researchers know this and may even save themselves the work of writing up and submitting a paper with a "null finding".

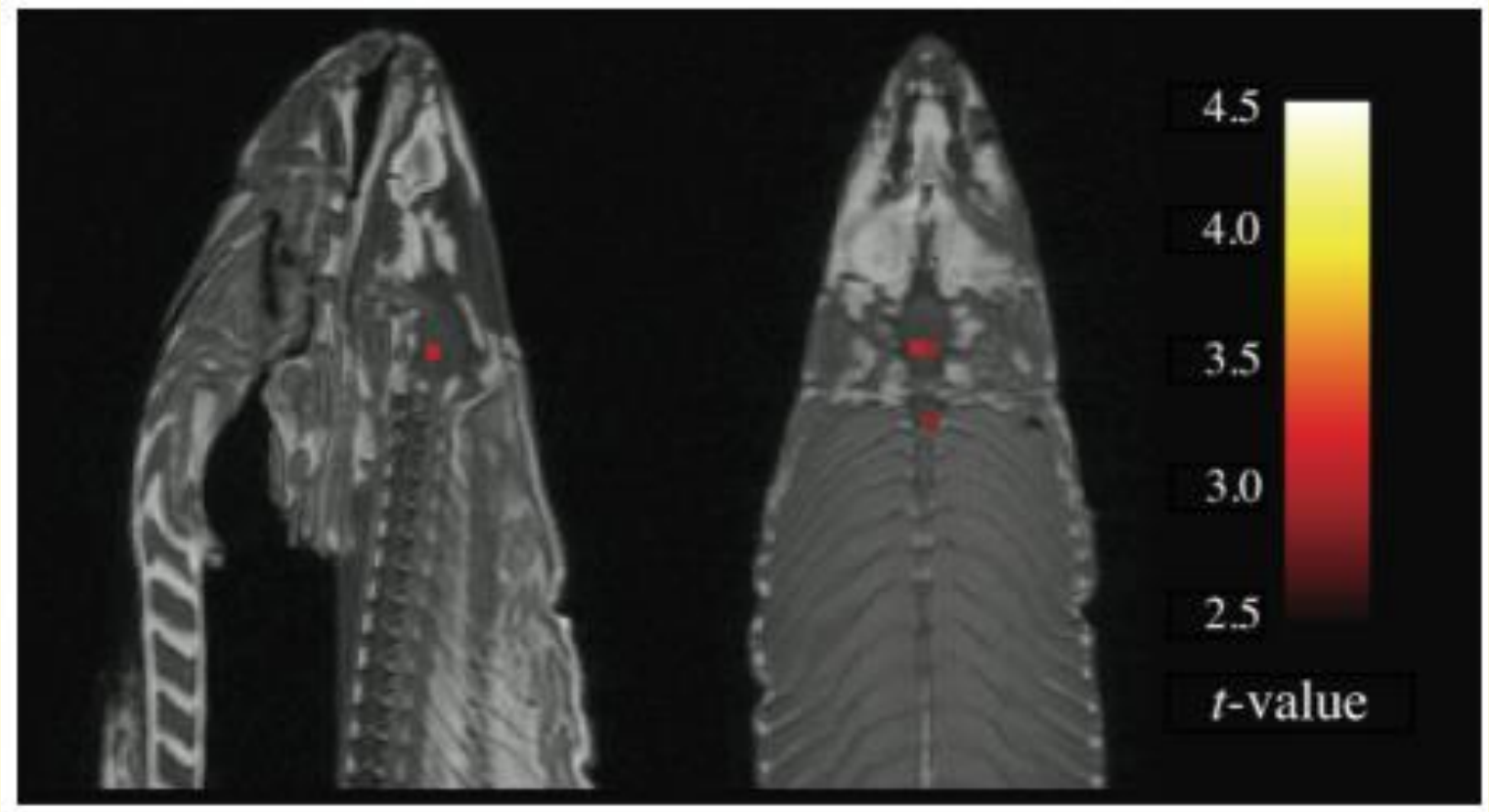
This leads to publication bias: for a given research question, looking only at published evidence creates a false impression of strong evidence in favour of an effect, while evidence to the contrary remains unpublished.

# Multiple tests and selective reporting

Increased risk of false findings can also result when studies selectively publish only analyses that yielded statistically significant results, and fail to report the analyses that yield non-significant result.

For example, in fMRI studies, if you scan sufficiently many regions of the brain, some will appear to exhibit “statistically significant” responses to stimuli. The more statistical hypothesis tests you conduct, the greater the risk of false positives. See next slide.

