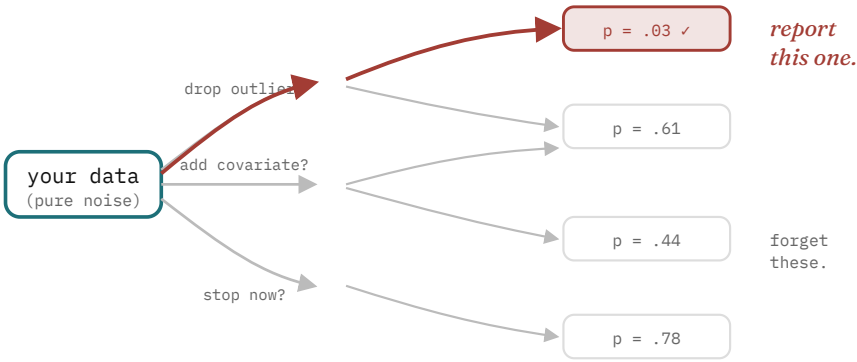
Statistics & the Art of *Not Fooling Yourself*

Given enough defensible choices, many datasets contain a path to p < .05. The whole discipline is the struggle not to walk it by accident.



The garden of forking paths. The data are pure noise — yet enough branches guarantee one ends in a "discovery."

In 2011 three psychologists set out to prove something impossible. They sat twenty undergraduates down, played half of them the Beatles' *"When I'm Sixty-Four"* and the other half a control tune, then asked a battery of questions — including each person's date of birth. After a perfectly ordinary statistical analysis, they announced their finding: listening to *"When I'm Sixty-Four"* made people **a year and a half**

younger. Not feel younger. Be younger — their birth dates said so, at $p = .04$.

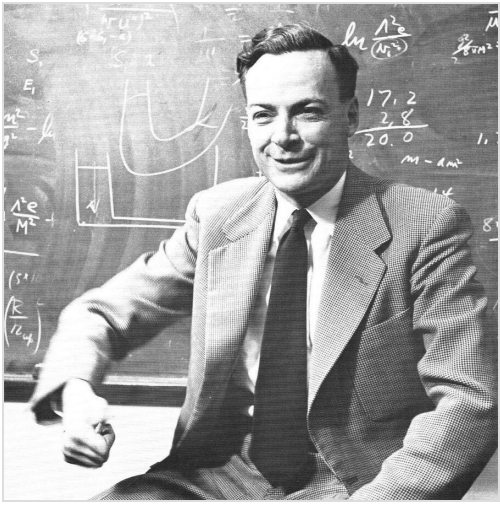
This is, of course, insane. A song cannot reach back and edit the year you were born. And that was exactly the point. Joseph Simmons, Leif Nelson, and Uri Simonsohn had used only standard, respectable tools, made only choices that any working scientist makes every week — and conjured a statistically significant impossibility. Their paper was a deliberately absurd demonstration designed to expose a vulnerability. Today we learn what it exposed, and how a science learns to stop fooling itself.

On **Day 2** we met the replication crisis as a set of bruising numbers (97% of original psychology studies "worked"; only 36% replicated) and named *p-hacking*[1] without yet opening it up. Today we open it up. This is the statistical engine room beneath that crisis — and beneath **Day 1**'s Gettier worry (a belief that's true by luck, not connection) scaled up to entire literatures. We'll lean hard on **Day 4**'s lesson that a *p*-value is not the probability the null is true (that confusion is the base-rate fallacy in disguise), and on **Day 5**'s warning that *which variables you control for* is itself a fork in the road. Threads today: *information* (signal vs noise) and *computation* (the lab as a fallible inference machine), with a quiet preview of *emergence* — science as an error-correcting system bigger than any one analyst.

THE HOOK

The easiest person to fool

> *"The first principle is that you must not fool yourself — and you are the easiest person to fool." — Richard Feynman, Caltech commencement, 1974*



Richard Feynman at a blackboard in 1959. The quote above comes from his 1974 Caltech commencement address on "cargo cult science." Image: *The Big T* 1959, public domain in the United States via Wikimedia Commons.

