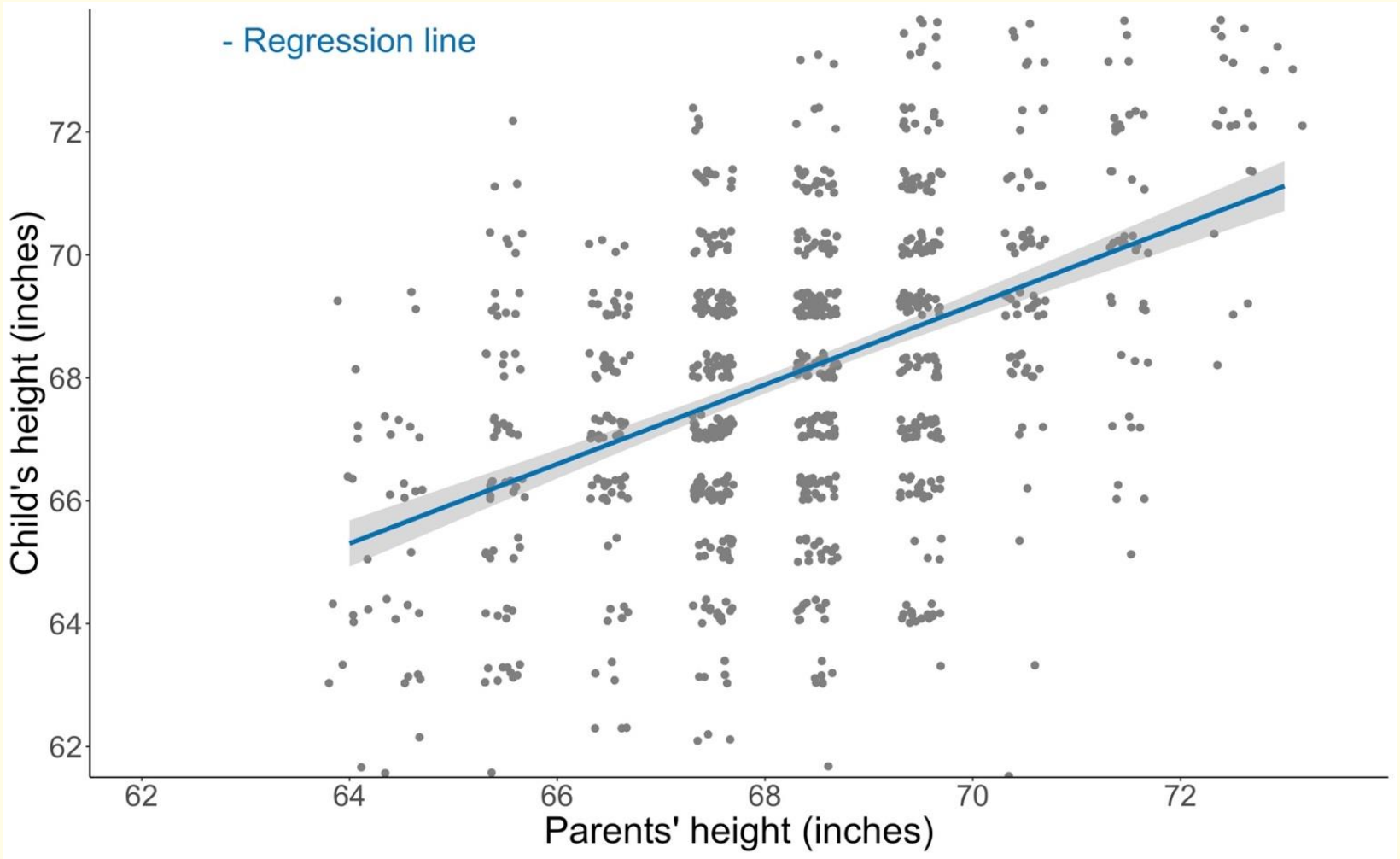
How should we teach “the world beyond $p < 0.05$”?

(to non-statisticians)

UK Conference on Teaching Statistics
Manchester, 13 June 2024

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

An example to start: mindless hypothesis testing



Data from Revelle (2015) and Galton (1886). A small jitter was applied to Galton's original data.