# An fMRI study finds that a dead salmon can understand human emotions




(Bennett, Baird, Miller, & Wolford, 2009)

# Fishing for statistical significance

If you design your study “cleverly”, you are almost guaranteed to get some p-value below the magic 5 % mark, even if all your null hypotheses are true.

For example, you might cherry-pick results: Analyse a large data set, test many different hypotheses, then select the ‘significant’ results for publication. Write your research report as if you had hypothesized the finding from the beginning.

This is called p-value hacking, or HARKing (Hypothesizing After the Results are Known), or “fishing for statistical significance”.

# Research finds that chocolate helps with weight loss


Image: Daily Star, 30/03/2015 (see: Bohannon, 2015)

The journalist John Bohannon intentionally used p-hacking to reveal how readily some media outlets publish research findings based on poorly conducted studies. His study was widely reported in many countries, before he revealed it as a spoof (Bohannon, 2015).

# Researcher degrees of freedom

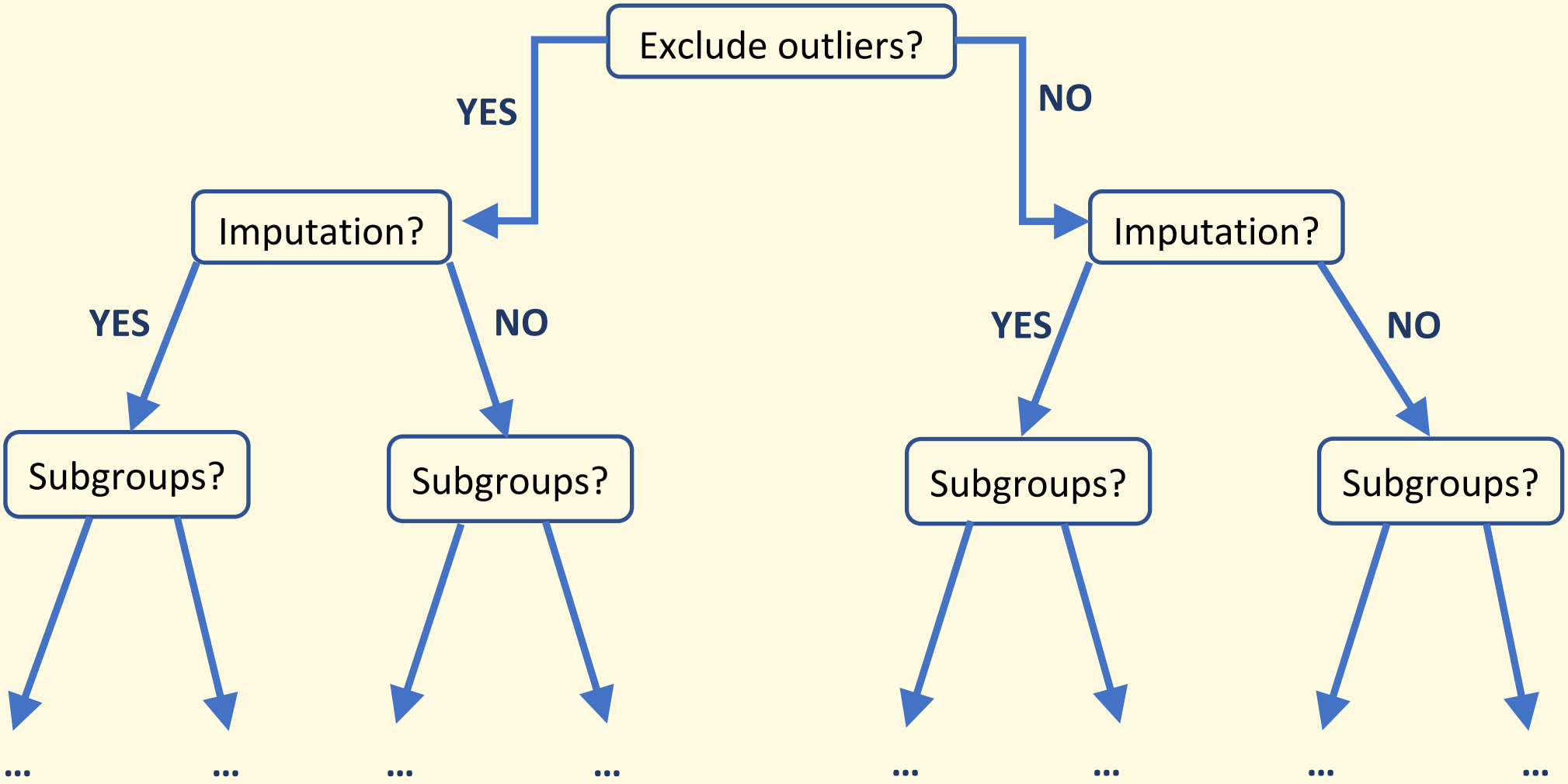
Data analysis involves many small decisions, such as:

- should I exclude outliers or not,
- should I transform variables before analysis or not,
- which subgroups should I analyse …?