Feynman was talking about a kind of integrity that no formula can supply. The cruel twist of modern statistics is that the formulas, used the way humans naturally use them, become accomplices. They lend the warm authority of mathematics to whatever you were already hoping to find.

The villain has a clinical name: *researcher degrees of freedom*[2]. Every real analysis hides dozens of small, defensible decisions. Which outliers to exclude? Should you control for age? Sex? Both? Do you stop collecting data now, or run twenty more participants? You measured three outcomes — which do you report? Each choice, taken alone, is innocent. Taken together, made *after* you've glanced at how the numbers are leaning, they become a machine for minting false discoveries.

Simmons, Nelson & Simonsohn proved this with arithmetic before they proved it with the Beatles. By simulation, they showed that bundling just four ordinary degrees of freedom — peeking at the data and adding more if needed; choosing between two related outcome measures; adding or dropping a covariate like gender; dropping one of three experimental conditions — can inflate your chance

of a false positive from the advertised **5%** to as high as **61%**. Flip enough innocent switches and a coin-flip's worth of noise becomes a near-certainty of "success."

The reference table below shows the same machine in static form: each added research freedom gives pure noise another route to look like a discovery.

The False-Positive Factory

ANALYTIC FREEDOM	WHAT CHANGES	WHY IT INFLATES FALSE POSITIVES
Two outcome measures	Measure two related results and report the one that works.	Noise gets two chances to cross $p <$ .05.
Optional stopping	Peek at n=20, then add participants if needed.	The stopping rule itself becomes another path through the data.
Flexible covariate	Try the overall result and subgroup splits.	Legitimate controls become multiple comparisons when chosen after seeing the data.
Drop a condition	Run three groups, then report the best pair.	The reported contrast is selected because it looks best.

Simmons, Nelson, and Simonsohn showed that combining four such freedoms can raise a nominal 5% false-positive rate to about 61% even when the data contain no real effect.

THE CORE MODEL

What a *p*-value actually says (and the six lies people tell about it)

To not be fooled, you have to know precisely what the instrument reports. Many readers, including trained scientists and a worrying number of the statistics instructors teaching them, get this wrong. So, carefully:

> DEFINITION · THE P-VALUE
>
> The *p-value*[3] is the probability — **computed on the assumption that the *null model*[4], the null hypothesis[5], and every modeling assumption are true** — of getting a result **at least as extreme** as the one you actually observed.

Concrete version: suppose the null model says two groups have the same mean. If a difference this large or larger would occur 3% of the time under that model, then $p = .03$. That is *not* a 3% chance the null is true.

Read it twice. It is a statement about *data, assuming a hypothesis* — written P(data | null). It is emphatically *not* a statement about the hypothesis given the data, P(null | data). Confusing those two is the exact inversion **Day 4** warned about with the disease-test problem: P(positive | sick) is not P(sick | positive), and no amount of staring turns one into the other without a prior. Here are the six misreadings to burn out of your reflexes — drawn from Greenland and colleagues' field guide to *twenty-five* of them:

THREE PROBABILITIES PEOPLE CONFUSE

Alpha[6] is the long-run Type I error rate of a testing procedure if the null model is true: with α = .05, about 5% of true-null tests will falsely cross the line. The *false discovery rate*[7] asks a different question: among the claims you actually declare significant, what fraction are false? That depends on base rates, power, and selection, not just α. A *posterior probability*[8], such as P(null | data), is different again: it needs a prior. A 5% alpha is neither a 5% false discovery rate nor a 5% chance the null is true.

x **"*p* = .03 means there's a 3% chance the null is true."**

No. That's P(null | data) — a Bayesian quantity needing a prior. The *p*-value already *assumes* the null.

x **"*p* = .05 means a 5% chance the result is just chance."**

No. Chance (the null) is the very assumption the calculation is built on; it can't also be the thing in doubt.

x **"*p* > .05 means there is no effect."**

No. Absence of evidence is not evidence of absence — an underpowered study fails to detect real effects all the time.

x **"1 − *p* is the probability the alternative is true."**

No. Neither *p* nor its complement is a probability about any hypothesis.

x **"*p* tells you the result will replicate."**

No. A low *p* from one noisy study says little about the next study.

x **"A significant result is a large or important one."**

No. With a big enough sample, a trivially small effect is "significant." Significance ≠ size ≠ importance.