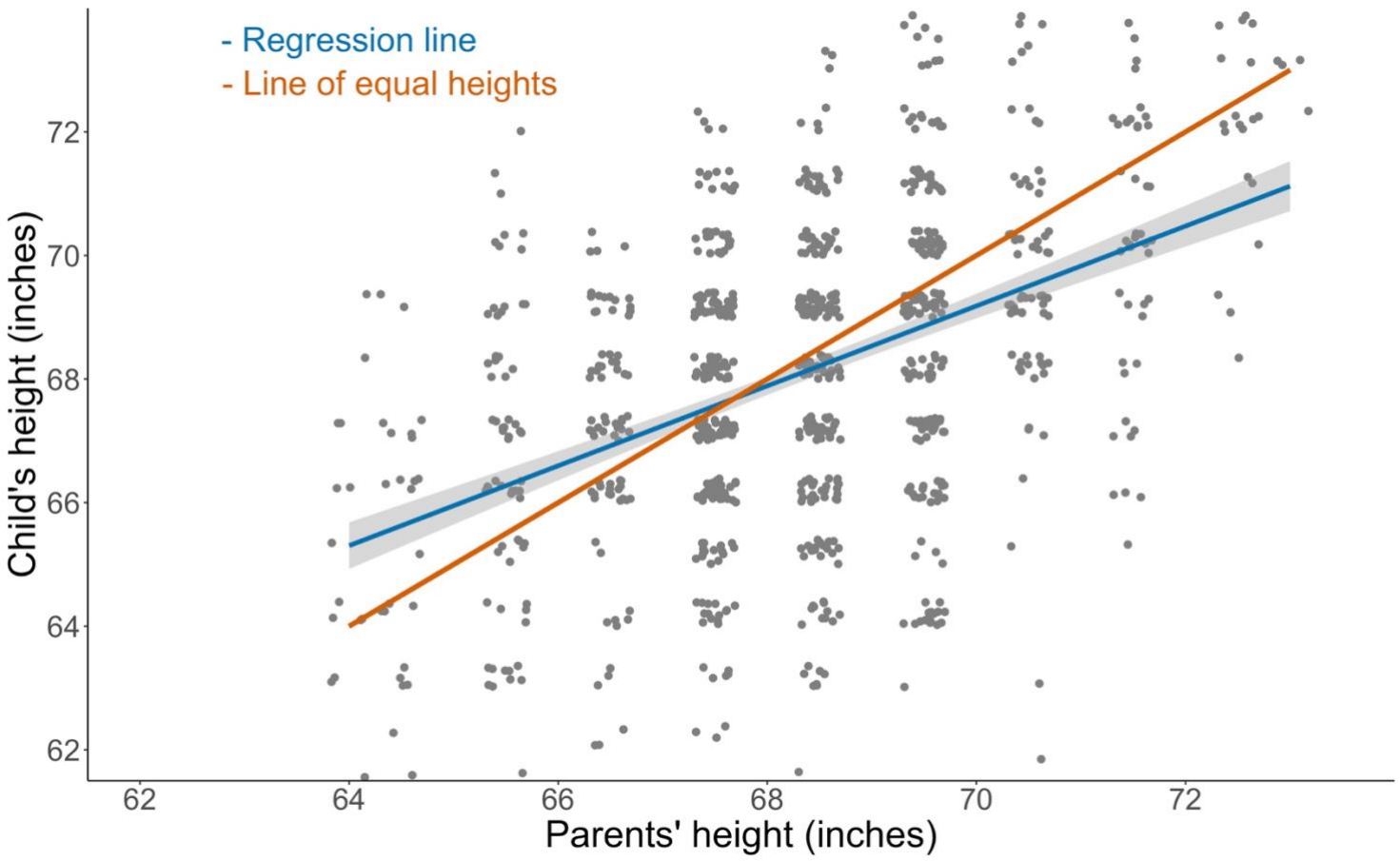
You can calculate correct p-values …
but utterly miss the scientific point

Table 1: Linear regression of child's height on parents' height (n = 928)

	Estimate	Std Error	t	df	p
Intercept	23.942	2.811	8.517	926	< 0.001
Parent's height	0.646	0.041	15.711	926	< 0.001

"Parental height has a positive association with children's height ($p < 0.001$)."

Galton’s demonstration of regression to the mean



Moving to a “world beyond p < 0.05”

The American Statistician, Volume 73, Issue sup1 (2019)

Statistical Inference in the 21st Century: A World Beyond p < 0.05

Editorial

Editorial

Moving to a World Beyond “$p < 0.05$” ›

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

Pages: 1-19

Published online: 20 Mar 2019

Screenshot from: https://www.tandfonline.com/toc/utas20/73/sup1?nav=tocList [2021-08-17]

So what to do?