Choices at different steps form a maze of potential combinations. The number of potential ways to analyse the data set can become quite large. Each analysis might yield a different p-value.

This phenomenon has been called “researcher degrees of freedom” (Simmons, Nelson, & Simonsohn, 2011), or the “garden of forking paths” (Gelman & Loken, 2013, 2014).

# The garden of forking paths



# The seductiveness of ‘statistical significance’

If a researcher expects to find a certain result, they may try to analyse the data in different ways until they find a combination of decisions under which their hypothesis is confirmed. They may be inclined to believe the “statistically significant” result more than other results.

This can lead to ‘false findings’, even if there is no intention to deceive.

If I use patterns in the data to decide which analyses to do or report, the p-values I find in the same data are invalid. ***A p-value is only a valid indication of the statistical evidence if it relates to a hypothesis that was specified before seeing the data.***

# Consequences of researcher degrees of freedom

Researcher degrees of freedom can have a large influence on findings.

Silberzahn et al (2018) gave the same data set to 29 research teams and asked them to investigate whether a football player's skin colour is related to the likelihood of the referee giving them a red card.

OR estimates ranged from 0.89 to 2.93. Twenty p-values were below 0.05, nine p-values were above 0.05.

Similar diversity in findings from different teams has also been reported in the analysis of fMRI data (Botvinik-Nezer et al., 2020).

# Part 6: So what can we do?

# Implications

- The p-value that you get from your statistical software of choice looks scientific, but may be entirely meaningless. It depends how you have conducted your study and analysis.
- It is not always possible to tell from published research reports how meaningful reported p-values are.
- Misinterpretation of p-values in scientific reports is common.
- Overreliance on hypothesis tests, over other statistical methods, is harmful to science (Gigerenzer, 2004; Gigerenzer & Marewski, 2014).
- The dichotomization of statistical evidence via a threshold (such as $p < 0.05$, $p < 0.01$) is useful in some contexts (eg regulation of medicines), but often misleading.

# Some proposed solutions

- **More emphasis on estimation** (confidence intervals)
- More emphasis on **other indicators of strength of evidence**
- **Transparency / open science**
- **Publication of study protocols and statistical analysis plans**
- **Registered research reports**
- **Improved design of observational studies**: e.g. emulate the target trial, directed acyclic graphs
- **Triangulation**
- **Bias analysis** and **sensitivity analysis** (Lash et al., 2014)
- **Negative controls** (Lipsitch, Tchetgen Tchetgen, & Cohen, 2010)