That last one is the quiet killer, so it gets its own model.

Significance is not size: meet the effect size

A *p*-value entangles effect size, noise, sample size, and model assumptions. Pour in enough participants and even a microscopic, meaningless difference crosses the line. So grown-up statistics insists on reporting the *effect size*[9] separately: *Cohen's d*[10] (a difference in means measured in standard deviations) or a correlation *r*. Jacob Cohen offered rough labels — $d \approx 0.2$ small, 0.5 medium, 0.8 large — while cheerfully admitting they were arbitrary: "Although arbitrary, the proposed conventions will be found to be reasonable by reasonable people." The point is not the labels. The point is that a number with no effect size attached has told you almost nothing worth knowing.

Significance versus size

SCENARIO	P-VALUE	95% CI FOR MEAN DIFFERENCE	COHEN'S D	PRACTICAL READING
Mean difference 0.30, n = 20, noise = 1.00	.34	[-0.32, 0.92]	0.30	Small effect, too imprecise to trust.
Mean difference 0.30, n = 500, noise = 1.00	< .001	[0.18, 0.42]	0.30	Statistically clear, still small in size.
Mean difference 0.80, n = 50, noise = 1.00	< .001	[0.41, 1.19]	0.80	Large by Cohen's convention.
Mean difference 0.30, n = 50, noise = 2.00	.45	[-0.48, 1.08]	0.15	Tiny relative to the noise.

The lesson is not "smaller p-values are better." A p-value can shrink because an effect is large, because the sample is huge, because noise is low, or because all three changed. Report the effect size and interval.

The interval you've been misreading too

Confidence intervals[11] are sold as the humane alternative to *p*-values, and they are better — but they hide their own trap.

> DEFINITION · THE 95% CONFIDENCE INTERVAL

> A 95% CI is produced by a **procedure** that, across many hypothetical repeat experiments, brackets the true value 95% of the time. The "95%" describes the long-run reliability of the *method*, not the odds for your one particular interval.

So the natural sentence — "there's a 95% chance the true value is in *this* interval" — is, in the frequentist world, simply false. Your interval either contains the truth or it doesn't; the dice were thrown when you ran the study. (A **Day 4** Bayesian *credible* interval *does* license that sentence — but only after you've committed to a prior. Different tool, different promise.) Nor does a CI mean 95% of your data live inside it, nor that values just outside are impossible. What it gives you, refreshingly, is a *range of effect sizes compatible with your data* — and that reframing, from "yes/no" to "how much, how precisely," is the heart of what reformers call the *new statistics*.

Why small studies can exaggerate what they detect

Two ways to be wrong: a *Type I error*[12] (false alarm — declaring an effect that isn't there) and a *Type II error*[13] (a miss — failing to see one that is). *Statistical power*[14] is your chance of catching a real effect; more data buys more power. You'd assume an underpowered study just yields shrugs and "no result." Worse: it actively poisons the literature. When power is low, the only estimates lucky enough to clear the significance bar are the wildly overblown ones — the *winner's curse*[15]. Pair that with publication practices that preferentially reward significant findings, and you manufacture a published record of effects that are both exaggerated and fragile. A sobering 2013 audit ("Power failure," Button et al.) put the median power across neuroscience at roughly **21%**; a 2017 reanalysis by Nord and colleagues usefully qualifies the headline, arguing that power is not uniformly low across every neuroscience subfield. The lesson survives the caveat: where studies are underpowered, real effects are missed and the effects that surface are often inflated.

THE DEEPER TRAP

You don't have to cheat to be fooled

Here is the part that should keep honest people up at night. You might read all of the above, swear off p-hacking, run exactly **one** pre-planned analysis, and report it faithfully — and *still* have an inflated false-positive rate. This is Andrew Gelman and Eric Loken's *garden of forking paths*[16].