Wasserstein, Schirm, & Lazar (2019) summarized their recommendations for good practice in statistical inference in four principles and seven words:

“Accept uncertainty. Be thoughtful, open, and modest.” [ATOM]

But how to teach this?

Guidelines on teaching statistics

Guidelines for Assessment and Instruction in Statistics Education (Carver et al., 2016), endorsed by the American Statistical Association:

1. **Teach statistical thinking.**
2. **Focus on conceptual understanding.**
3. **Integrate real data with a context and a purpose.**
4. Foster active learning.
5. Use technology to explore concepts and analyze data.
6. Use assessments to improve and evaluate student learning.
7. **Teach statistics as an investigative process of problem-solving and decision-making.**
8. Give students experience with multivariable thinking.

Link statistical inference to its role in the process of science

1. Link Type 1 / Type 2 errors to direct replication
2. Link unobserved biases to conceptual replication
3. Explicitly challenge misconceptions about p-values and “significance”
4. Show how shortcomings of NHST lead to distortions in published results
5. Emphasize the role of statistical inference within the scientific process
6. Demonstrate the importance of estimation
7. Encourage thoughtfulness about hypotheses

1. Link direct replication as a principle of good science with types of errors in statistical inference

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

Type 1 error:
Even if you do everything right, you may still be wrong

Carney et al (2010) published results from a small experiment (n = 42), which found claimed to have found evidence that holding a “high-power pose” for two minutes results in higher testosterone (p = 0.045).

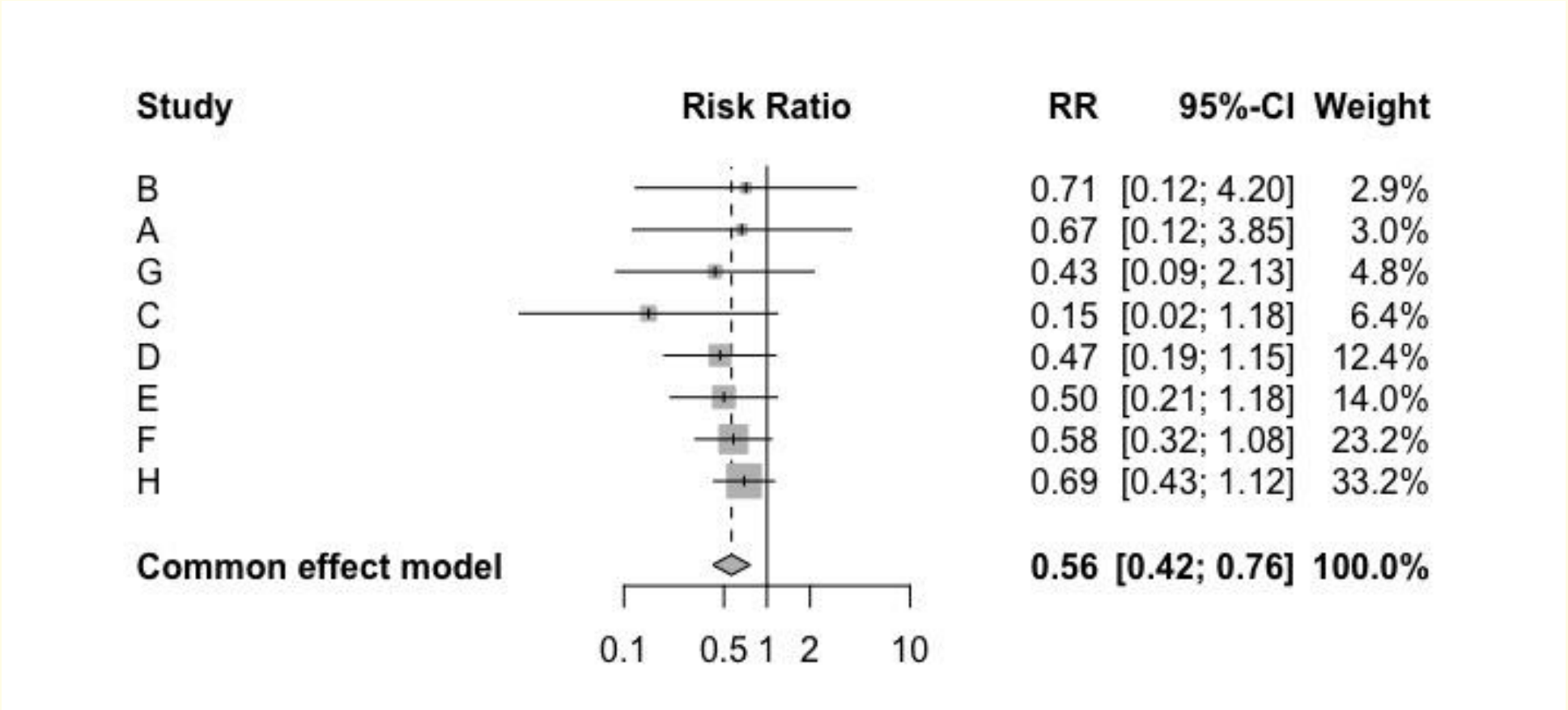
Power posing was then promoted as a way to boost performance in a popular TED talk by one of the co-authors, Amy Cuddy (pictured on previous slide).