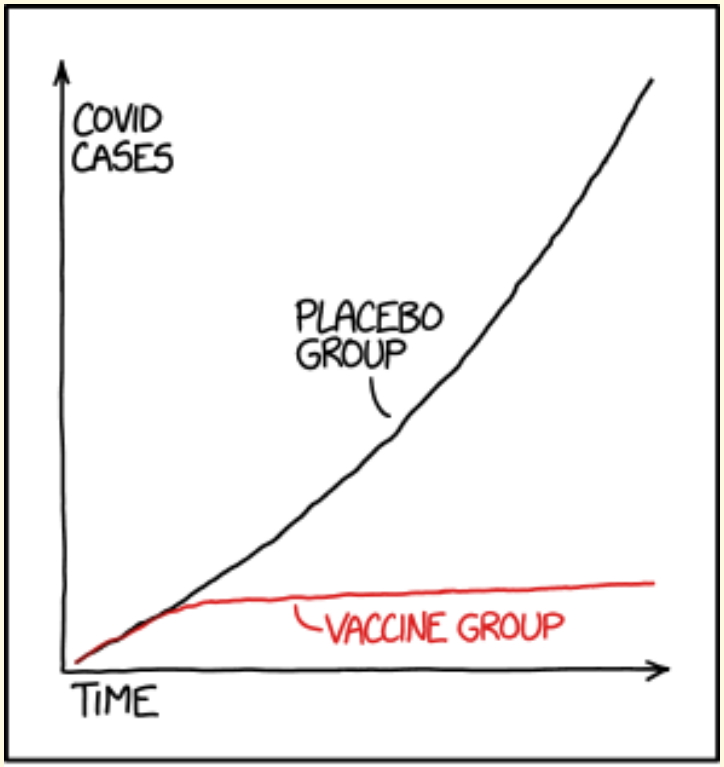
# A change in attitude

"[I]t seems to me that statistics is often sold as a sort of alchemy that transmutes randomness into certainty, an '**uncertainty laundering'** that begins with data and concludes with success as measured by statistical significance. [...]

[T]he solution is not to reform p-values or to replace them with some other statistical summary or threshold, but rather to move toward a greater acceptance of uncertainty and embracing of variation."

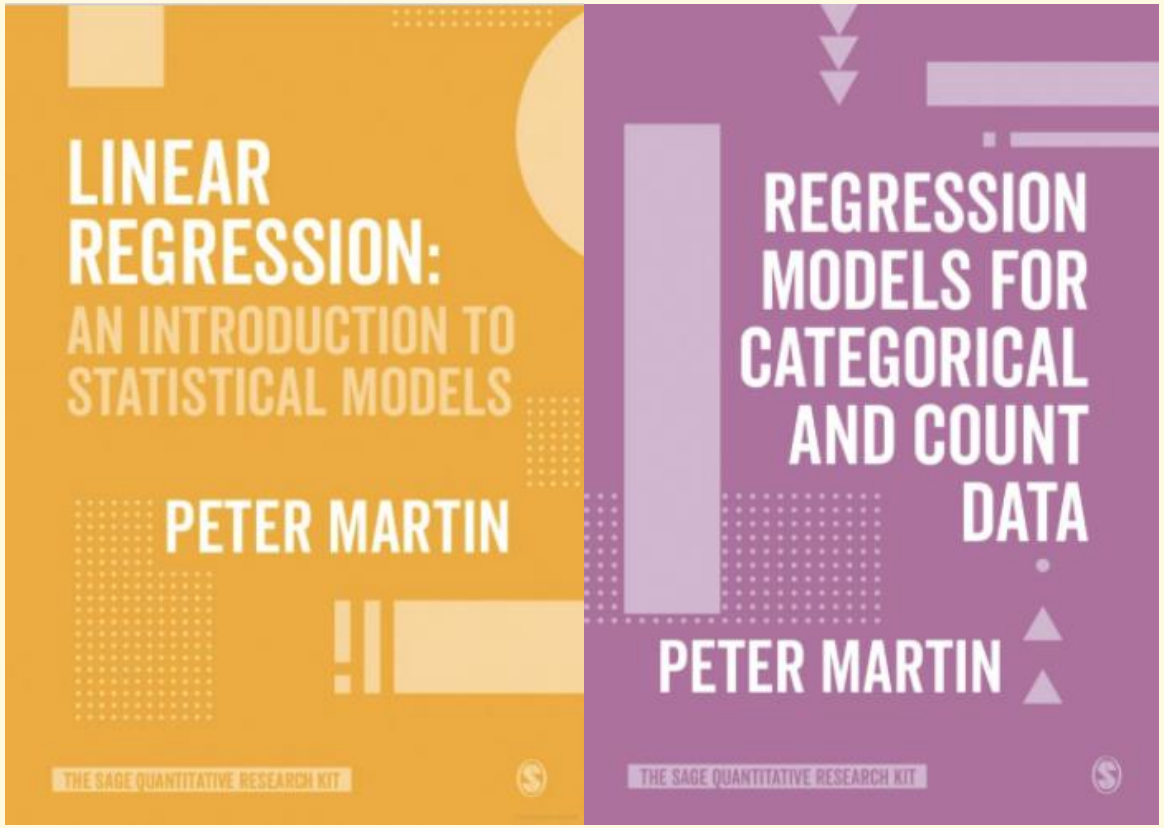
Andrew Gelman (2016, p. 2, emphasis added)

# A controversial idea?




( https://xkcd.com/2400/ )

## Implementing these ideas:




*The Sage Quantitative Research Kit, Vols 7 & 8* (Martin, 2021a, 2021b).
(Parts of this talk are based on Chapter 6 in the **purple book**.)