Their insight is subtle and a little haunting. The damage doesn't require running many analyses. It only requires that the single analysis you ran was *contingent on the data you happened to see*. Suppose your data had come out slightly differently — a touch noisier here, a cluster there. You'd have made a different, equally reasonable choice: controlled for age instead of income, compared women instead of the whole sample, used a median instead of a mean. The other analyses are invisible because you never ran them. But the universe of paths you *would have* taken, in nearby worlds, is exactly the multiplicity that inflates your error rate. As Gelman and Loken put it, you can get a false positive with no "fishing expedition" and no p-hacking at all, even with the hypothesis fixed in advance.

Notice how this rhymes with **Day 5**. There, the lurking decision was *which variables to condition on* — and conditioning on a *collider*[17] could *manufacture* a correlation out of nothing. The forking paths are the same hazard generalized: every defensible analytic choice is a junction, and the data quietly nudge you down the branch that flatters your hypothesis. The honest analyst and the p-hacker can produce the identical false result. The difference is only that one of them knows they did it.

THE FIX THAT NAMES THE DISEASE

If the problem is that choices get made *after* seeing the data, the cleanest cure is to make them *before*: *preregistration*[18] (write down your analysis plan, publicly, in advance) and *Registered Reports*[19] (journals peer-review and accept the *plan*, before any results exist). We met these on **Day 2**; here's *why* they bite. They freeze the confirmatory path; deviations become exploratory rather than hidden, converting a garden of "reasonable choices I made while peeking" into a single committed road.

THE DEBATE

Burn the threshold, or just lower it?

If $p < .05$ is so abusable, what should replace it? Here the field splits — not into crank versus expert, but into camps of serious statisticians who genuinely disagree. It helps to lay them on a single line, from "patch it" to "torch it."

DIAGRAM · THE REFORM SPECTRUM

Four ways to live with *p*

Everyone on this line agrees the status quo is broken. They disagree about how radical the cure must be.

Redefine — Benjamin et al. (2018) keep the threshold idea but move it: call $p < .005$ a "discovery," and .005–.05 merely "suggestive." Simple, blunt, and (critics say) treats a symptom.

Justify your alpha — Lakens et al. (2018) reject any universal number. Pick your threshold deliberately, case by case, weighing the real costs of false alarms versus misses, and *show your reasoning*.

Retire significance — Amrhein, Greenland & McShane (2019), with 800+ co-signers in *Nature*, want to scrap the "significant / not significant" dichotomy itself — the habit of treating $p = .04$ and $p = .06$ as different worlds. Not a ban on *p*-values; a ban on the bright line.

The field's statistical society — the American Statistical Association weighed in twice: a careful 2016 statement of six principles, then a more radical 2019 editorial urging us to stop saying "statistically significant" at all. (That 2019 piece was the editors' view, not formal ASA policy — a distinction that itself caused a row.)

What unites the whole spectrum, from cautious to incendiary, is a single migration: away from a yes/no verdict at a magic number, toward *reporting how big an effect is, how uncertain you are, and how fragile the answer is to the choices you made*. That last idea — fragility — is where the live 2020s frontier lives.

THE FRONTIER · 2024–2026

Making fragility visible — and the hype filter

If a single analysis can mislead, the modern move is brutally simple: *run them all*, and show the spread. Two named methods do this, and a third line of work tests the whole idea by turning loose armies of real scientists on the same data. As always, each claim gets a label.

Edge 01 [ESTABLISHED] [CONTESTED]

Multiverse & specification-curve analysis

Instead of defending one analysis, you enumerate *every* reasonable combination of choices — keep outliers or trim them, log-transform or not, control for this covariate or that — and compute the result under each. *Multiverse analysis*[20] (Steegen, Tuerlinckx, Gelman & Vanpaemel, *Perspectives on Psychological Science*, 2016) displays the whole cloud of outcomes. *Specification-curve analysis*[21] (Simonsohn, Simmons & Nelson, *Nature Human Behaviour*, 2020) sorts hundreds of specifications into a single curve and asks: across all the defensible ways to slice this, does the effect hold up, or does it evaporate the moment you wiggle a choice? It's honest, vivid, and increasingly expected by reviewers. The interactive below lets you feel it.