Replications summarized by Cesario et al (2017) conclude that there is no evidence for an effect of power posing on hormones.

All indications are that the original study was methodologically robust and well-conducted (but see: Carney n.d.).

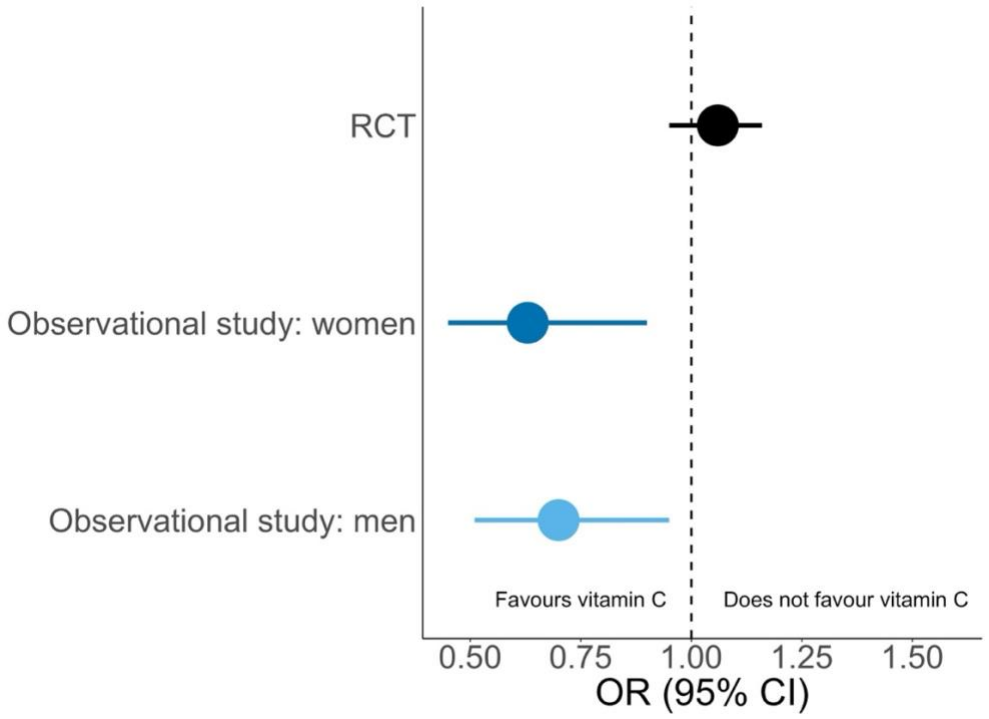
Type 2 error: “Null findings”

(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.)*

2. Link the importance of conceptual replication with biases that cannot be seen by the analyst of one data set



Estimates for the effect of vitamin C intake on the risk of coronary heart disease, adapted from Lawlor et al (2004). If there is bias in the data, even the best analysis gives a biased result.

3. Explicitly challenge misconceptions

Exercise: A glioma is a type of brain tumor. Statins are a group of medicines that reduce the production of low-density lipoprotein ("bad") cholesterol in the liver. Two case-control studies claimed to have found evidence that long-term use of statins has a protective effect against glioma. The following effect sizes were reported:

- **OR 0.72 (95 % CI 0.52 – 1.00;** Ferris et al 2012**)**
- **OR 0.76 (95 % CI 0.59 – 0.98;** Gaist et al 2013**)**