# References


Arbuthnott, J. (1710). An argument for Divine providence taken from the constant Regularity observ'd in the Births of both Sexes. *Philosophical Transactions of the Royal Society of London*, *27*, 186–190.

Bennett, C. M., Baird, A. A., Miller, M. B., & Wolford, G. L. (2009). Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: an argument for multiple comparisons correction. *Neuroimage*, *47*(Suppl. 1), S125.

Bohannon, J. (2015). I fooled millions into thinking chocolate helps weight loss. Here's how. Retrieved February 14, 2017, from http://io9.gizmodo.com/i-fooled-millions-into-thinking-chocolate-helps-weight-1707251800 .

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., … Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Camerer, C., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2018). Evaluating the Replicability of Social Science Experiments in Nature and Science. *Nature Human Behaviour*, *2*, 637–644.

Carney, Dana R., Cuddy, A. J. C., & Yap, A. J. (2010). Power Posing. *Psychological Science*, *21*(10), 1363–1368. https://doi.org/10.1177/0956797610383437

Carney, Dana Rose. (n.d.). My position on "Power Poses." Retrieved May 27, 2022, from https://faculty.haas.berkeley.edu/dana_carney/pdf_my position on power poses.pdf

Cesario, J., Jonas, K. J., & Carney, D. R. (2017). CRSP special issue on power poses: what was the point and what did we learn? *Comprehensive Results in Social Psychology*, *2*(1), 1–5. https://doi.org/10.1080/23743603.2017.1309876

Dibben, G., Faulkner, J., Oldridge, N., Rees, K., Thompson, D. R., Zwisler, A. D., & Taylor, R. S. (2021). Exercise-based cardiac rehabilitation for coronary heart disease. *Cochrane Database of Systematic Reviews*. John Wiley and Sons Ltd. https://doi.org/10.1002/14651858.CD001800.pub4

Errington, T. M., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Challenges for assessing replicability in preclinical cancer biology. *Elife*, e67995. https://doi.org/10.7554/eLife

Gelman, A. (2016). The problems with p-values are not just with p-values. *The American Statistician*, *70*(2), 1–2.
Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. https://doi.org/dx.doi.org/10.1037/a0037714
Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(November/December), 460–465.
Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033
Gigerenzer, G., & Marewski, J. N. (2014). Surrogate Science: The Idol of a Universal Method for Scientific Inference. *Journal of Management*, *41*(2), 421–440. https://doi.org/10.1177/0149206314547522
Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3
Ioannidis, J. P. A. (2005a). Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. *Journal of the American Medical Association*, *294*(2), 218–222. Retrieved from https://jamanetwork.com/
Ioannidis, J. P. A. (2005b). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8). https://doi.org/10.1371/journal.pmed.0020124
Jones, D., Molitor, D., & Reif, J. (2020). What do workplace wellness programmes do? Evidence from the Illinois Workplace Wellness Study. *The Quarterly Journal of Economics*, *134*(4), 1747–1791. https://doi.org/10.1093/qje/qjz023.Advance
Lash, T. L., Collin, L. J., & Van Dyke, M. E. (2018). The Replication Crisis in Epidemiology: Snowball, Snow Job, or Winter Solstice? *Current Epidemiology Reports*, *5*(2), 175–183. https://doi.org/10.1007/s40471-018-0148-x
Lash, T. L., Fox, M. P., Maclehose, R. F., Maldonado, G., Mccandless, L. C., & Greenland, S. (2014). Good practices for quantitative bias analysis. *International Journal of Epidemiology*, *43*(6), 1969–1985. https://doi.org/10.1093/ije/dyu149
Lawlor, D. A., Smith, G. D., Kundu, D., Bruckdorfer, K. R., & Ebrahim, S. (2004). Those confounded vitamins: What can we learn from the differences between observational versus randomised trial evidence? *Lancet*, *363*(9422), 1724–1727. https://doi.org/10.1016/S0140-6736(04)16260-0

Lipsitch, M., Tchetgen Tchetgen, E., & Cohen, T. (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, *21*(3), 383–388. https://doi.org/10.1097/EDE.0b013e3181d61eeb.Negative

Martin, P. (2021a). *Linear regression: an introduction to statistical models*. London: Sage.

Martin, P. (2021b). *Regression models for categorical and count data*. London: Sage.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon Statistical Significance. *The American Statistician*, *73*(S1), 235–245.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., … Nosek, B. A. (2018). Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Stigler, S. M. (2016). *The Seven Pillars of Statistical Wisdom*. Cambridge, MA & London: Harvard University Press.