One discipline matters more than the picture suggests: some "reasonable specifications" change the *estimand*[22]. Controlling for W is not just another estimate of the same question unless W is a causally justified control; rank-transforming, trimming outliers, or dropping a subgroup can shift the target from a mean difference to an ordinal association, a robust effect, or a population-specific effect. Specification curves expose that fork, but they do not decide which question you meant to ask.

The **hype filter** matters here, though. These are superb *transparency* devices and weak *inferential* ones. The authors of specification-curve analysis concede the

catch themselves: deciding which specifications are "reasonable" is a judgment call no algorithm can make, and "the goal to eliminate subjectivity is unattainable." A determined arguer can still curate the multiverse. And there is no settled way to compute a single valid conclusion from a curve — which is why 2024 brought new machinery (PIMA, below) trying to supply one. [ESTABLISHED] [REVIEW]

Edge 02 [ESTABLISHED]

Many analysts, one dataset — the most damning evidence in the room

Here is the experiment that should change how you read every headline. Take one dataset, one clear question, and hand identical copies to dozens of expert teams. Do they converge on the same verdict? Often, no.

In Silberzahn et al. (*Advances in Methods and Practices in Psychological Science*, 2018), **29 teams** (61 analysts) were asked a single question — do soccer referees give more red cards to dark-skin-toned players? Their estimated odds ratios ranged from **0.89 to 2.93**; twenty teams found a significant positive effect, nine did not. Tellingly, the analysts' prior beliefs and even their statistical expertise did *not* explain who found what. In Botvinik-Nezer et al. (*Nature*, 2020), **70 teams** (180 researchers) analyzed the same brain-imaging dataset against nine pre-set hypotheses — and no two teams used the same pipeline; their yes/no conclusions diverged sharply even when their underlying statistical maps were highly correlated. Breznau et al. (*PNAS*, 2022) put the same data and hypothesis (does immigration erode support for social policy?) to **73 teams** and watched estimates scatter from clearly negative to clearly positive. Even finance has its version: Menkveld et al.'s "Nonstandard Errors" (*Journal of Finance*, 2024) had **164 teams** test the same hypotheses on the same market data, and found the cross-team variability rivaled ordinary statistical error — shrinking, encouragingly, when extra peer-review stages were added.

This is **Day 1**'s Gettier problem made flesh, inside a single dataset. Each team's belief is "justified" by a competent analysis; whether it's *true* — connected to reality rather than to their particular path through the garden — varies team to team. One analysis is one branch. Treat it accordingly.

Edge 03 [PROMISING] [ESTABLISHED]

Plugging the inference gap — and Day 4's quiet alternative

The open problem in all of the above is honest *inference* across a multiverse: how do you draw one valid conclusion from a thousand entangled analyses? A 2024 entry, *PIMA* (Post-selection Inference in Multiverse Analysis; Girardi et al., *Psychometrika*, 2024), offers a sign-flipping test that aims to give the multiverse a proper joint error guarantee, reaching beyond the linear-model limits of the 2020 specification curve. It's peer-reviewed and genuinely interesting, but new and not yet standard practice — a **promising hint**, not a settled tool.

Meanwhile the boring, durable fixes keep spreading: the Registered Reports format — where journals accept your *plan* before your results exist — is now offered at **several hundred journals**, and preregistration is becoming a default rather than a virtue. And recall **Day 4**'s *e-values*[23] and "testing by betting" (Grünwald, Ramdas, Shafer): an alternative to the *p*-value that stays valid even if you peek at your data and stop whenever you like — aimed squarely at the optional-stopping degree of freedom you toggled in the factory above. Still a research frontier rather than mainstream practice — **promising**, watch this space.

Specification Curve Summary

CHOICE	EFFECT ON THE ANALYSIS	INTERPRETIVE RISK
Trim outliers	Exclude extreme values before estimating the association.	May stabilize a result or selectively remove inconvenient points.
Control for W	Adjust for a lurking background variable that drives both X and Y.	Often shrinks a spurious association, but the control must be justified causally.
Rank-transform	Replace raw values with ranks before computing the association.	Can reduce sensitivity to distribution shape while changing the estimand.
Drop subgroup	Analyse only a subset of the data.	Can test a real boundary condition or create a flattering subset.

The honest report is the distribution across reasonable specifications; the dishonest report is the prettiest point on the curve.

OPEN QUESTIONS

What's genuinely unsettled

- **Should "statistical significance" survive at all?** Reformers are split between disciplining the threshold and abolishing it. Decisions still have to be made somewhere — a drug is approved or it isn't — and abolitionists owe an account of how.
- **Can a multiverse ever yield a single honest verdict?** Or is the choice of "reasonable specifications" an irreducibly subjective act that just relocates the

forking paths one level up? PIMA and friends are early attempts; the jury is out.

- **Is variation in *conclusions* the same as variation in *reality*?** A subtle 2023 reanalysis noted that in some many-analyst studies the headline-grabbing disagreement was about *significance*, while the underlying effect sizes were quietly consistent and small. Disagreement about a verdict can exceed disagreement about the number. Don't over-learn "everything is hopeless."
- **Do *p*-values or e-values deserve the future?** The betting-based tools elegantly solve optional stopping but can demand more data and import a new modeling burden (which bet?). Coexistence looks likelier than conquest.
- **And the question waiting in the AI block:** when a model trained on millions of papers reports a "robust" result, has it learned to reason about evidence — or to imitate the very forking-path habits that got us here? (**Days 138–145**.)

◆ THE DAY IN THREE SENTENCES

BIG IDEA

Statistical tools don't protect you from fooling yourself — used with ordinary human flexibility they help you do it, because the dozens of defensible choices in any analysis (the "garden of forking paths") let noise masquerade as discovery, so the cure is transparency, preregistration, effect sizes with uncertainty, and robustness across reasonable analyses rather than a verdict at a magic threshold.

BEST ANALOGY

The False-Positive Factory: hand yourself enough innocent freedoms and a coin-flip's worth of nothing becomes a 61% chance of "success" — and the garden of forking paths, where the analyses you *would have* run in nearby worlds inflate your error even if you only ran one.

LIVE CONTROVERSY

Whether to lower the threshold ($p < .005$), justify it case by case, or retire "statistical significance" altogether — with multiverse and many-analyst studies revealing how alarmingly a single dataset's verdict depends on who analyzes it.

THREADS TODAY › information (separating signal from noise; the *p*-value as a misread evidence measure) · computation (the lab as a fallible inference engine; the multiverse as brute

enumeration) · emergence (science as an error-correcting system larger than any analyst — preregistration, many-analyst crowds, meta-analysis).

TOMORROW → DAY 07

Information Theory

Today was about not mistaking noise for signal. Tomorrow we learn to *measure* signal exactly. Claude Shannon's bit, entropy as surprise, the game of twenty questions, channel capacity — and the startling bridge to physics in Landauer's principle, the irreducible energy cost of erasing a single bit. The *information* thread we've been quietly pulling since Day 1 finally gets its own mathematics.

NOTES

1. P-hacking means trying enough analyses, exclusions, outcomes, or stopping rules that random noise can be made to look statistically significant.
2. Researcher degrees of freedom are the legitimate-looking analytic choices a researcher can make about data collection, cleaning, modeling, and reporting.
3. A p-value is a tail probability for data under a specified null model, not the probability that the null hypothesis is true.
4. A null model is the world assumed by the p-value calculation, usually a model with no effect plus all stated sampling and modeling assumptions.
5. A null hypothesis is the baseline claim being tested, often that there is no effect or no difference.
6. Alpha is the false-alarm rate you commit to before looking at the data, conditional on a true null model and a specified testing procedure.
7. The false discovery rate is the fraction of declared discoveries that are false.
8. A posterior probability is the probability of a hypothesis after combining data with prior information.
9. An effect size reports how large a difference or association is, instead of only whether it crossed a significance threshold.
10. Cohen's d expresses a difference between two means in units of standard deviation.
11. A confidence interval is a range produced by a method with a stated long-run coverage rate, such as 95%.
12. A Type I error is a false positive: rejecting a true null hypothesis.

13. A Type II error is a false negative: missing a real effect.
14. Statistical power is the probability that a study will detect an effect of a given size if that effect is real.
15. The winner's curse is the tendency for selected significant estimates from noisy studies to exaggerate the true effect.
16. The garden of forking paths is the set of analyses a researcher might reasonably have chosen after seeing different patterns in the same data.
17. A collider is a variable influenced by two other variables; conditioning on it can create a misleading association between its causes.
18. Preregistration means recording hypotheses, data handling, and analysis plans before looking at the results.
19. Registered Reports are journal articles accepted in principle after peer review of the research question and method, before the results are known.
20. Multiverse analysis runs many defensible analysis choices and reports the distribution of results rather than one preferred path.
21. Specification-curve analysis sorts many defensible model specifications by their estimated effect to reveal robustness or fragility.
22. An estimand is the exact quantity a study is trying to estimate, such as a raw association, a causal effect, or an effect for a specific population.
23. E-values measure evidence as the payoff of a betting strategy against a null hypothesis.

SOURCES

Sources & further reading

1. Feynman, R. P. (1974). "Cargo Cult Science." *Engineering and Science* 37(7): 10–13. Caltech's 1974 commencement address; source of the opening quote. calteches.library.caltech.edu

2. Open Science Collaboration. (2015). "Estimating the reproducibility of psychological science." *Science* 349(6251): aac4716. doi:10.1126/science.aac4716; source for the 97% original significant / 36% significant replication statistic. doi.org/10.1126/science.aac4716

3. 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. doi:10.1177/0956797611417632. doi.org/10.1177/0956797611417632 – the "When I'm Sixty-Four" demonstration; the 5%→61% simulation.

4. Gelman, A. & Loken, E. (2014). "The Statistical Crisis in Science." *American Scientist* 102(6): 460–465. – the "garden of forking paths"; earlier as a 2013 Columbia working paper. americanscientist.org

5. 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. doi:10.1007/s10654-016-0149-3. doi.org/10.1007/s10654-016-0149-3 — the canonical list of 25 misreadings.

6. Haller, H. & Krauss, S. (2002). "Misinterpretations of Significance: A Problem Students Share with Their Teachers?" *Methods of Psychological Research Online* 7(1): 1–20. — survey evidence that p-value misunderstandings persist even among instructors teaching statistics. epub.uni-regensburg.de/34338

7. Badenes-Ribera, L., Frias-Navarro, D., Iotti, B., Bonilla-Campos, A. & Longobardi, C. (2016). "Misconceptions of the p-value among Chilean and Italian Academic Psychologists." *Frontiers in Psychology* 7: 1247. doi:10.3389/fpsyg.2016.01247. doi.org/10.3389/fpsyg.2016.01247

8. Wasserstein, R. L. & Lazar, N. A. (2016). "The ASA Statement on p-Values: Context, Process, and Purpose." *The American Statistician* 70(2): 129–133. doi:10.1080/00031305.2016.1154108. doi.org/10.1080/00031305.2016.1154108 — the six principles.

9. 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. doi.org/10.1080/00031305.2019.1583913 — editorial to a 43-article special issue; "stop saying statistically significant" (editors' view, not ASA policy).

10. Benjamin, D. J., Berger, J. O., Johannesson, M., et al. (2018). "Redefine statistical significance." *Nature Human Behaviour* 2(1): 6–10. doi:10.1038/s41562-017-0189-z. doi.org/10.1038/s41562-017-0189-z

11. Lakens, D., Adolfi, F. G., Albers, C. J., et al. (2018). "Justify your alpha." *Nature Human Behaviour* 2(3): 168–171. doi:10.1038/s41562-018-0311-x. doi.org/10.1038/s41562-018-0311-x

12. Amrhein, V., Greenland, S. & McShane, B. (2019). "Scientists rise up against statistical significance." *Nature* 567(7748): 305–307. doi:10.1038/d41586-019-00857-9; 800+ signatories. doi.org/10.1038/d41586-019-00857-9

13. Button, K. S., Ioannidis, J. P. A., Mokrysz, C., et al. (2013). "Power failure: why small sample size undermines the reliability of neuroscience." *Nature Reviews Neuroscience* 14(5): 365–376. doi:10.1038/nrn3475; median power ≈ 21%. doi.org/10.1038/nrn3475

14. Nord, C. L., Valton, V., Wood, J. & Roiser, J. P. (2017). "Power-up: A Reanalysis of 'Power Failure' in Neuroscience Using Mixture Modeling." *The Journal of Neuroscience* 37(34):

8051–8061. doi:10.1523/JNEUROSCI.3592-16.2017; caveat on the uniformity of low power across neuroscience. doi.org/10.1523/JNEUROSCI.3592-16.2017

15. Center for Open Science. "Registered Reports." Official COS overview and participating-journal list; source for "over 300 journals." cos.io/initiatives/registered-reports

16. Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. (2016). "Increasing Transparency Through a Multiverse Analysis." *Perspectives on Psychological Science* 11(5): 702–712. doi:10.1177/1745691616658637. doi.org/10.1177/1745691616658637

17. Simonsohn, U., Simmons, J. P. & Nelson, L. D. (2020). "Specification curve analysis." *Nature Human Behaviour* 4(11): 1208–1214. doi:10.1038/s41562-020-0912-z. doi.org/10.1038/s41562-020-0912-z

18. Silberzahn, R., Uhlmann, E. L., Martin, D. P., et al. (2018). "Many Analysts, One Data Set." *Advances in Methods and Practices in Psychological Science* 1(3): 337–356. doi:10.1177/2515245917747646; 29 teams, ORs 0.89–2.93. doi.org/10.1177/2515245917747646

19. Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., et al. (2020). "Variability in the analysis of a single neuroimaging dataset by many teams." *Nature* 582(7810): 84–88. doi:10.1038/s41586-020-2314-9; 70 teams. doi.org/10.1038/s41586-020-2314-9

20. Breznau, N., Rinke, E. M., Wuttke, A., et al. (2022). "Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty." *PNAS* 119(44): e2203150119. doi:10.1073/pnas.2203150119; 73 teams. doi.org/10.1073/pnas.2203150119

21. Mathur, M. B., Covington, C. & VanderWeele, T. J. (2023). "Variation across analysts in statistical significance, yet consistently small effect sizes." *PNAS* 120(3): e2218957120. doi:10.1073/pnas.2218957120; caveat on reading many-analyst disagreement through effect sizes and intervals, not significance labels alone. doi.org/10.1073/pnas.2218957120

22. Menkveld, A. J., Dreber, A., Holzmeister, F., et al. (2024). "Nonstandard Errors." *The Journal of Finance* 79(3): 2339–2390. doi:10.1111/jofi.13337; 164 teams. doi.org/10.1111/jofi.13337

23. Girardi, P., Vesely, A., Lakens, D., et al. (2024). "Post-selection Inference in Multiverse Analysis (PIMA): An Inferential Framework Based on the Sign Flipping Score Test." *Psychometrika* 89(2): 542–568. doi:10.1007/s11336-024-09973-6. doi.org/10.1007/s11336-024-09973-6

24. Ramdas, A., Grünwald, P., Vovk, V. & Shafer, G. (2023). "Game-theoretic statistics and safe anytime-valid inference." *Statistical Science* 38(4): 576–601. doi:10.1214/23-STS894; e-processes, testing by betting, and optional-stopping validity. doi.org/10.1214/23-STS894

25. Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum. — effect-size conventions, offered as deliberately arbitrary.

END OF DAY 06 · 174 DESCENTS REMAIN