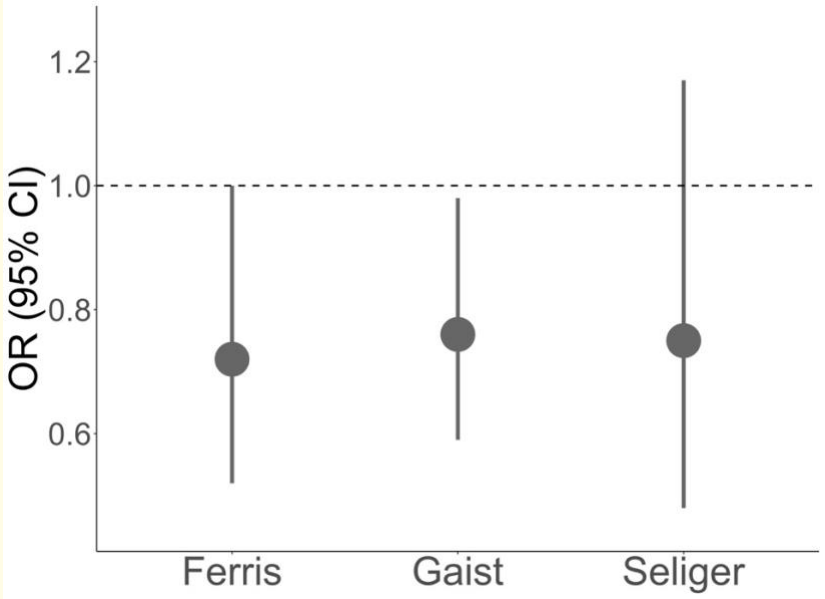
Seliger et al (2016) conducted a new case-control study and found:

- **OR 0.75 (95 % CI 0.48 – 1.17)**

Seliger et al (2016) concluded that **"[o]ur findings do not support previous sparse evidence of a possible inverse association between statin use and glioma risk**." (p. 947, highlight added)

Comment on this conclusion. Do you agree?

The cognitive habit of “checking for statistical significance” can blind us to the obvious



In fact, the new study supports, rather than contradicts, the previous studies (Lash, Collin, & Van Dyke, 2018). Presumably Seliger et al (2016) used the 95 % CIs to determine “statistical significance” and equated the lack of it in their own study with a lack of evidence for an effect.

4. Demonstrate shortcomings of null hypothesis significance testing

Give specific examples of misleading conclusions or fake results in the scientific publication record caused by narrow focus on p-values as indicators of strength of evidence:

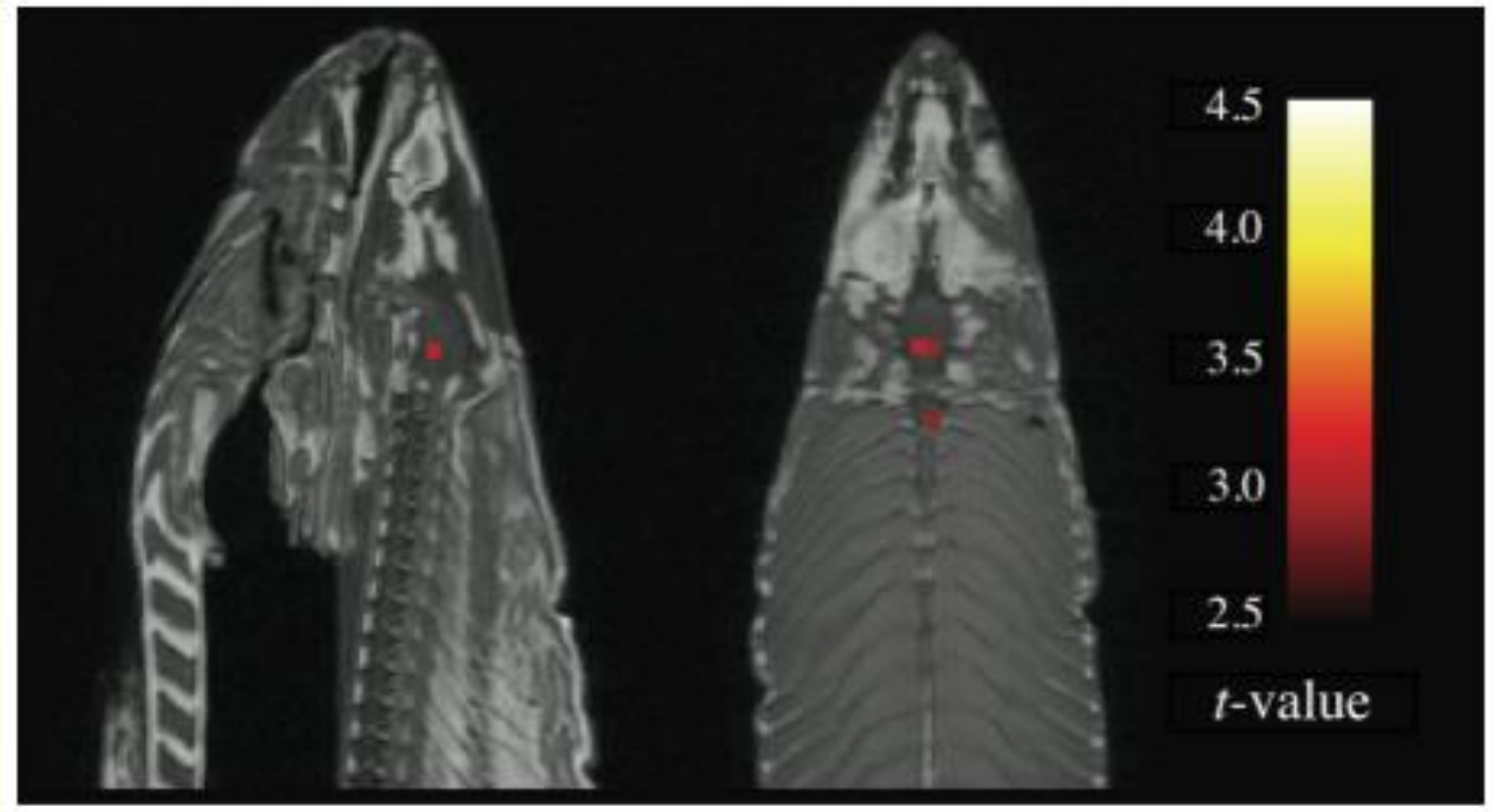
- publication bias
- multiple testing and selective reporting
- p-hacking
- “researcher degrees of freedom” (Simmons, Nelson, & Simonsohn, 2011), aka “the garden of forking paths” (Gelman & Loken, 2013)

P-values can be gamed & most people don't pay attention to the conditions under which p-values are valid

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).

Dangers of multiple testing: an fMRI study finds that a dead salmon can understand human emotions



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

5. Emphasize the difference between statistical inference and inference to a scientific explanation

John Arbuthnot (1667 –1735) observed that males were more likely to die before they could marry than females. But shouldn't God have arranged it so that there is one man for every woman?

Arbuthnot analysed 82 years of London birth statistics. More boys than girls were born in every year. If the boy-girl ratio was 1, this observation would have probability:

$$p = \left(\frac{1}{2}\right)^{82} = 0.00000000000000000000000002068.$$

[More detail in: Stigler (2016) and Gigerenzer & Marewski (2014).]

Your finding may be correct, but your theory may yet be uncertain

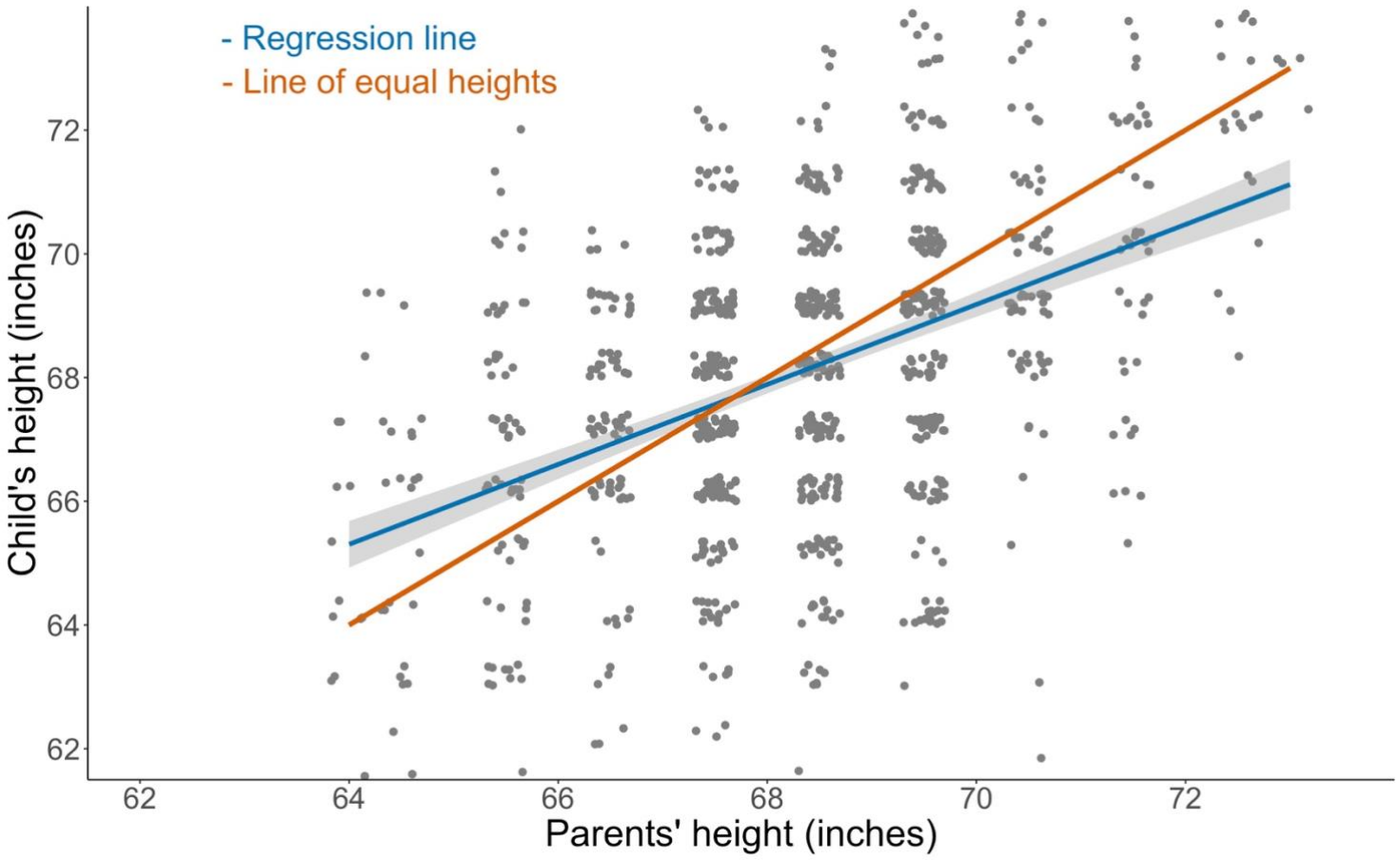
“Among innumerable Footsteps of Divine Providence to be found in the Works of Nature, there is a very remarkable one to be observed in the exact Ballance that is maintained, between the Numbers of Men and Women; for by this means it is provided, that the Species may never fail, nor perish, since every Male may have its Female, and of a proportionable Age. This Equality of Males and Females is not the Effect of Chance but Divine Providence, working for a good End […]” (Arbuthnott, 1710, p. 186).

6. Show the power of estimation

Arbuthnot could have used his data for a more interesting purpose: to estimate the sex ratio at birth. The best estimate is currently about 1.05 (male to female), although there are well-evidenced variations over time ("returning soldiers effect"). There may also be some regional variation. Strong divergence from 1.05 in some regions may be evidence for sex-selective abortion (Chao, Gerland, Cook, & Alkema, 2019).

"Rejecting the null hypothesis" might be a lot less interesting than estimating a parameter.

7. Encourage thoughtfulness about hypotheses



How about testing hypotheses other than “no effect”?

In this example, we could more meaningfully test H_0: $\beta = 1$, and thus find evidence to support Galton’s conclusion of “regression to the mean” (t = 8.599, df = 926, p < 0.0001).

Note that the statistical software you teach (e.g. Stata, SPSS, R), may not be capable of calculating *p*-values for user-specified null hypotheses for most types of statistical models, and almost certainly will assess ‘no difference/no effect’ null hypotheses by default. This offers opportunity to discuss the limits of (default settings in) software.

Summary:
thoughts on communicating statistical inference

- Tell stories
- Link the process of statistical inference to the process of science
- Emphasize uncertainty
- Give concrete examples:
  - **negative examples** of things that have go wrong because of mis-use of statistical inference procedures
  - **positive examples** of meaningful use of statistical inference
- Be consistent and clear in your language – reinforce good habits and thoughtful interpretation (cf. good suggestions in: Watt, 2020).
  - More difficult than it might seem …

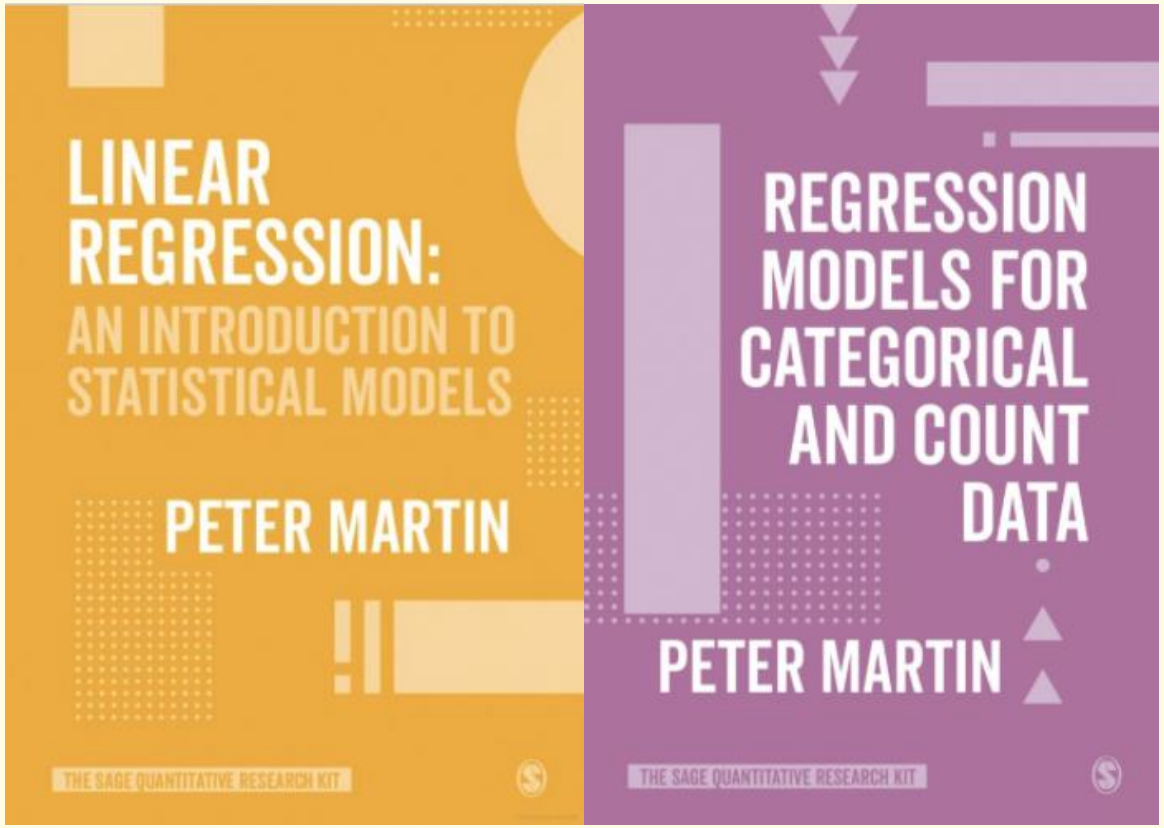
Uncertainty laundering

“[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)

Attempts at implementing these ideas are published as:



The Sage Quantitative Research Kit, Vols 7 & 8 (Martin, 2021a, 2021b)

Bibliography

Original graphs were made in R *(R Core Team, 2019) using the* ggplot2 *package (Wickham, 2016).*

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 .

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

Carver, R., Eversen, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., … Wood, B. (2016). *Guidelines for Assessment and Instruction in Statistics Education: College Report 2016*. *American Statistical Association*. Retrieved from http://www.amstat.org/education/gaise

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

Chao, F., Gerland, P., Cook, A. R., & Alkema, L. (2019). Systematic assessment of the sex ratio at birth for all countries and estimation of national imbalances and regional reference levels. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(19), 9303–9311. https://doi.org/10.1073/pnas.1812593116

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

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246–263.

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

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

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

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

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

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

R Core Team. (2019). R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Revelle, W. (2015). psych: Procedures for Personality and Psychological Research. Evanston, Illinois: Northwestern University. Retrieved from http://cran.r-project.org/package=psych

Seliger, C., Meier, C. R., Becker, C., Jick, S. S., Bogdahn, U., Hau, P., & Leitzmann, M. F. (2016). Statin use and risk of glioma: population-based case–control analysis. *European Journal of Epidemiology*, *31*(9), 947–952. https://doi.org/10.1007/s10654-016-0145-7

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.

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

Watt, H. C. (2020). Reflections on modern methods: Statistics education beyond 'significance': novel plain English interpretations to deepen understanding of statistics and to steer away from misinterpretations. *International Journal of Epidemiology*, 1–6. https://doi.org/10.1093/ije/dyaa080

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag.