The Statistical Crisis in Science

Stijn Debrouwere (author), Prof. Dr. Els Goetghebeur (advisor)

Preface

What can we say about the statistical crisis in science that has not already been said? Eminent scientists like Paul Meehl, Jakob Cohen and John Ioannidis have already done a great job pinpointing the problems with P-values, null hypothesis significance testing, nonreplicable research, sloppy statistics and fishing expeditions. At the other end of the spectrum journalists at The Atlantic, The New Yorker, FiveThirtyEight and The New York Times have done a surprisingly good job of translating these issues to a lay audience. What is left?

What is left is the space in between: for many working scientists the arguments in the statistical literature are too abstract and I have a sneaking suspicion that nobody has ever bothered to double-check the somewhat convoluted calculations in John Ioannidis' otherwise wonderful *Why Most Published Research Findings Are False*. On the other hand lay accounts do not treat the problem with enough fidelity to satisfy those who might not be statistical experts as such but do regularly have to interpret t-tests and regression analyses. Yet ultimately it is these students and doctors and working scientists who are the key audience for publications about the statistical crisis in science, because they decide what tomorrow's statistical practices will look like.

I have used this dissertation as an exercise in statistical writing for such an audience of not-quite-experts, not-quite-laymen. To the dissertation committee reading this dissertation I therefore ask that you please do not get frustrated when it takes five paragraphs to explain what *power* or *sensitivity* or *positive predictive value* means. Instead I hope you appreciate the effort that went into finding ways to explain these concepts without dumbing them down, which to me was one of the core challenges in writing this work; perhaps readers can find inspiration for their own teaching among the explanations, anecdotes and analogies interspersed throughout this work.

The dissertation is split into two equal parts. The first part is a mostly theoretical look at the shortcomings of P-values and null hypothesis significance testing. The second part is a broader overview of how bad statistics and questionable research practices have led to a crisis of unreplicable research. These two parts are fully independent of each other and can be read in either order. Occasionally the material overlaps, but repetition has been kept to a minimum.

You will read in this work a lot of statements along the line that *some* or *many* or *a lot of* scientists do something or other. I appreciate quantification as much as any statistician so I did feel queasy when writing these kinds of vague statements. At the same time, I don't feel that when talking about statistical practices across a wide range of scientific disciplines, it necessarily makes sense to specify, say, that 12.8% of scientists confuse P-values for effect

sizes and that 23.1% of scientists are not aware of how Type II error might affect their studies. References are made to research that estimates the prevalence of various practices whenever appropriate.

Acknowledgements

I still remember the first meeting I had with my dissertation advisor Els Goetghebeur. With a concerned look in her eyes, she said something along the lines of "well, this all sounds very interesting... but this is a statistics program, you're not going to turn your dissertation into a philosophical treatise, are you?". I thank my promotor for signing off on a dissertation that is one third philosophy, one third sociology and one third (but a crucial one third!) statistics.

This dissertation relies on a great many case studies and examples of bad statistics and dubious science. In addition to a lot of my own research, I found many great anecdotes through the blog and academic work of Andrew Gelman (Columbia University), papers and columns by Howard Wainer (University of Pennsylvania) and Retraction Watch which is run by Adam Marcus and Ivan Oransky. My thanks to them.

The Statistical Crisis in Science is the overarching title of this dissertation; I first encountered this way of describing the replication crisis in an article of the same title by Andrew Gelman and Eric Loken in the American Scientist. *Why Most Published Research Findings Are False* is a 2005 paper by John Ioannidis. As its underlying model of publication bias forms the core of the second chapter of this dissertation, I've kept the title and have loosely structured the chapter as an elaboration on Ioannidis' compelling but perfunctory explanations. My thanks to these authors and my apologies for the lack of creativity.

All graphs and simulations in this thesis have been coded from scratch unless otherwise noted. Some of the graphs do mimic work by other authors. I first saw P-value prior/posterior plots in a course by Stijn Vansteelandt (Ghent University). The effect exaggeration plots are based on work by Andrew Gelman. A plot showing the compound effect of low power and publication bias is based on a visualization by Tal Yarkoni (University of Texas). Any errors are mine.

The Cult of P

Stijn Debrouwere (author), Prof. Dr. Els Goetghebeur (advisor)

Contents

P-values

When you hear the word *statistics*, are you reminded of bar charts and medians and standard deviations? I imagine most people are, but that's actually not the kind of statistics college students have such a hard time with and not the kind of statistics that scientists obsess about.

Statistics finds its roots in the collection of lots and lots of measurements, like the measurements of human weight and height that led Adolphe Quetelet to the body mass index. But *your* dataset of heights and weights might not look anything like *my* dataset and in our separate analyses we might even be driven to opposite conclusions. So we need a means to judge how much trust we can put in our findings and how likely it is that we will continue to find a similar pattern when we collect more data. Statistics thus slowly revised its mission from the measurement of people, nations and economies to the measurement of uncertainty and confidence.

Many years after he had figured out the physical equations that govern ocean tides, Pierre-Simon Laplace wondered whether the moon might affect the atmosphere the same way it affects the oceans. To find out, in 1827 Laplace cross-referenced atmospheric pressure readings recorded at the Paris Observatory with the moon's orbit and found that the moon's phases did appear to have a modest effect on barometric readings. But how convincing was this finding? As a way to double-check his findings, Laplace calculated the probability of seeing data such as the records he had at his disposal or hypothetical records that would seem to show an even greater lunar influence *presuming that* the moon in fact had no influence on atmospheric pressure at all. The answer, it turned out, was that two times out of three random fluctuations in atmospheric pressure would be enough to produce an illusory effect like the one Laplace had found, and the theory of atmospheric tides was abandoned (Stigler 2016, ch. 3).

Throughout the 18th and 19th century enterprising scientists would make calculations such as these, but they become ubiquitous only after Ronald Fisher's *Statistical Methods for Research Workers* shows researchers how to calculate them for different kinds of experiments and gives the P-value a solid theoretical underpinning, using the eponymous Fisher information to determine the standard error of any maximum likelihood estimate.

Fast forward almost two hundred years from Laplace's calculations, and across scientific disciplines, the strength of statistical conclusions is communicated exclusively using P-values, which answer the question *"What if it were due to chance?"* It's a useful question to ask, because patterns found in datasets often appear much more convincing than they really are, especially with few observations, noisy measurements or when there's a lot of diversity among individuals, and we must be careful not to read too much into what might turn out to be spurious correlations and imaginary causes and effects. A scientific theory is accepted only if the empirical observations in its support would be too unusual or surprising to chalk up to coincidence or happenstance. Only then do we consider a scientific finding to be *statistically significant*.

The P-value hides its imperfections well, but despite its great appeal and its almost universal use in modern science, the P-value and the hypothesis testing routines that go with it are

not as harmless as they seem. The P-value is used to make ineffective drugs look like wonder treatments, to lend credence to fringe science like extrasensory perception and to confirm every researcher's pet theory. Hypothesis testing is supposed to give us confidence that our scientific findings will stand the test of time, but when scientists repeat their colleagues' experiments more often than not they find that whatever phenomenon was supposed to be there has vanished.

How can such a harmless idea have such a devastating impact on science? Why do scientists and scientific journals display an almost cult-like adherence to hypothesis testing when its flaws have been known for almost half a century and when even Ronald Fisher, the original promoter of P-values and null hypothesis testing, chastised his fellow statisticians for bastardizing his ideas? Who is to blame: the P-value or the way we use, abuse, interpret and misinterpret it?

P-values only protect against random error

Among the many definitions of a P-value that you might hear from researchers is this one: *a p-value is the probability that I'm wrong.* This definition confuses likelihood with probability, mistakes corroboration for verification and ignores non-stochastic forms of uncertainty. We will talk about each of these misconceptions in turn, but for now let us concentrate on the difference between stochastic and empirical uncertainty.

Stochastic uncertainty is the uncertainty introduced by variable phenomena. If you flip a coin and it ends up tails five times in a row, that's suspicious but not quite suspicious enough to say with any confidence that the coin is biased. P-values give us something to hold onto when dealing with randomness, but it is silent about all other forms of variation and uncertainty:

- Will the finding generalize to other populations? A different way of teaching might work for fourth graders, but will it work for third graders?
- Did we identify cause and effect? Countries where olive oil is a staple have fewer cardiovascular problems, but is it really about the oil?
- Does the statistical analysis fit the research question? Is our model appropriate to the data at hand? If we study the wrong outcome variable, any conclusion we make is void (Rotello, Heit, and Dubé 2014). If our model is defective, then any P-value conditional on that model is meaningless.
- Was our sample representative? An internet poll on a left-leaning news website will not be representative of what all voters think.
- Did any mistakes slip in during data collection?
- Were statistical calculations performed without error? It's easy to get the degrees of freedom wrong or to make a mistake when importing data.
- Did we measure what we thought we did? Is measuring IQ really the same thing as measuring intelligence?

Statistics is not altogether helpless in answering these questions. Regression can get rid of many confounding factors, and randomized experiments avoid them altogether. Mediation

analysis can determine whether A causes B by means of our proposed mechanism or some other way. But most of the factors that determine whether a research finding can be trusted is outside of the statistician's control.

It is the job of the peer reviewer to check whether the research makes sense, regardless of P-values. Reviewers do the best they can, but sometimes things do fall through the cracks, from faulty calculations (Bakker and Wicherts 2011) to using the wrong statistical test (Scales et al. 2005) to not accounting for important confounders.

For example, dozens of studies have been published in prestigious journals claiming that moderate alcohol consumption promotes heart health, especially red wine. But in reviewing the literature, Kaye Fillmore and her colleagues found very few studies that properly took into account that among those who abstain from alcohol is a large group of former alcoholics and patients on medication, leading to unfair comparisons (Fillmore et al. 2007).

Because the P-value is the central piece of evidence in scientific publications, researchers can be led to believe that P-value calculations inoculate them against all sorts of methodological issues that in fact it is powerless against. As Judea Pearl once pointed out, "The opacity of probability theory avoids argument, the clarity of causal statements invites it" (Pearl 2003, 288) and unfortunately much of statistics diverts attention from measurements, methods and causal connections that ultimately decide whether the statistics we calculate make sense. Andrew Gelman and Erik Loken have called this "laundering uncertainty" (A Gelman and Loken 2014): making the uncertain look much more certain than it really is through statistical wizardry.

The false sense of security that P-values provide can thereby lead to moral licensing – the idea that a good deed entitles you to a bad one – where the researchers assume that good statistics will compensate bad experimental setups and far-fetched interpretations. As statistics gets smarter, does science become dumber?

P-values are only as good as the hypotheses that prompt them

Predictions solve an important problem for scientific reasoning: any fool can come up with a theory that fits the facts. There's an infinite amount of theories that can explain the same set of facts, and an infinite amount of ad-hoc adjustments you can make to that theory whenever new data comes in that doesn't support it – we've all seen pundits on television who magically appear to be able to explain everything and even when they make the wrong predictions, they'll explain *that* too. Making a good guess *beforehand* is much harder than explaining something *afterwards*, so this is what we require to put trust in scientific findings.

Sometimes predictions link up quite naturally with the underlying theory, but more often a scientific theory requires a whole host of auxiliary assumptions to be turned into a testable hypothesis.

Let's say we wish to study the effect of solving Sudoku puzzles on intelligence. Administering repeated IQ tests would be very cumbersome, so instead we might use working memory as

a proxy variable, as working memory is known to correlate with intelligence. To measure working memory, we will ask people to remember a series of random numbers, and see whether a Sudoku-solving group can remember more of them than the non-Sudoku solving group. Because we can't track people until they die, we will let the experiment run for a couple of weeks, and then assume that any effect we find will persist or even grow larger over time, as long as the subject sticks to it.

If we've found a statistically significant difference in the average working memory capacity between the two groups, we will conclude our discussion section by saying that Sudoku seems to be a great way to keep your brain young. But every step of the way we had to make assumptions and simplifications, and the result is that the P-value supports the theory that prompted it only in a very roundabout way, as one small part of a long chain of probabilities.

An alarming trend in the social and behavioral sciences is that increasingly the bulk of scientific argument is shifted towards the discussion section. After a perfunctory statistical analysis which shows a relationship between A and B that is not likely to be due to chance, the discussion section is used to wax philosophical about exactly what is going on and what implications it has for other research and society. In an insiduous bait and switch, strong evidence of a correlation between A and B is co-opted to make whatever claims we wish, often only *after* seeing the data, as long as they vaguely align with that correlation.

A P-value is only as good as the translation between research hypothesis and statistical hypothesis.

P-values confound sample size and effect size

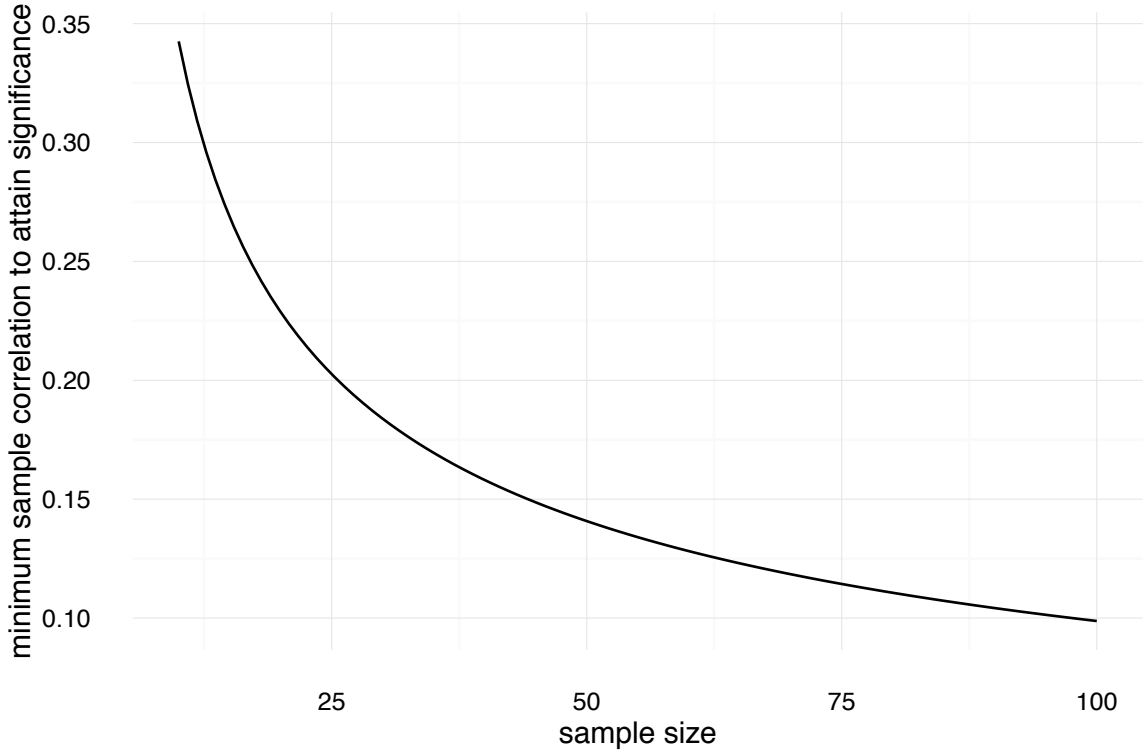
The p-value exists to distinguish between happenstance and those events that did not arise due to chance.

What factors allow us to make this distinction?

- Big effects are less likely to be the result of natural variation and random fluctuations. The introduction of vaccines for measles, polio and smallpox reduced infections from as high as 1 in 100 to near zero in western countries; no statistical confirmation is needed beyond the *intraocular trauma test*: the result hits you right between the eyes (Panhuis et al. 2013 as visualized in DeBold and Friedman (2015) for the United States).
- Low variation under the alternative hypothesis. It is not hard to prove that parachutes have a statistically significant effect on the mortality rates of people jumping out of airplanes: rarely does anyone survive a jump without a parachute (Smith and Pell 2003).
- Reliable measurement. It can be hard to detect supernovae thousands of light years away, and just as hard to learn anything about diets that work when they rely on unreliable self-reported food questionnaires. But just two precise observations of a solar eclipse were sufficient to convince physicists of general relativity.
- A lot of data. It is possible to glean information from even the weakest signals. For example, it is possible to statistically recover passwords and other secret messages from

a sufficiently long recording of keystrokes, because the minute differences between how fast the user hits subsequent keys and minute differences in the sound between keys depend on where on the keyboard these keys are located and which fingers we use to hit them.

A low P-value can mean the effect is big, the phenomenon is not very variable, the sample size is big or the measurements were precise. A significant effect with $p = .01$ can be tiny but still readily distinguishable from chance due to highly accurate measurement apparatus. But that same $p = .01$ effect could be an earth-shattering new finding that even noisy measurement cannot hide.



It is very rare to see research where the only thing we care about is whether or not something exists, regardless of effect size. Imagine if Isaac Newton had concluded the *Philosophiae Naturalis Principia Mathematica* by writing that force and mass have a statistically significant effect on acceleration, and then retired from physics.

Point estimates, confidence intervals, predicted probabilities, common language effect sizes – statistics has many good indicators of effect size at its disposal. Still researchers are tempted to treat the P-value as an indication of effect size, which it was never meant to be. Even when studies mention effect sizes and confidence intervals, they don't always use them to judge the importance of a finding. Summarizing her review of statistical practices in medical journals that do in fact require effect size estimates, Fiona Fidler writes that "compliance was superficial" [Fidler2004, p. 119]: authors made the necessary calculations but did not use them to quantify their conclusions, instead falling back on P-values.

Some scientists have grown so frustrated with this lack of attention paid to effect sizes that they have even published guides such as *Computing contrasts, effect sizes, and counternulls on other people's published data*, *How to calculate effect sizes from published research* and the 254-page book *Statistics with Confidence: Confidence Intervals and Statistical Guidelines*. Statistical tools have been built specifically for statistical analysis using confidence intervals.

Much of this can be explained by the simple fact that in most disciplines and most scientific journals, interpretation of effect size is considered to be optional, whereas P-values are not. Some researchers may consider statistics nothing but a bunch of opaque calculations of dubious value, so they simply calculate and summarize the experiment in whatever way they think will get them published. Finally, neither SAS nor R provides confidence intervals for built-in analyses unless explicitly asked for.

Yet sometimes lack of effect sizes is an indication of malice or at the very least wishful thinking: researchers will sometimes be aware of the small effect size, but try to jazz up their findings by just talking of *significant* or *very significant* effects, de-emphasizing the actual quantitative results.

For example, smaller class sizes might contribute *significantly* to student test scores, but to evaluate the importance of this finding a researcher must

- put a number on how much test scores have increased,
- compare the increase to alternative interventions,
- calculate the cost of the teachers needed to teach these smaller classes.

Without this context, "class size significantly affects test scores" is a meaningless statement, communicating only that the effect of smaller class sizes on test scores is not exactly zero point zero.

Every scientist would like to say that they've found something big, something real, something *significant* and so researchers can get a little carried away when describing the results of their statistical tests.

Others realize that sample size factors into it, and will argue that the statistically significant P-value they found in their small experiment is *extra special*: if you see a light through thick fog, it's probably a pretty powerful beam. But P-values can behave quite erratically at low sample sizes. A single observation can cause a large drop or spike in P-value. This is most evident in discrete distributions like the binomial. Assuming a fair coin flip, throwing 8 out of 10 heads results in $p = 0.11$ but throwing 9 out of 10 heads results in $p = 0.02$. Throw another die which lands on tails and the P-value bumps back up to $p = 0.07$. But continuous distributions have problems of their own: P-values that rely on the normal distribution (Wald tests) or even the t-distribution only approximate the true null distribution at low sample sizes.

Because P-values are influenced by so many different factors, attempting to attach any interpretation to them beyond a simple *chance alone can explain what we found* quickly turns into a guessing game.

P-values are not posterior probabilities

Unfortunately, there exists no straightforward way to calculate the probability that a scientific theory is correct using only data. But we can get close with some help from Bayes' theorem, which tells us that the posterior probability that a hypothesis based on a theory is true can be updated according to:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

The denominator $P(D)$ can be ignored because regardless what your hypothesis might be, it stays the same. We can also get rid of $P(H)$ because it asks us to rate the hypothesis as intrinsically more or less probable, before having seen any of the data... but we've gone to great lengths to collect all this data precisely because we wish to know whether our theory is or isn't probable!

What remains of the original equation is $P(H_A|D) \propto P(D|H_A) = 1 - P(D|H_0)$, and the right hand side of this equation is our good friend, the P-value.

To make sure we don't confuse the probability of the data with the probability of the hypothesis, statisticians prefer to call $P(D|H)$ the *likelihood* of the hypothesis, whereas $P(H|D)$ is the *probability* of the hypothesis.

Mixing up likelihoods and probabilities is known as *confusion of the inverse*, and it's a serious logical error.

Imagine winning the lottery. The probability of winning solely due to the luck of the draw is astronomically small. If, however, you had somehow found a way to cheat, your probability of winning would actually be quite high. With these likelihoods in mind, we can conclude that most lottery winners are cheats. After all, $P(\text{winning}|\text{chance}) < P(\text{winning}|\text{cheating})$. Clearly, reasoning from likelihoods isn't always so wonderful and we were a bit overeager in throwing out $P(H)$ from our equation: ignoring the prior probability of a hypothesis can mislead.

Despite the fact that statistics professors around the world warn students of confusion of the inverse, it turns out to be a very hard difference to grasp, and even those statistics professors themselves sometimes mistake the P-value for a posterior probability (Castro Sotos et al. 2007).

According to Bayes theorem, the posterior probability of a hypothesis is determined not just by the scientific evidence provided by an experiment or by observational data, but also by all other subject-matter considerations that may make a result more or less probable.

Occasionally we *do* have a good sense of whether a hypothesis is likely to pan out. For example, when Psychological Science published research showing evidence of extrasensory perception in 2011, they should have probably taken into account that there's no way to account for ESP unless everything we know about human biology and the physical universe is wrong (Schimmack 2012). We cannot even begin to imagine the physical means by which

thoughts would float through the air and the biological means through which our brains would send and read these signals, so this kind of fringe science is best ignored.

Other fields of research are still so new that we don't know if there are likely to be many effects to be found, or none at all, as in epigenetics which studies whether stresses and experiences of a parent can carry through in the genes of children (Heijmans and Mill 2012). Because the validity of the field of research is still in doubt – the field as a whole, not individual studies – P-values can be optimistic and should be adjusted upwards.

Finally, keep in mind that when we write $P(D|H)$ what we actually mean is $P(D|H, M)$: a P-value can only be calculated conditional on a model, and when that model is inappropriate, so are the P-values. In rare cases, faulty models can even lead to very small P-values and unrealistically narrow confidence intervals, in particular when there is only a very narrow range of parameter values that are even the least bit plausible.

But mostly the problem is simply one of interpretation: a P-value of 0.03 is not a 3% probability that your theory is wrong and evidence in favor of the investigator's alternative hypothesis might still be quite weak, depending on the ratio $P(H_A)/P(H_0)$.

Because a P-value is determined by both sample size and effect size, research on small samples often won't find any relationships or differences, even if they are really there.

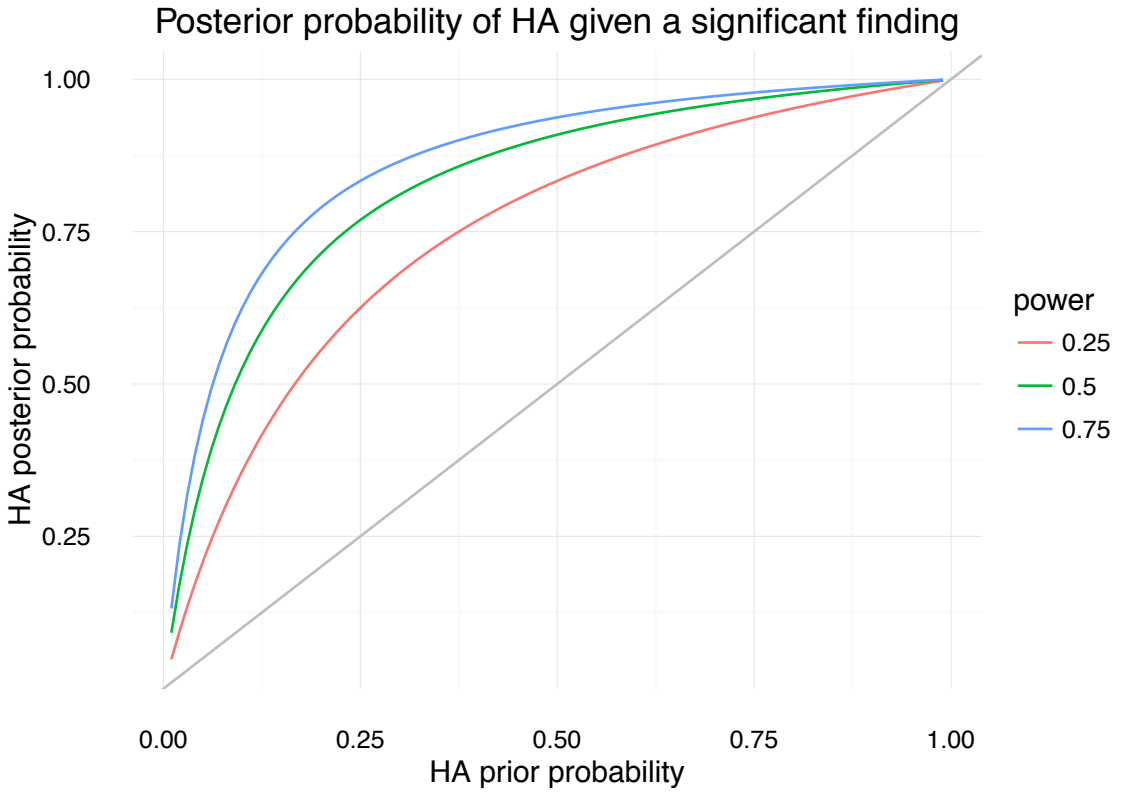
The simplest way to show this is to calculate the confidence interval around a point estimate. Let's say a difference between groups has a 95% confidence interval from a 0% increase all the way up to a 100% increase. Then we must admit that it's possible that the intervention

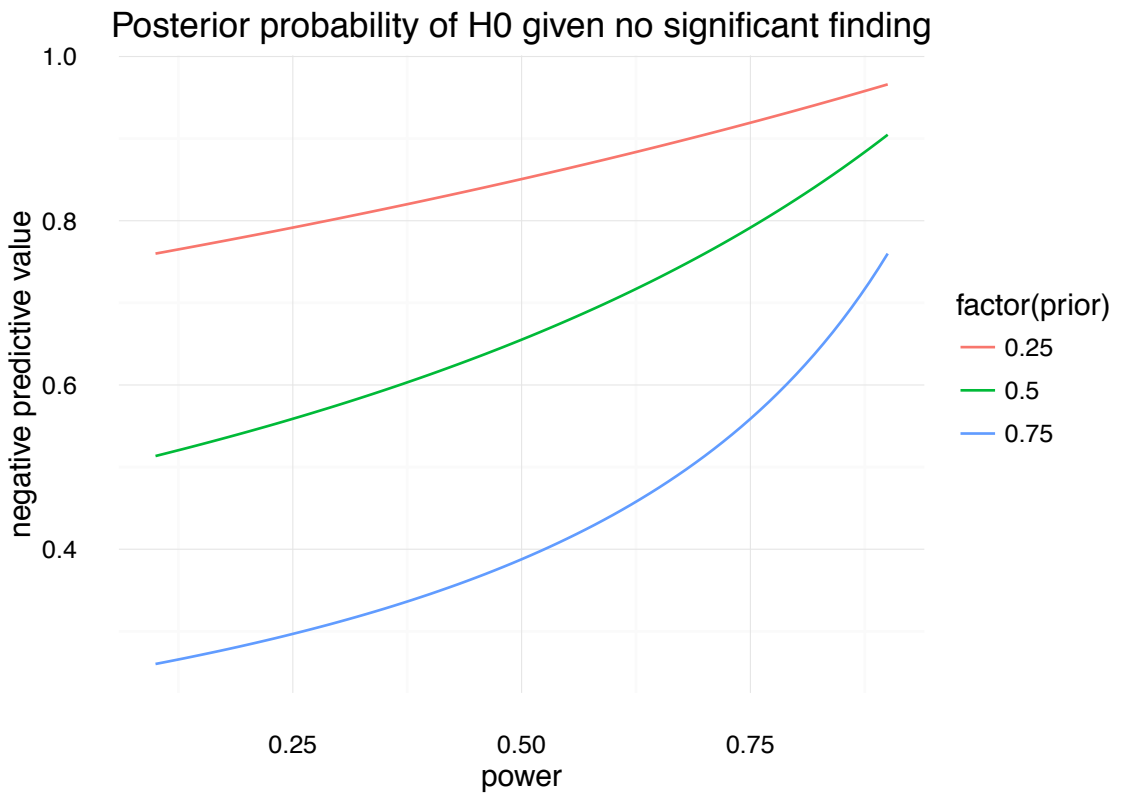
makes no difference at all... but it's equally possible that it leads to a 100% increase.

Because of this, we say that a high p-value is *consistent* with no effect, not that there actually is no effect.

This is a particularly salient point considering that studies often don't have the sample size they really should have. But even large studies can have low power, because the effect that the researcher wishes to detect is small.

The chart below shows that when power is low, not attaining statistical significance makes practically no difference to our prior belief in $P(H_0)$.





One reason statisticians often prefer likelihoods over posterior probabilities – even though it's posterior probabilities that ultimately align best with our research questions – is because likelihoods are much easier to calculate and only since the 90s have computers become sufficiently fast and cheap to make Bayesian analyses anything but a royal pain in the neck.

Another reason statisticians prefer likelihoods is because they are considered to be *objective*. P-values and frequentist analyses in general don't generally talk about prior probabilities, but you can think of them as having a non-informative prior. It depends on the analysis, but often the non-informative prior is a simple uniform distribution: every parameter value is equally likely.

Treating every possible outcome as equally probable sounds fair and balanced and, well, *scientific*, but strictly speaking P-values and non-informative priors are not *objective*, they're *neutral*: they do not give an advantage to any outcome. A neutral prior can be terribly opinionated: you'd have to be insane to think that an unproven cancer drug is equally likely to cure 100% of patients than it is to cure 10% of them, and given everything we know about the universe extrasensory perception is *not* just as likely to exist as not, yet this is how non-informative priors work.

It is also sometimes thought that frequentist estimates provide a conservative lower bound on effect estimates and that opinionated priors (also known as subjective priors) would just make us more optimistic about new treatments, but there is no such relationship between using or not using priors and the conservatism or liberalism of a statistical conclusion. In fact, a common use of priors is in shrinkage estimators, which, as the name suggests, shrink

parameter estimates towards zero or no effect.

Non-informative priors don't always make sense on purely technical grounds as well: if $\theta \in [0, 1]$ then it's perfectly plausible to assume theta is distributed according to $U(0, 1)$, but this would not be a sensible prior distribution for θ^2, as, barring any additional information, the probability density of θ^2 will be higher near 0 than near 1.

Neutral priors are not always a bad idea: they work well when you have little prior evidence in favor or against a hypothesis, when the sample size or power of the study is high (a reasonable prior will be swamped by the evidence in the data anyway) or when used to produce a rough approximation of the true posterior probability. But it remains an approximation and often not a very good one. P-values are likelihoods, not posterior probabilities.

P-values can confirm hypotheses but not help find them

The ideal of science is the brilliant scientist who comes up with a revolutionary new theory. The revolutionary new theory is used to come up with bold predictions, and when those hypothesized outcomes are confirmed, so is the theory.

Real science doesn't quite work that way.

Instead, a researcher will continuously go back and forth between theory and fact: you read some interesting research by colleagues, you craft your own hypothesis, you run an experiment or collect observational data, evaluate the results and consider whether these results make sense given your hypothesis, you tweak the theory a little to account for any discrepancies, and this process of learning and testing goes on ad infinitum.

In this process, researchers have a lot of leeway in how they interpret and analyze the data. For example, let's imagine a study that purports to show that women wear more red when ovulating. If true would establish an interesting parallel between human behavior and the sexual swellings noticable in female baboons and chimpanzees, a valuable scientific finding. (The example is loosely inspired on a controversial study from 2013, which we won't name.) Here's how our hypothetical researchers proceeded:

- The researchers formulate a hypothesis which states that women will wear more red during ovulation. A questionnaire is designed, responses are collected and the data is analysed. However, the effect of ovulation on clothing choices looks to be small or non-existent.
- The researchers wonder if they might have been too strict when coding the data. They revise the definition of *red clothing* to also include pink. After all, pink is a shade of red. They also wonder whether the 12th through 17th days of the menstrual cycle is the right way to identify women "at high risk of conception". A couple of hours browsing around on Google Scholar shows that the medical literature sometimes mentions the 6th through 14th days instead. They try both. The effect still has a rather large P-value, but it looks more promising.

- The researchers get together for another brainstorm and consider that it's really only normal that menopausal women, pregnant women and women on birth control do not experience the same hormonal triggers that other women would. They exclude these groups and the P-value drops a little more.
- The researchers worry that their sample size might be to small to find much of anything, so they run a second questionnaire. They're getting quite anxious by this time, so they rerun their analyses multiple times as more responses come in. After a while the p-value drops below 0.05. $p < .05$ is is acceptable to most scientific journals, so the researchers decide to save time and money and they cancel the afternoon session.
- Looking at the effect size, it appears women wear only marginally more red during ovulation than at other points during the menstrual cycle, but because the data is so noisy it seems safer to just report that *significantly more red is worn during peak fertility* without putting a number on it, and this is the result they publish.

This sort of hunt for effects with small P-values is often disparaged. Statisticians will talk of *fishing expeditions*, *hypothesizing after the results are known* (HARKing), *data dredging* and so on, and students are warned to never, ever succumb to it.

Fishing expeditions are problematic because of regression to the mean: any significant result we see is the result of a mix of intrinsic factors together with random variation around those factors. The largest apparent effects are produced when random error is high, and it is precisely these apparent effects that we've been fishing for. Anyone else trying to find the same effect would be out of luck, as this random component is much more likely to hover around zero next time you take measurements.

Researchers have a lot of ways in which they can tailor an analysis (John, Loewenstein, and Prelec 2012; Andrew Gelman and Loken 2013): tweak the expected outcome, try out different statistical methods including inherently biased methods such as stepwise regression (Whittingham et al. 2006), determine outliers by looking only at how they affect the outcome. Even something seemingly harmless like log-transforming a variable that does not look linear makes P-values uninterpretable and the coefficient of determination R^2 of regression analyses overly optimistic. (See Bakker, Dijk, and Wicherts 2012 a number of simulations.)

Such data-dependent decisions result in very weird null distributions. The interpretation of a P-value becomes something along the lines of "the result we would find given that no effect exists, when we discard data that seems faulty but only if this improves the P-value, and where we pick the most vivid outcome from a bunch of related ones, and where we then pick the statistical method that seems most 'appropriate' which we also decide based on seeing the data." There's very little a scientist can still infer from such a P-value.

But are any of the individual decisions made by the researchers from our hypothetical example really so bad? Is it unreasonable to worry about whether miscategorized data might obscure an effect that is truly there? Is it unreasonable to broaden the outcome from red clothing to *reddish* clothing? Is it unreasonable to exclude women that, in hindsight, they should not have invited to participate in the first place?

To some extent, yes, such an expedition *is* unreasonable, because if the decisions were all so obvious, then the researchers could have detailed them in a protocol before collecting

any data, and any claim that these choices were made on purely scientific grounds is a form of self-delusion. The fact that P-values break down in the face of dubious practices by researchers can hardly be seen as a slight against the P-value; no measure of evidence can protect against questionable research practices.

On the other hand it would be absurd to expect that the only way in which we can ever learn anything from data is by making a bold hypothesis and then if the hypothesis is not clearly and unambiguously supported by the data, to discard the whole enterprise. Exploratory research is a very important part of science.

Is the problem here one that is intrinsic to P-values, or are they simply being used in exploratory research when it should be clear that they can only be used in confirmatory research? A little of both. Researchers *do* need a guide to help them explore data and point out how poignant an effect is when it's not immediately graphically visible. This is not in fact impossible. Two very different approaches available to researchers right now are hierarchical analyses for which analyses are run for *all possible subgroups*, but estimates are regularized to the overall mean to get rid of the optimism. For larger datasets, holdout samples, and for slightly smaller datasets, cross-validation, are an excellent way to mess around with data and explore it in depth while maintaining the capacity to run inference afterwards and subject your guesses to a more rigorous test.

P-values are becoming less and less relevant in world full of data-driven analysis, multiple testing and creative research designs. It is up to statisticians to figure out safer ways to do this kind of exploratory research.

Null hypothesis significance testing

The P-value is a measure of surprise. If you play a couple of sets of tennis with a friend, then even if you're equally matched it's not so surprising that every once in a while you or your friend will win all six games in a set. After all, being equally matched doesn't mean your games will neatly alternate between one win for you, one win for your friend, one win for you, and so on. But if you and your friend are truly equally matched, it *would* be unusual and surprising for your friend to win all games in ten consecutive matches.

Up until now we have talked about P-values as a continuous measure of how *surprising* a statistic would be if we knew for certain that nothing special was going on. But it can sometimes be useful to make hard decisions: if there's a 5% chance that it's going to rain today, you won't bring a jacket to work; if there's a 90% probability that your car will break down in the next month, you will bring it in for maintenance. Instead of suffering through ambiguity and uncertainty on an ongoing basis, we prefer turning maybe into *yes* or *no*.

To return to our tennis match, exactly how many games must your friend win, how wide apart must your scores be before you conclude that your friend is the better player? The p-value is a *measure* but to use it, we must also have a *criterion*, a guide for how to interpret it. To do that, we ask the question: well, let's assume me and my friend were truly equally matched, how *rare* would it be to see my friend take all these wins? If you'd see such an

extreme discrepancy at most one time out of twenty, it might be time to admit that your friend outskills you. If on the other hand a series of unchallenged wins is not unusual even for players of equal skill, you should perhaps give yourself the benefit of the doubt.

This kind of reasoning is known as *null hypothesis significance testing*. The null hypothesis is the idea that you and your friend are equally matched. The alternative hypothesis is that he's the better player. The significance test is the calculation you perform to figure out how unusual those game scores would be if the null hypothesis is in fact true – the P-value – and the decision you make to either categorize the outcome as *significant* or not. The more unusual and thus the lower the probability of seeing those scores *if* you're equally matched, the harder it becomes for you to hold on to the idea that you're matched in skill, and once that probability sinks below 5%, the customary significance level, it's time to face facts and admit you're not as good a player as he is.

We've already seen how P-values can be misleading because they give us the probability of the data (or data that is more extreme) on the assumption that the null hypothesis is in fact true, whereas what we're interested in knowing is the probability of one hypothesis or the other being true. The probability of the data *given* a hypothesis might sound like the same thing as the probability of a hypothesis *given* the data, but it isn't!

But let's put the issues with P-values aside for a moment and focus not on the *measure* but on the *criterion*, the way in which we decide that a result is surprising enough that we can't chalk it up to chance.

The underlying idea of null hypothesis significance testing is that we don't want to jump to conclusions. On TV or in newspapers we sometimes see mentions of worrying trends and alarming connections, but if you actually do the math it turns out that it could just as easily be explained by natural fluctuations. In 2015 my alma mater saw a 14% increase in people starting a civil engineering undergraduate degree. Does this mean that STEM is on the rise, or will registrations be back to the usual levels the upcoming academic year? It's hard to tell. Scientists are wary of such premature conclusions.

But even though null hypothesis significance testing (NHST) sounds perfectly reasonable, it really is not.

First off, NHST is not reasonable because most of the time science doesn't require scientists to make a decision after every experiment and determine *right now* whether something works or whether it doesn't. Instead, science works by shifting the weight of the evidence in favor or against certain ideas over the course of many experiments and data analyses.

Secondly, NHST encourages us to categorize hypotheses as either true or unproven, but this dichotomy is really not very useful. Let's say a study finds that blueberries are good for your health, does that mean they add years to your life expectancy, or just days? If scientists have invented a better battery, are you still interested if that new technology would make your laptop run for half a minute longer or if it reduces manufacturing costs by just 17 cents? You'd probably be a bit underwhelmed, but as far as null hypothesis significance testing is concerned, these are *significant* effects.

Also, scientists tend to equate *unproven* with *false*, but not finding any evidence in favor of a hypothesis is not enough to say with any great confidence that it's false, and confusing the

two can sometimes lead us to prematurely give up on promising research avenues.

Thirdly, the null hypothesis is usually that the effect of an intervention on an outcome is *exactly zero*, or that an intervention and a control have *exactly equal odds* of resulting in a particular outcome. This is not a useful distinction in social science, where almost everything has an effect on everything else, albeit usually an effect that is too tiny too matter. Statistical significance does not imply practical significance.

We will now discuss these problems in more detail.

Significance testing forces decisions where none are needed

Few scientists put much stock in a single study. Instead, science works by shifting the weight of the evidence in favor or against certain ideas over time and getting a better idea of the magnitude of an effect and how it interacts with other phenomena.

Null hypothesis significance testing runs counter to both the ideals and the practice of science as a constantly shifting body of knowledge. Instead of growing knowledge over time, it imposes on science a decision-making context: the idea that after every individual scientific study, we must gather around and decide *on the spot* whether we have found something of interest or have not.

If we take the clinical trial as the archetype of scientific research, then a decision framework is natural: at some point we must decide whether or not we will approve an experimental drug and recommend its use. Any delay in this decision could cause people to receive suboptimal care or die, and pharmaceutical companies and universities design studies in precisely such a way as to enable a fair decision at the end of the ride, which gives pharmaceutical companies at least a little bit of predictability in the otherwise very risky business of pharmacological research and development.

But research in psychology, economics, geography and physics is usually not quite so pressing, and there is no immediate reason

It also leads researchers to desperately try to get a finding into the "significant" category, either through questionable research practices (John, Loewenstein, and Prelec 2012) or through verbal gymnastics: *a tendency toward statistical significance*, *on the verge of significance*, *narrowly evaded statistical significance* and so on. (These gems and many more can be found in a list that epidemiologist Matthew Hankins put together in 2013, cataloguing all the flowery language scientists use to talk about effects that are tantalizingly close to significance but did not reach the 0.05 significance treshold (Hankins 2013).)

It is impossible for an effect to be anything other than significant or not significant, so talking of an effect *flirting with conventional levels of significance* shows a poor understanding of how null hypothesis significance testing actually works. At the same time, though, it is possible to have a little or a lot of evidence in favor of a hypothesis, as well as every degree of empirical support in between. It isn't unreasonable for researchers to ask us not to reject an idea out

of hand and to consider the possibility that an effect might really exist even if the evidence is weak. Scientific standards should not force researchers to dichotomize the evidence.

Furthermore, while null hypothesis significance testing doesn't require any particular significance level, in practice it is almost impossible to deviate from $\alpha = 0.05$, which means that we require preliminary and exploratory research to adhere to the same standards as a phase III clinical trial on which lives depend.

During World War II, Robin Plackett and Peter Burman tried to develop better proximity fuses for anti-aircraft shells. The accepted wisdom at the time was to *only ever change one thing at a time*, but Plackett and Burman did the exact opposite: they designed experiments in which it is impossible to isolate the effect of every individual factor, but because you can test so many factors at a time, you can quickly narrow down your search to a number of likely candidates, which you can then follow up with a more rigorous experiment. These kinds of *fractional factorial experiments*, as they are known, trade precision for speed and are now commonly used in industrial settings. (Box, Hunter, and Hunter 2005, 281)

Conversely, CERN scientists realize that the discovery of the Higgs boson would have an outsized impact on physics and that the enormous amount of data they have at their disposal makes false discoveries likely, so they've decided to put the significance level not at the $5e-2$ social scientists are used to, but rather $3e-7$, a difference of five orders of magnitude.

In those rare cases when a decision framework does make sense, it should at least be possible to set a standard of evidence appropriate to the problem, something that is now the exception rather than the rule.

Significance testing is an empirical steamroller

An analgesic might reduce the intensity of a headache a little or a lot or not at all, and what we would like to know is how many points the average patients drops on a pain intensity scale. Radioactivity is toxic, and we would like to know how many sievert of radiation has what kind of harmful effects (changes in the odds of contacting various ailments, days or years subtracted from one's life expectancy). "Analgesic good, radioactivity bad" is the kind of statement we'd expect from a caveman but it is unfortunately surprisingly close to scientists' tendency to categorize interventions into real and bogus effects.

It is for example common in sensory science, the science of how consumers perceive food and drink, to use significance testing to judge whether consumers can perceive changes in ingredients and production method. The statistical workhorse of sensory science is the triangle test, in which tasters are presented with three food samples, two identical ones and one that's different, and they're asked to identify the odd one out. The hypothesis test asks: *can people identify the odd one out more often than would be expected due to chance, that is, more than 1 out of 3?* It sounds sensible, but the statistical hypothesis is actually very far from the questions that have economic relevance to the food industry: how many people will notice the difference and of those, how many people will and won't like the change and how many won't care? The old statistical adage applies: ask the wrong questions, get the wrong answers.

We are now so familiar with terms like "false positive" and "Type II error" that it doesn't even occur to us that these are loaded terms: most effects are not *true* or *false*, they are small or large or tiny or anything in between, and consequently the real issue is not with false positives but with exaggerated positives. Andrew Gelman refers to these as Type M error: errors of magnitude (Andrew Gelman and Carlin 2014).

The null hypothesis cannot be rejected is a sentence that most scientists know by heart. It is what you write when an effect is not statistically significant and the reason we are so fond of this odd little declaration is because there is almost no other correct way to describe an effect with a p-value that's higher than the desired significance level α.

We would rather not talk of an insignificant effect, because that carries the connotation of unimportant and meaningless. We can't say that the effect does not exist, because our study might have missed it due to a low sample size. We never *accept* the null hypothesis for that same reason. The epistemology of the statistician: *never say never*.

Nevertheless, by anointing some results but not others with the label of statistical significance, we turn a continuous measurement of evidence into a dichotomous judgment: either we have all the evidence we need, or we have no admissible evidence at all.

This bifurcation of scientific findings sometimes leads to strange decisions. Given two hypothetical strategies to reduce global warming, one that promises a 1-3% reduction (95% confidence interval) will merit publication in a leading scientific publication, whereas another other strategy that could lead to a 0-14% reduction will have quite a bit of trouble getting past peer review. Given a choice between a guaranteed five dollars or a 50/50 chance to win 100 dollars, the statistician would prefer the lottery (with an expected outcome of 50 dollars) but the *significista* will take the five.

No serious statistician would ever tell you that a correlation with a P-value above the magical .05 treshold is wrong or does not exist. No textbook will teach you that a significant finding is always of practical importance. A keen observer might even point out that when one level of a factor is significantly different from zero but another level isn't, the *difference* between these levels might not itself be significant (Andrew Gelman and Stern 2006).

But what statisticians say and what they teach ultimately doesn't much matter because every working scientist can *see* all around them that significant findings are treated very differently from findings that are not. In 2015, Blakeley McShane and David Gall put this to the test and sent a questionnaire to researchers who had recently published in the *New England Journal of Medicine*. One of the scenarios describes a random selection, random assignment study of two drugs and in this study 52% of subjects who got the first drug recovered while only 44% of those who got the second drug did, albeit with the difference between these treatments indistinguishable from chance ($p = .175$). Asked which treatment they would recommend, all else being equal, more than 1 in 5 respondents said that they could not recommend one or the other because both treatments were equal; they did not seem to grasp that regardless of whether a difference in the efficacy of two treatments is statistically significant, absent any further information the maximum likelihood estimates are equal to the sample means and thus the first treatment is still a better bet. Remarkably this was no different for participants who received a slightly different scenario with a p-value of

$p = .075$, which is still not significant but does shift the evidence quite a bit in favor of the first treatment.

These results are unfortunate, but not entirely unsurprising considering the growing literature that's out there about misinterpretations of p-values. But Blakeley McShane and David Gall found something much more worrying. They presented a very similar scenario to the last one: the hypothetical outcomes of two randomly assigned last-ditch cancer treatments. The group receiving the first treatment survived for another 8.2 months on average whereas those receiving the second treatment survived for only 7.5 months on average, $p = .27$. The question they then asked was whether, "speaking only of the subjects who took part in this particular study", those who received the first treatment lived on average longer than those who received the second treatment, whether there was no difference or whether we can't determine if there is. Fewer than 1 in 5 researchers responded that *of the subjects in this study*, those in the first group lived longer on average. Let this sink in for just a moment: thinking in terms of P-values has apparently become such an ingrained habit that doctors and medical scientists can no longer see the difference between the numbers 7.5 and 8.2.

Do the exercise yourself: go through a random sample of scientific papers and see how often the abstract mentions significant results and how often the abstract mentions an actual effect size.

Perhaps in spirit null hypothesis significance testing was only ever meant to filter out particularly noisy evidence, as was Fisher's habit, but today it has become an empirical steamroller that pancakes quantitative, continuous effect sizes down to a single bit of information (*it works* or *we don't know if it works*), puts studies with large margins of error in the same pile as studies with small or non-existent effects and sweeps aside posterior distribution, likelihood function, maximum likelihood and anything else that could quantify the weight of the evidence.

Significance is guaranteed and the null hypothesis is never true

Scientific theories are intricate contraptions that they cannot themselves be tested, but we get around this inconvenient fact by testing their *consequences* instead.

The theory of general relativity predicts that light will bend around stars and other heavy objects. Albert Einstein calculated that under general relativity we'd expect light to bend around the sun by precisely 1.75 seconds of declination whereas Newtonian physics would expect a bend of only 0.87 seconds. During a 1919 solar eclipse, coordinated observations in Brazil and on the island São Tomé and Principe validated Einstein's theories and proved Newton wrong.

To get an idea of just how impressive and just how precise Einstein's prediction was: a second of declination is the angle corresponding to 1,296,000th of a circle.

Because of physicists' brilliance at setting up or figuring out situations in which outside influences are minimized and because their theories take the form of mathematical equations

which allow for precise quantative predictions, physicists are usually warranted in equating a high probability of seeing an outcome given a hypothesis with a high probability that the hypothesis is true given the data. That is, in physics we can substitute $P(D|H)$ for $P(H|D)$ without having to worry so much about confusion of the inverse. No other hypothesis could account for what was observed so we have almost no choice but to accept that Einstein's mechanics are superior to Newton's.

Robert Merton's theories about deviant behavior long enjoyed the same stature in criminology that Einstein's theories do in physics. Merton's basic insight was that a person will show deviant behavior when they have no clear and legitimate path towards a culturally highly valued goal. A burglar wants money and nice things just like anybody else in a capitalist economy. A student might cheat on a test if school is important to them or if they want to please their parents, whereas a student who really didn't care about school would simply not study for the test and accept whatever grade they received.

But while formulating hypotheses and predictions to test a theory works great in the natural sciences, its extension to sociology, economics, geography, medicine and a host of other scientific disciplines by means of *null hypothesis significance testing* is not without its problems (Meehl 1978).

1. Lack of exactitude: the human psyche and the social world are incredibly messy, everything can affect everything else, and as a result even the best social science is limited to "if you do x, then people might feel a little bit more y" without the ability to put bounds on how small or how large the effect could be.
2. Lack of universality: usually caveats arise. To give a bit of a facetious example, if your theory is that people will get angry if you punch them, that will probably be true most of the time, but a playful jab from a friend might have the opposite effect, in a sense "disproving" your theory. If more and more of these exceptions accumulate over time we might reject the theory, but we won't throw away a promising theory just because in one particular scenario it doesn't pan out.
3. Lack of interdependence: in physics, tweaking one particular equation might allow you to explain unusual experimental data that no other theory can account for, but almost all straightforward adjustments would then be in direct opposition to everything else we know about how nature works. Even in medicine, where we have a vast amount of anatomical and biochemical insight into the human body, we can often formulate hypotheses that have no bearing at all on this base of knowledge. Because our freedom to formulate hypothesis is much higher, there's also a much higher chance that we'll be wrong.

In the natural and the social sciences alike, it it often impossible to make a precise quantitative prediction. The best we can do is predict that A will have an effect on B *of any magnitude*. As predictions become less and less specific, they also provide a diminishing amount of evidence for the theory that prompted the prediction, and null hypothesis significance testing is at the very extreme of this spectrum.

One can even argue that in many disciplines, Type I error, the risk of finding illusory effects, simply does not exist.

In social science, uncountable factors influence someone's behavior. The fact that I've just had a cup of tea instead of a cup of coffee might conceivably albeit only in some minute way affect how I write this paragraph. A largely ineffective drug might once in a blue moon interact with the patient's state, environment and genetics in just the right way to kill the bacteria or get the immune system moving again. Everything, ultimately, affects everything else. (Keep in mind that when we reject the null hypothesis, we reject that the effect is *exactly* zero or that the odds are *exactly* one, so even the tiniest of effects count.) Geneticist David Lykken christened the phenomenon "the ambient noise level of correlations". (Lykken 1968, 154 via Meehl (1990))

The absence of Type I error leads to a methodological paradox that was first described by Paul Meehl in 1967. Meehl's paradox (Meehl 1967) states that as a discipline's research gains in precision, with the invention of better methods of measurement or through increased sample sizes, the value of a significant finding and the degree to which it acts as evidence for a theory dwindles in proportion. Increased precision in measurement with no increased precision in the statistical hypothesis will lead to 100% of effects being classified as statistically significant, and if effects are uniformly distributed in $\rho \in [-1, 1]$ then 50% of true (but mostly unimportant) effects will be in the direction that was predicted by the investigator.

In physics, advances in scientific instruments make it ever harder for physicists to prove a hypothesis, because it's easier to spot when their predictions don't *exactly* match up with the recorded measurements. But when their predictions *do* align with the hypothesis, the corroboration is much stronger than it would have been if less sophisticated and less precise lab equipment had been used. In the biomedical and social sciences, in stark contrast, better methods paradoxically weaken the evidence.

Significance cripples self-correction

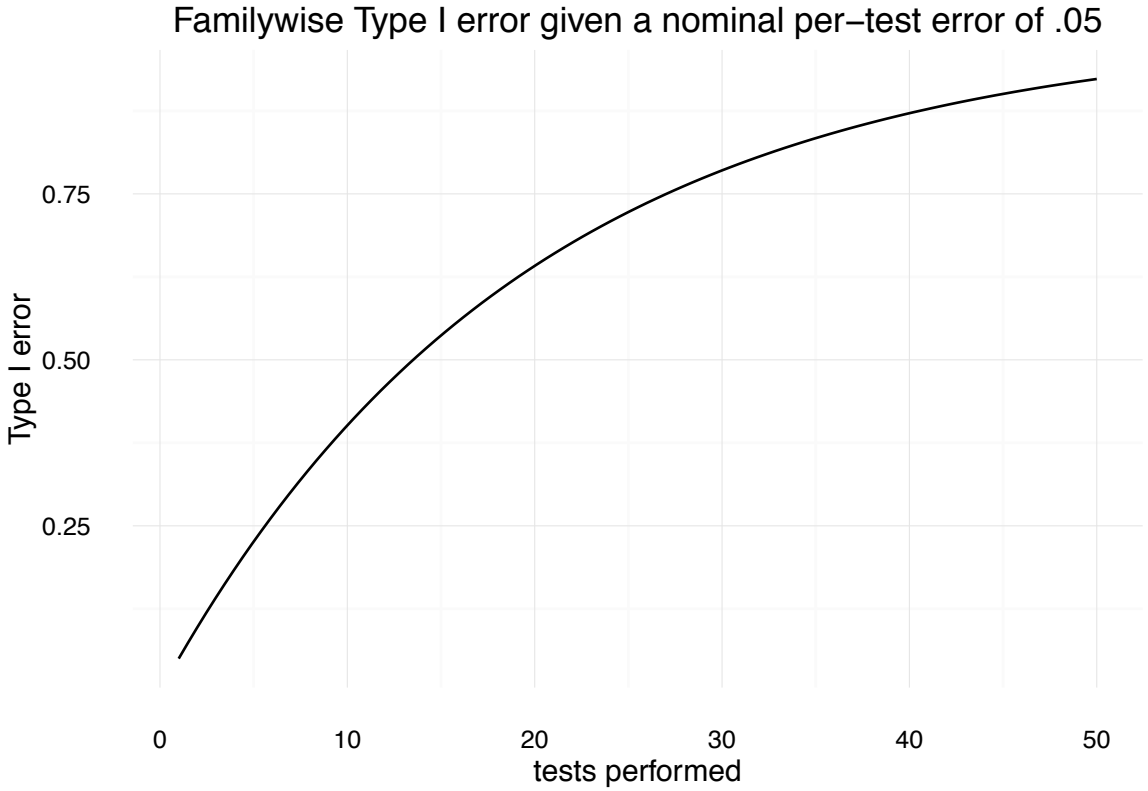
The P-value is first and foremost a measurement of noise. A high p-value means that there's so much noise that any pattern that pops up is the statistical equivalent of an image of Jesus on a piece of toast – not quite the miracle you were hoping for.

To stick to the noise metaphor for a moment, what can you infer when you're tuned to a particular frequency and you're not picking up a channel? Well, you might be getting all of that noise because your antenna isn't strong enough to pick up the signal. It's definitely out there, but you will need a better radio. (Bigger sample, better measurements.) But noise can also mean that there is simply nothing there, no channel to tune into.

It should not come as a surprise that journal editors are wary of publishing scientific research without any statistically significant effects. While it's possible that the researchers found the absence of an effect, which would be useful information in line with that famous Thomas Edison quote that "I haven't failed. I've just found 10,000 ways that won't work", there is also the possibility that the quality of the research is subpar: nonsignificant results, as Fisher suggested, should be *ignored*. This ambiguity in how to deal with results that don't reach significance, together with the fact that not finding anything – Edison notwithstanding

– makes for a boring read, explains how publication bias can take hold: significant results are published, everything else is not.

The result is a multiple testing fiasco on a discipline-wide scale: if many different teams investigate the same phenomenon and these teams have no way and no incentive to broadly communicate failings, then every new publication is another potential source of error: if the nominal Type I error of a finding is 5%, then for n studies that becomes $1 - 0.95^n$.



Given that the even studies that have been decidedly refuted continue to be cited and relied on (Lenfant 2003) the effects of publication bias on the trustworthiness of science are devastating.

Traditionally, meta-analysis was hailed as the solution to unreliable and underpowered research: just take a weighted average of the effect sizes reported by many different studies, and the forest plot or the weighted average will give you an idea of the state of the evidence. Unfortunately, publication bias corrupts even meta-analysis (Kicinski 2013): if only significant results are published, then any average of these results will be biased upwards too. Ingenious attempts have been made at trying to estimate and correct for bias in meta-analyses (e.g. Simonsohn, Nelson, and Simmons 2014, Assen, Aert, and Wicherts (2014)), but it is doubtful that the phenomenon can be corrected mathematically (Andrew Gelman and O'Rourke 2013).

Conclusion

It pays to recall how Ronald Fisher, the statistician who popularized the P-value, suggested we use them:

- We *ignore* evidence because it is of insufficient quality to say anything conclusive. (It has a high P-value.) This is not the same as saying that there's no effect, and it doesn't imply that it can't be fruitful to try again with a bigger sample or a better means of measurement.
- We *accept* evidence when it is of a high enough standard to be admissible (it has a low P-value), but without thereby making any judgment as to its ultimate relevance or importance. We accept evidence when it's worth looking at.

Fisher wanted to make sure we didn't read signal into noise.

Even for this purpose the P-value is ill-suited – the standard error indicates precision, the confidence interval or credible interval indicates the range of plausible parameter values, the point estimate indicates what is most likely – but what is most striking is how the P-value has become in fact the way that we end up confusing signal and noise.

In statistical analyses by Daniel Bernoulli, Pierre-Simon Laplace and Ronald Fisher, the P-value always functions as a sanity check, never as the crux of the argument. Today, it is considered to be the core piece of scientific evidence in favor of a theory.

But to fill this role, we have to interpret the P-value in a more than dubious fashion, and we have to accept a number of assumptions that might look harmless yet are everything but.

1. We had to translate our theory into a hypothesis, and once we've calculated our p-value we have to translate back from statistical evidence for the hypothesis to scientific evidence for the theory. They're not the same thing.
2. Because human behavior and complex physical and biological systems alike are inherently unpredictable, we had to give ourselves the leeway to predict not a precise outcome but rather *any effect at all*, which is terribly vague and often not useful.
3. We avoided specifying when one theory or hypothesis is more or less probable, preferring instead to say any outcome is equally likely. What if we do have some pre-existing knowledge that makes some states of the world more likely than others?

These issues have led to serious misunderstandings of what P-values really are, but they also make P-values intrinsically less useful than they appear to be. The P-value is only useful insofar as it can approximate the questions that scientists are really interested in. But the idea that a P-value is *good enough* and that, if only we interpret it correctly, it works as intended, increasingly looks like wishful thinking:

- A p-value gives us $P(D|H0)$ but we really, really want to know about $P(H_0|D)$ and $1 - P(H_A|D)$ so we just pretend that it is.

- P-values and confidence intervals tell us something about the evidence *given* a certain statistical model, *given* an uninformative prior and *given* that only stochastic but not empirical uncertainty is a factor, but we really really just want to know the chances of being wrong or right, so we assume that a P-value provides this seal of approval when it doesn't.
- A significant result is (incomplete) evidence for an effect different from zero, but we really truly want to know whether the effect is of practical significance and don't want to make the potentially subjective effort to ascertain what the lower boundary of practical significance would be, and neither do we want to submit ourselves to stricter standards of significance when attaining statistical significance can already be a challenge, so we just pretend statistical significance is practical significance.
- Scientific theories are too broad and have too many implications to test straight-up, so instead we must limit ourselves to one of those consequences and convert it into a statistical hypothesis; because translating statistical hypotheses back to the original research hypothesis can be challenging, we assume significant results are intrinsically sufficient to support the theory.
- Confirmatory research works by putting bold, falsifiable hypotheses to the test, but in the social sciences quantitative precision is unattainable so we must satisfy ourselves with weaker hypotheses that some phenomenon or other will occur more or less frequently given a particular intervention, and when this hypothesis pans out we celebrate that our bold conjecture pans out when in fact it had a 50/50 chance of being corroborated.
- We want to talk about causes but must deal with correlations and associations – even clinical trials are observational studies waiting to happen – so we make a half-hearted attempt to warn readers about the difference between correlation and causation and then proceed to speak entirely and unabashedly in causal terms.
- No evidence doesn't mean evidence of nothing... but we really don't want to spend any more time on research that won't pan out, so we drop the investigation no matter how low the power of the study might have been.

As the psychologist Jacob Cohen said about null hypothesis significance testing: "it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!" (Cohen 1994, 997)

But perhaps this is an overly critical assessment of what is ultimately just a particular kind of calculation, and insofar as this calculation was properly done, a P-value cannot strictly speaking be *wrong*. Only poor craftsmen blame their tools. It's not p-values that are problematic but their misinterpretation and abuse. The collection of essays *What If There Were No Significance Tests?* (Harlow, Mulaik, and Steiger 1997) is a good example of this kind of thinking.

Let us compare the P-value to two different technologies that are considered to be dangerous and with which society has dealt in a radically opposed fashion: asbestos and electricity.

When safely tucked away inside a wall, asbestos is a wonderful insulator and poses no health risk. The problem is not asbestos, the problem is how it's used: contractors working on renovations don't always realize it's there and don't take the necessary precautions, engineers

use it in car brakes from where it easily finds its way into the air we breathe, mining operations ignore regulations that ensure safe handling. Whether you consider asbestos to be an *intrinsically* unpleasant building material or one that is misunderstood and misused devolves into a semantic discussion with little value.

The high voltage alternating current that provides our homes with electricity is terribly dangerous and unlike asbestos it *cannot* be tucked away because it is not useful unless you can access it to plug in appliances. However, with the aid of earth grounding, circuit breakers, ground fault interrupters and strictly regulated plug and socket designs, mains electricity is now fairly idiot proof. This is a good thing, because unlike asbestos for which many superior alternatives are available, the next best thing to our AC grid are pneumatic tools powered by air compressors such as those used by the Amish.

Is null hypothesis significance testing dangerous but manageable if we just try a little harder, like electricity, or is it an unmanageable toxic mess like asbestos?

The problem is that p-values are begging to be misunderstood. It is impossible to use a p-value to make a decision without stretching the interpretation of that p-value and confusing the inverse. Students' misinterpretations of P-values are actually the correct interpretation of the P-value *as it is used*. Correctly describing a P-value forces the author into a kind of double talk where you write one thing but hope readers will read it as a posterior probability anyway.

Perhaps the most tragic aspect of having the P-value as the quintessential example of statistical inference is that come to treat them as a ritual, a hurdle to cross rather than as a genuine help in scientific enquiry.

What now?

Very little, actually. Some statisticians have proposed switching to Bayes' factors and credible intervals, others have noted that preregistration can avoid problems with hypothesizing after the results are known, yet others want to see more funding for replications.

All of these are interesting ideas, but to some degree orthogonal to the current discussion. The most important change is to start talking about effect sizes again.

Use confidence intervals to indicate uncertainty, the margin of error to quantify the precision of measurements and power analysis to avoid unacceptably high amounts of Type II error. None of these techniques are perfect – confidence intervals do not account for prior information, the margin of error needs to be compared to the level of precision that is practically relevant, power analysis is susceptible to Meehl's paradox – but their flaws are minor compared to the use of the P-value as a one-number summary of statistical evidence.

In particular, confidence intervals are strictly superior to P-values because they do not confound sample size and effect size.

This isn't always easy, as the appropriate effect size measures are different for different kinds of research (Fidler et al. 2004, Ellis (2010)) and while most scientists know of power analysis they might be uncomfortable actually performing it. Moreover, less attention going to P-values doesn't automatically imply that researchers will pay more attention to causality, confounding, selection bias, experimental design, validity and the challenges of analyzing

messy real-life data. But it also does not require sweeping changes to scientific publishing or statistical education.

It seems that often the main reason to keep using P-values is familiarity and a wish not to rock the boat too much. For example, the *Task Force on Statistical Inference* of the American Psychological Association published an excellent guide in 1999 titled *Statistical Methods in Psychology Journals: Guidelines and Explanations* that covers study design, randomization and causality, psychometrics, sample size, reporting assumptions, what good tables and graphics look like... as well as the encouragement to report confidence intervals and to interpret effect sizes in addition to p-values (Wilkinson and American Psychological Association Task Force on Statistical Inference 1999). But why report P-values at all? And why should interpreting effect sizes be *encouraged* instead of establishing it as a strict requirement?

Statisticians are unfortunately complicit in this status quo. As George Cobb points out in *The Introductory Statistics Course: A Ptolemaic Curriculum?* (Cobb 2007), statisticians have a tendency to just teach whatever academia and industry seems to demand, and this includes a heavy dose of null hypothesis significance testing as well as questionable techniques such as stepwise regression. It is easy to end up in a circular logic where we end up teaching outdated techniques because that is what we believe students will use. But why do scientists use the t-test instead of calculating confidence intervals using Monte Carlo simulations or the bootstrap? Because the t-test is what they were taught in school. And so nothing ever changes.

What to think of the journal *Basic and Applied Social Psychology*, which banned statistical inference altogether (Trafimow and Marks 2015), with the exception of certain Bayesian methods?

To me, it makes sense from a statistical point of view, as it moves us away from having to declare an effect as either significant or not significant, rather than looking at the preponderance of the evidence and building knowledge over multiple experiments. There is no reason at all to force scientists to come to a hard and final conclusion at the end of every investigation.

P-values have also led to an exclusive focus on stochastic uncertainty at the expense of all of the other kinds of uncertainty researchers are faced with: do I have the theory to back this up, am I actually measuring what I am trying to measure, is this coherent with other findings in the field? I do think the editors of Basic and Applied Social Psychology might be right when they say banning P-values will make the quality of the research go up, not down.

But banning null hypothesis significance testing might also be the only *honest* choice we can make.

As it stands, scientists informally apply Bayes' theorem every time they read a scientific study. Few scientists trust the results of a single study, and when especially bold claims are being made, they trust them even less, regardless of whether the P-value is one in two or one in a million. Statistics, after all, can only control for *stochastic uncertainty*, it does not correct for *empirical uncertainty* – uncertainty about whether we've got the right model but also uncertainty about whether those performing the scientific study did sloppy work, or even just a general uncertainty about what we might be missing.

If nobody trusts a single study anyway, we might wonder what the big problem is: as poor studies, such as those with low power or high researcher degrees of freedom, become more common, scientists will adjust the importance they attach to each individual study downwards and an equilibrium is maintained. But of course, if we take this line of reasoning to its logical conclusion, there is no need for any inferential statistics at all: just calculate the group means, compare them, and by taking a look at the sample size and thinking a bit about how variable and how large the effect is that you'd expect, and an experienced scientist will be able to distinguish differences that are real from those that are flukes fairly consistently. So the most advanced statistics anyone really needs is the sample mean and the sample standard deviation and perhaps enough fluency with computers to draw scatterplots. No?

Statistics as a discipline exists because at a certain point scientists realized that it makes sense to put the same rigor and diligence in the way they analyze data that they put in their experiments and theories. But if rigor is what we want, then a P-value that says that given chance there's only a 2% probability of seeing a particular outcome or one that's more extreme *better correspond to precisely that 2% probability if the outcome were due to chance.* If we report a 95% confidence interval, *then 95 out of a 100 of these kinds of confidence intervals had better contain the population parameter.* The fact that we must and in fact do mentally adjust all of these estimates downwards and adjust P-values upwards suggests that they do not achieve their nominal level of confidence. If P-values and confidence intervals don't do what's on the tin, we might wonder what the point even is of reporting a confidence interval with bounds with two decimal places when we are then going to transmogrify it into a more plausible ballpark estimate and thereby show ourselves indifferent to the accuracy.

The economist Paul Romer, talking about the role of mathematical models in economics, writes about "the new equilibrium: empirical work is science; theory is entertainment. Presenting a model is like doing a card trick. Everybody knows that there will be some sleight of hand. There is no intent to deceive because no one takes it seriously." (Romer 2015, 93) The use of inferential statistics in science – from psychology to geology – has not devolved to the point where it can be called a card trick; but every day it becomes harder and harder to argue that it is more than a harmless ritual: before we get down to the real and actual science, we bring an offering of inferential statistics to the pantheon of Fisher, Nyman and Pearson, peace be upon them.

In 1974, Richard Feynman expressed his worry about scientists who followed all the *forms* of science, but not its spirit, and compared this kind of science to a cult (Feynman 1974). What he describes is painfully close to the ritualized statistics of null hypothesis significance testing.

References

Assen, M. van, R. van Aert, and J. Wicherts. 2014. "Meta-Analysis Using Effect Size Distributions of Only Statistically Significant Studies." *Psychological Methods*, no. Advanced online publication.

Bakker, Marjan, and Jelte M Wicherts. 2011. "The (mis)reporting of statistical results in psychology journals." *Behavior Research Methods* 43 (3): 666–78. doi:10.3758/s13428-011-0089-5.

Bakker, Marjan, Annette van Dijk, and Jelte M Wicherts. 2012. "The Rules of the Game Called Psychological Science." *Perspectives on Psychological Science* 7 (6): 543–54. doi:10.1177/1745691612459060.

Box, George E.P., J. Stuart Hunter, and William G. Hunter. 2005. *Statistics for Experimenters: Design, Innovation and Discovery*. Second edi. Wiley-Interscience. doi:10.1198/tech.2006.s379.

Castro Sotos, Ana Elisa, Stijn Vanhoof, Wim Van den Noortgate, and Patrick Onghena. 2007. "Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education." *Educational Research Review* 2 (2): 98–113. doi:10.1016/j.edurev.2007.04.001.

Cobb, George W. 2007. "The Introductory Statistics Course: A Ptolemaic Curriculum." *Technology Innovations in Statistics Innovation* 1 (1).

Cohen, Jacob. 1994. "The Earth Is Round ($p < .05$)." *American Psychologist* 49 (12): 997–1003.

DeBold, Tynan, and Dov Friedman. 2015. "Battling Infectious Diseases in the 20th Century: The Impact of Vaccines." http://graphics.wsj.com/infectious-diseases-and-vaccines/.

Ellis, Paul D. 2010. *The Essential Guide to Effect Sizes*. Cambridge University Press.

Feynman, Richard P. 1974. "Cargo Cult Science." *Engineering and Science* 37 (7): 10–13.

Fidler, Fiona, Neil Thomason, Geoff Cumming, Sue Finch, and Joanna Leeman. 2004. "Editors can lead researchers to confidence intervals, but can't make them think: statistical reform lessons from medicine." *Psychological Science* 15 (2): 119–26.

Fillmore, Kaye Middleton, Tim Stockwell, Tanya Chikritzhs, Alan Bostrom, and William Kerr. 2007. "Moderate Alcohol Use and Reduced Mortality Risk: Systematic Error in Prospective Studies and New Hypotheses." *Annals of Epidemiology* 17: S16–23. doi:10.1016/j.annepidem.2007.01.005.

Gelman, A, and E Loken. 2014. "The AAA tranche of subprime science." *Chance* 27 (1): 51–56. doi:10.1080/09332480.2014.890872.

Gelman, Andrew, and John Carlin. 2014. "Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors." doi:10.1177/1745691614551642.

Gelman, Andrew, and Eric Loken. 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." http://www.stat.columbia.edu/\protect\T1\textbraceleft~\protect\T1\textbracerightgelman/research/unpublished/p\protect\T1\textbraceleft/_\protect\T1\textbracerighthacking.pdf.

Gelman, Andrew, and Keith O'Rourke. 2013. "Discussion: Difficulties in making inferences about scientific truth from distributions of published p-values." *Biostatistics*, no. Advance Access: 1–6. doi:10.1093/biostatistics/kxt034.

Gelman, Andrew, and Hal Stern. 2006. "The Difference Between 'Significant' and 'Not Significant' is not Itself Statistically Significant." *The American Statistician* 60 (4): 328–31. doi:10.1198/000313006X152649.

Hankins, Matthew. 2013. "Still Not Significant." https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/.

Harlow, Lisa L., Stanley A. Mulaik, and James H. Steiger, eds. 1997. *What If There Were No Significance Tests?* Psychology Press.

Heijmans, Bastiaan T., and Jonathan Mill. 2012. "Commentary: The seven plagues of epigenetic epidemiology." *International Journal of Epidemiology* 41 (1): 74–78. doi:10.1093/ije/dyr225.

John, L. K., G. Loewenstein, and D. Prelec. 2012. "Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling." *Psychological Science* 23 (5): 524–32. doi:10.1177/0956797611430953.

Kicinski, Michal. 2013. "Publication bias in recent meta-analyses." *PLoS ONE* 8 (11): 1–10. doi:10.1371/journal.pone.0081823.

Lenfant, Claude. 2003. "Clinical Research to Clinical Practice — Lost in Translation?" *The New England Journal of Medicine* 349 (9): 868–74.

Lykken, D T. 1968. "Statistical significance in psychological research." *Psychological Bulletin* 70 (3): 151–59. doi:10.1037/h0026141.

Meehl, Paul E. 1967. "Theory-testing in psychology and physics: a methodological paradox." *Philosophy of Science* 34: 265–66.

———. 1978. "Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology." *Journal of Consulting and Clinical Psychology* 46 (113): 806–34.

———. 1990. "Why Summaries of Research on Psychological Theories Are Often Uninterpretable." *Psychological Reports* 66 (1): 195–244. doi:10.2466/pr0.1990.66.1.195.

Panhuis, Willem G. van, John Grefenstette, Su Yon Jung, Nian Shong Chok, Anne Cross, Heather Eng, Bruce Lee, and Vladimir Zadorozhny. 2013. "Contagious Diseases in the United States from 1888 to the Present." *New England Journal of Medicine* 369 (22): 2152–58. doi:10.1016/j.molcel.2007.05.041.A.

Pearl, Judea. 2003. "Statistics and Causal Inference: A Review." *Test* 12 (2): 281–345.

Romer, By Paul M. 2015. "Mathiness in the Theory of Economic Growth." *American Economic Review* 105 (5): 89–93.

Rotello, Caren M., Evan Heit, and Chad Dubé. 2014. "When more data steer us wrong: replications with the wrong dependent measure perpetuate erroneous conclusions." *Psychonomic Bulletin & Review*. doi:10.3758/s13423-014-0759-2.

Scales, Charles D, Regina D Norris, Bercedis L Peterson, Glenn M Preminger, and Philipp Dahm. 2005. "Clinical research and statistical methods in the urology literature." *The Journal of Urology* 174 (October): 1374–79. doi:10.1097/01.ju.0000173640.91654.b5.

Schimmack, Ulrich. 2012. "The Ironic Effect of Significant Results on the Credibility of Multiple-Study Articles." *Psychological Methods*, no. Advance online publication (December). doi:10.1037/a0029487.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. 2014. "P-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results." *Perspectives on Psychological Science* 9 (6): 666–81. doi:10.1177/1745691614553988.

Smith, Gordon C S, and Jill P Pell. 2003. "Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials." *BMJ (Clinical Research Ed.)* 327: 1459–61. doi:10.1177/154510970400300401.

Stigler, Stephen M. 2016. *The Seven Pillars of Statistical Wisdom*. Cambridge: Harvard University Press.

Trafimow, David, and Michael Marks. 2015. "Editorial." *Basic and Applied Social Psychology* 37 (1): 1–2. doi:10.1080/01973533.2015.1012991.

Whittingham, Mark J., Philip a. Stephens, Richard B. Bradbury, and Robert P. Freckleton. 2006. "Why do we still use stepwise modelling in ecology and behaviour?" *Journal of Animal Ecology* 75 (5): 1182–89. doi:10.1111/j.1365-2656.2006.01141.x.

Wilkinson, Leland, and The American Psychological Association Task Force on Statistical Inference. 1999. "Statistical methods in psychology journals: Guidelines and explanations." *American Psychologist* 54 (8): 594–604. doi:10.1037/0003-066X.54.8.594.

"Why Most Published Research Findings are False", explained

Stijn Debrouwere (author), Prof. Dr. Els Goetghebeur (advisor)

Contents

Most Published Research Findings are False

Promising cancer treatments go nowhere. There's as many opinions on minimum wage as there are economists. Last year olive oil and blueberries were the key to longevity, now we're not so sure. Psychologists repeat famous experiments but end up with very different results. What, exactly, is going on?

Aside from conceptual criticism ("I don't think this means what you think it means."), scientific research can be double-checked through verification, re-analysis, replication and reproduction (adapted from Gómez, Juristo, and Vegas 2010), and published research findings that are put to the test survive surprisingly few of *any* of these checks.

A **verification** investigates the same data using the same method. A succesful verification shows that no fraud was committed and that no technical errors were made in how the data was encoded and analyzed. In 1982 the *Journal of Money, Credit and Banking* started routinely asking study authors to provide the original data and analyses alongside their manuscripts; when a couple of years later they finally got around to repeating some of these analyses, they found that many authors had made errors in their statistical calculations or had transformed the original data without mentioning precisely how (Dewald, Thursby, and Anderson 1986). In a working paper from 2015 for the Federal Reserve System, Andrew Chang and Philip Li went through a similar exercise: 59 papers from 13 journals were reviewed using author-provided data and code, and even with help from the original authors the re-runs for just 29 papers of those papers produced the same numbers that were used in the published article, which is barely half of them (Chang and Li 2015).

A **re-analysis** investigates the same data using a different method. A successful re-analysis shows that the results do not depend on an ideosyncratic way of looking at the data, but will hold when modeled differently or when better methods are used to deal with outliers and missing data. Much of the early imaging research in neuroscience and cognitive psychology used makeshift methods that all but guaranteed statistical significance even when study subjects responded the same to control and experimental conditions and few of those older studies reach today's publication standards (Vul et al. 2009).

A **replication** collects new data but sticks to the same methods of the original research. A successful replication shows that an effect generalizes beyond the original setting – incredibly useful because experiments must often resort to convenience samples. But many highly touted psychological findings and even visual illusions that were thought to be universal disappear when tested in non-western and non-industrialized cultures (Henrich, Heine, and Norenzayan 2010), and when Brian Nosek and his 269 collaborators tried to replicate 100 high-profile psychological findings, more than half of the high-powered replications did not reach statistical significance, and even when it did the estimated size of the effects was nearly always lower than what the original study had found (Open Science Collaboration 2015).

A **reproduction** collects new data and also changes the analysis or the experimental setup. A successful reproduction shows that different kinds of investigations still lead to the same conclusion, proving that the effect is not an artifact of a particular methodological quirk or limited to a non-realistic setting. In 1988 Linda Mayes and her co-authors published a list of 56 medical topics for which case-control studies had found conflicting evidence (Mayes,

Horwitz, and Feinstein 1988) and in 2013 a controversial article, *A Decade of Reversal*, analyzed a large amount of scientific studies published between 2001 and 2010 regarding medical best practices and showed that, of 363 articles testing the standard of care, 146 recommended reverting back to earlier methods that, in hindsight, worked better than newer practices that were originally hailed as great improvements to the standard of care (Prasad et al. 2013).

	same method	different method
same data	verification	re-analysis
different data	replication	reproduction

In *Why most published research findings are false*, John Ioannidis lists the main culprits: small studies, research into treatments and interventions with smaller expected effects, studies that investigate many different relationships at the same time, ambiguity or flexibility in the analysis and interpretation of experiments, prejudice and conflicts of interest and too many researchers investigating the same phenomena.

Ioannidis' article is a tour de force and at this point just about every scientist has read or at least heard of it. His work has also found its way into the mainstream press, with profiles in The Atlantic and The New Yorker among other places (Freedman 2010, Lehrer (2010)).

Why most published research findings are false packs a lot of information into just over five pages, including a somewhat convoluted series of calculations. The sheer density of information is exhilarating but also makes for a daunting read, especially for scientists whose statistics are a bit rusty and even more so for the interested layman who is wondering what is really going on with this *replication crisis* their PhD friend keeps blabbing about.

What I intend to do here is to explain concepts like *Type II error* that Ioannidis glosses over, provide annotated tables and graphs that explain how different factors like a study's sample size affect whether it's likely to be true and provide easier-to-grasp alternatives to formulas such as $c(R + \alpha - \beta R + u - u\alpha + u\beta R)/(R + 1)$ that are sprinkled throughout Ioannidis' 2005 paper.

Didn't statistics fix this?

When we say there's a *crisis* in science we don't mean to imply that science used to be more rigorous or that yesterday's scientists were smarter than those of today. When in 1835 the physician H.C. Lombard tabulated occupational hazard statistics, he only barely had the common sense to exclude *student* from the list of most deadly occupations (students don't die young yet those who die young are likely to be students) (Wainer 1999); today social scientists routinely adjust for confounding factors and reweight survey results to make them more representative of the target population. Galileo Galilei's famous experiments showing that heavy things do not in fact drop faster than lighter ones were in fact thought experiments which he couldn't be bothered to verify empirically; today this would be considered fraud. In

the early 20th century psychoanalysis thrived despite a total lack of scientific evidence in favor of its approach; now it has become impossible to publish much of anything without a statistical analysis to back it.

The deluge of unreplicable research constitutes a crisis because it has made us realize that scientific findings are much less certain than they appear to be – a sudden realization that has taken us by surprise. Why? Because statistics is supposed to allow us to draw strong conclusions from highly variable data, through statistical tests and randomization and random selection and regression analysis and all of those wonderful things... but it appears it doesn't! 19th century scientists made a lot of stupid mistakes, but they did not casually claim to be at least 95% confident in their results. 21st century scientists have not stopped making mistakes, but with mathematical statistics to back us up, we do claim that *objectively*, it'd be almost impossible for us to be wrong. *Empirically*, however, this confidence in the accuracy of published scientific work is no longer tenable. That makes the replication crisis a statistical crisis, too (Gelman and Loken 2014, from which this dissertation borrows its overarching title as well as angle).

In our defense, statistical promises are never quite as bold as scientists may make them out to be. Yes, for over a hundred years we have been able to routinely calculate the probability of seeing a particular sample correlation between two variables that are assumed not to be associated with each other, and if this probability is very low, then, why, this does lend credence to the alternative explanation, namely that there exists an association after all. But in drawing this conclusion, we forget that statistical tests protect only against uncertainty due to random chance, also known as *stochastic uncertainty*, it does not protect against everything else that might throw a scientist for a loop, like non-representative samples, miscalculations or faulty assumptions.

Even if for the sake of argument we assume that the only mistake we can make is reading too much into the random walks of variable phenomena, our hypothetical 21st century scientist still confuses a 95% true positive rate, which is what statistics guarantees, with a 95% positive predictive value, which it cannot guarantee: positive predictive value depends not on the statistical test but also on the fertility of a particular field of study, the prevalence of true associations among those we choose to investigate.

On the other hand, the statistical methods that are taught to graduate students are still mostly predicated on an increasingly outdated view of what scientific experimentation actually looks like. Let's take a look at some of these newer but potentially troublesome methods and the problems they pose for statistics.

The price of ingenuity

The cutting edge is often seen as the place where the very best research happens, but it's quite the opposite.

Conservative, accepted research practices provide strong guarantees against mistakes and bias: we've had a lot of time to hone them in response to previous mistakes. Modern randomized

controlled clinical trials, for example, have long used placebos to blind the control group to their treatment arm. We now know that we must also blind the attending physician and have learned the hard way that if we don't, the physician will behave differently towards the treatment group and introduce bias by channeling patients with more severe symptoms to the experimental arm or through subtle changes in bedside manner and coding depending on who they treat. Over time pitfalls such as these are discovered and dealt with.

Newer methods might allow us to see and measure things that before were outside of our reach, but because we're not yet familiar with the drawbacks of the method and because we haven't quite figured out when the newer method is better than the older methods, the cutting edge of innovative research is also a scientific wild west.

Subgroup analyses. It wasn't so long ago when randomized controlled trials would try out new drugs exclusively on men (Murthy, Krumholz, and Gross 2004), to avoid whatever effect an experimental drug might have on the female reproductive system among other concerns. We now realize that men and women respond differently to drugs, as do patients of different age and race. It seems natural to categorize drugs into *it works* and *it does not work*, but we now realize that it can pay to be a little more specific and look at the patient's genetics and this realization has led to the promising new field of personalized medicine. Herceptin is one of the bigger early successes: in 3 out of 10 breast cancer patients the ERBB2 gene is overexpressed, which means the gene creates an overabundance of growth factor receptors that in turn encourage breast cancer cells to proliferate; Herceptin blocks these receptors and works wonderfully in this subgroup of patients.

Unfortunately, our knowledge of the human genome is still patchy and generally the easiest way to figure out whether a treatment might work better for one particular group is to run a clinical trial with a broad range of participants and then look at which subgroups respond best. The catch is that such analyses are susceptible to random flukes: group people by particular combinations of age, gender, race, genetic and diagnostic markers and you end up with hundreds of groups, and if it then turns out that the group of middle-aged blue-eyed black females responds well to the treatment but most other groups do not, that's usually just a bizarre coincidence, not a promising new personalized treatment. It does not help that pharmaceutical companies routinely run these kinds of analyses as a way to salvage unpromising research, which is very easy to do: just pick the group of people who responded best and figure out what, if anything, they have in common. (This is known as the Texas sharpshooter strategy: fire a shot in a random direction, draw a target centered around where it hit and congratulate yourself on the unsurpassed accuracy.)

Multiple testing. Functional MRI brain scans face a similar issue. fMRI scans allow us to see which parts of the brain are most active at any given time, and by carefully honing particular tasks or presenting study participants with certain stimuli, we can learn a lot about human emotion, behavior, language and cognition. An *old school* experiment with a treatment and a control group often needs just a single t-test to show whether the treatment was effective, but a brain scan is a three-dimensional picture of blood flow in the brain subdivided into thousands of little cubes known as voxels. Blood flow is a noisy measurement, regions of the brain are not in exactly the same spot for every human being and some voxels in a region can be highly active while others are not, and these scans require laborious

preprocessing and thousands of statistical tests to make sense of. But if you run thousands of tests and each has a tiny probability of returning the wrong result, those tiny probabilities add up and any conclusion we make rests on quicksand.

Presumably exasperated by the outlandish claims of so many fMRI studies, psychologists from the University of California at one point put a dead salmon through an fMRI scan while showing the salmon "human individuals in social situations with a specified emotional valence" and showed that even the tiny brain of a dead salmon can be used to produce highly statistically significant results (Bennett et al. 2011). Neuroscientists have long been aware of the problem and they have come up with different methods to keep the false discovery rate low, but these methods are not consistently or not correctly applied and the correlations found in fMRI studies remain "puzzlingly high" (Vul et al. 2009).

Higher-order effects. When studying human health or social phenomena, often one factor has a pretty straightforward effect on the other and so for example a sociological study might conclude that for every additional year of study, you can expect to earn an additional x euros or dollars. It is amazing how well linear relationships like the above manage to approximate complex systems... but not always. A lot of things interact with a lot of other things. Grapefruit doesn't contain any cholesterol, but it deactivates the statin drugs many people use to lower their cholesterol. According to the latest research salt does not actually affect blood pressure, but it's still not a good idea to eat a lot of it when you already have high blood pressure. The effect of income on political identification is different for those in their thirties, fifties and the retired. These are examples of interactions, moderating effects, varying effects and higher-order effects – really just different names for the same thing. As epidemiology and sociology, among other disciplines, have gotten more advanced, researchers increasingly want to figure out what these varying effects are like, instead of pretending that any outcome is a simple sum of factors.

Statistics can deal with varying effects quite easily, and all of the major statistical software packages support higher-order regression models. But to tease apart these varying effects, we need data for as many different combinations of predictors as possible: we don't just want young and old study participants and participants who vote liberal or conservative, no, we specifically need young conservatives, young liberals, middle-aged conservatives, middle-aged liberals, old conservatives and old liberals. As the amount of potential combinations go up, the amount of participants in each category goes down and some categories might have no data at all. As a result, higher-order factors are much harder to find and scientists' underestimate how many observations or study participants they need to have a reasonable chance of finding an effect if it really exists.

Large-scale exploratory research. Epidemiologists have a wealth of long-term observational studies at their disposal with many, many variables they can study. Sometimes as a scientist you start with a theory but sometimes you have no idea what you're looking for, and you just let the data guide you. These kinds of exploratory analyses can provide inspiration for new drugs, effective social interventions or really just any scientific theory. It's not hard: just run a kitchen sink regression that includes every single variable you measured and let the software figure out which variables are significant. But because exploratory research is, by its very nature, guided by decisions that depend on the data, because the most vivid effects in

the data are likely to be those that are overestimated (a phenomenon known as regression to the mean), and because observational research can never rule out confounding variables, whatever you find in an exploratory analysis can never be used to conclude with any degree of certainty that x causes y and what the magnitude of that relationship might be.

A commonly held misconception among non-statisticians is that big data is reliable by nature of the sheer amount of information that it works on, but in fact large sample sizes only reduce random error and cannot magically correct for biases.

Nonlinear effects. The key insight of toxicology is that everything is a poison in large enough doses, even water, and conversely most toxins have a treshold below which they are harmless. As a result, you wouldn't expect every additional milligram of TCDD (the poison they used on former Ukrainian president Viktor Yushchenko) to increase the probability of death by a flat percentage. Instead, the effect looks more like a hockey stick.

Similarly, economists often work with effects that peter out like price elasticities, effects that have a sweet spot (not too high, not too low) and all other imaginable shapes of curves. It is common in economics to use polynomial regression to account for these effects because polynomials can approximate lines of any shape. Unfortunately, the resulting equations are wonky and polynomial regression has the notorious habit of producing wildly different fitted curves for even small changes to the data, so it's hard to trust these models.

Natural experiments. By randomly assigning subjects to the control and experimental groups, we can ensure apples-to-apples comparisons between different treatments without having to worry so much about unmeasured confounders. Randomized experiments must however trade external for internal validity: the apples-to-apples comparisons give us insight into cause and effect, but they usually limit the intervention to a small group of study subjects which might not be representative of the population at large. To get around this limitation, epidemiologists and economists sometimes try to combine the best of observational and experimental research by analyzing what are known as natural experiments. A very early example is the 1960 analysis by Donald Thistlethwaite and Donald Campbell that compared students who had received a certificate of merit with students who had only just missed the mark and instead received a letter of commendation (Thistlethwaite and Campbell 1960). The differences between these two groups are akin to those winning gold and those winning silver: so small that we can just about consider students randomly assigned to the certificate group or the commendation group and analyze the data as we would any other experiment. If all goes well, a natural experiment is truly the best of both worlds: it avoids the confounding of observational studies and the selection bias of experimental ones. But it can also be little more than wishful thinking, *of course these groups are the same!*, *of course this variable mimics random assignment!* and such wishful thinking leads researchers to much stronger conclusions than are really warranted.

This ain't your grandma's science. We have all of these new research methods at our disposal, a spectrum of personal preferences about how to design an observational or experimental study and numerous statistical techniques to analyze the results. And that's a good thing: across all of science, technological and methodological innovations are making a lot of new things possible. But statistics and methodology often has to play catch-up, and as a result we're continually flooded with unreliable scientific research.

When 31 teams of researchers were asked to investigate whether soccer referees are more likely to give red cards to dark skinned players, they got back 29 different probabilities with 20 teams finding some degree of racism but the other 9 not finding statistical significance (Silberzahn et al. 2014).

The poor quality of today's science is not because researchers are too convervative, but because they always try to stay on the cutting edge. As the statistician Andrew Gelman quips: n is never large, because as soon as your sample size is large enough to comfortably prove your point, there will always be that healthy temptation to go just that one step further (Gelman and Hill 2006, 481). We do need scientists to be bold, but the unfortunate fact is that, when it comes to science and statistics, being bold can get you in trouble.

Practices such as large-scale exploratory research and unvetted research methods result in bias, and bias affects the positive predictive value of research, that is, it affects the probability that the discoveries you read about in academic journals are actually true.

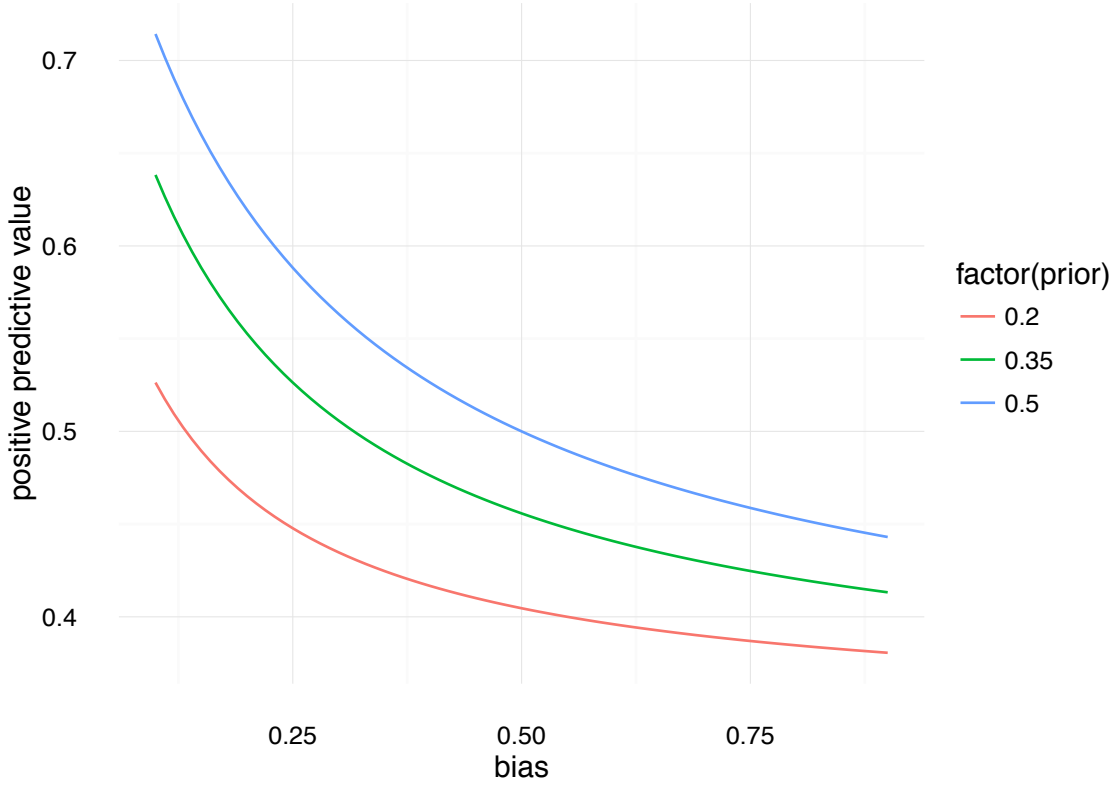


Figure 1: Positive predictive value given medium power (0.5), different levels of fertility (0.2, 0.35, 0.5) and different levels of bias

Newer fields like personalized medicine run a different risk: we don't know yet whether or not there's much to find. And you can spend days and days looking for the monster of Loch Ness, but if it doesn't exist you won't find it. When a field of study has low *fertility* (which we won't know), research in that field will have low positive predictive value.

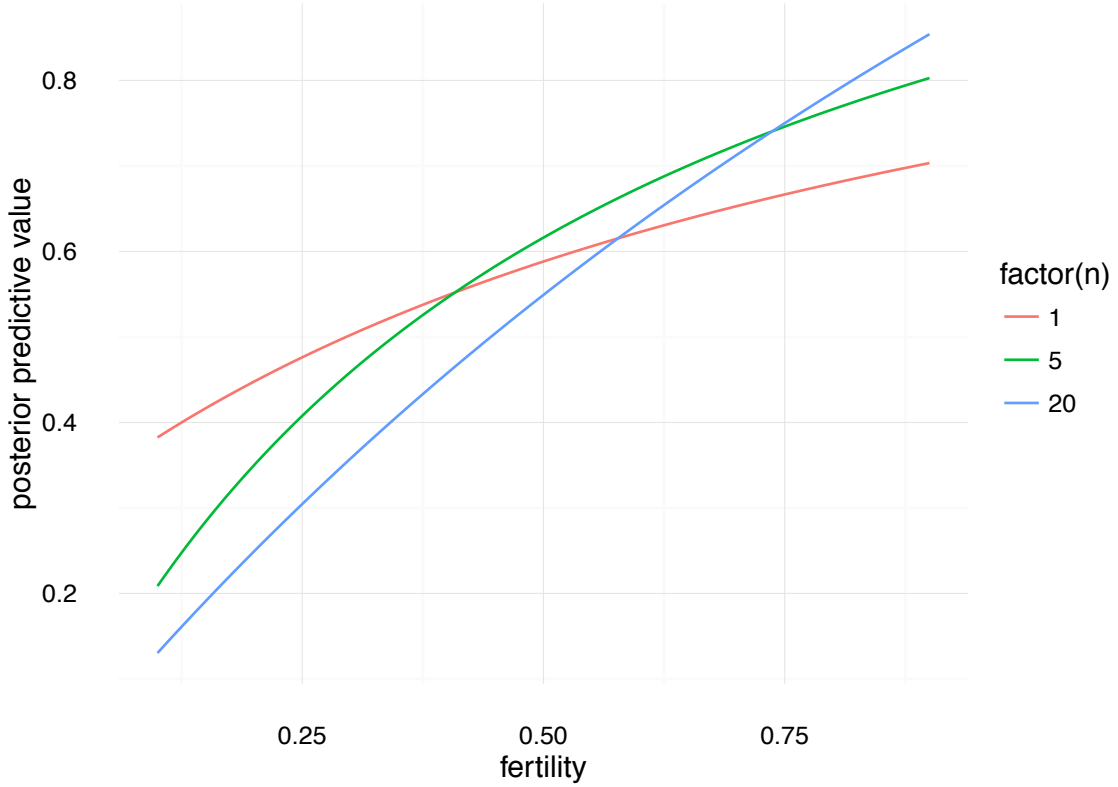


Figure 2: Positive predictive value given low bias (0.25), medium power (0.5) and 1, 5 or 10 investigative teams per association

Take a look at the `roc` function for the R programming language, found in the appendix. It allows you to play around with various values for Type I error, Type II error, fertility, bias and multiple testing (of which more later) and see the results on positive predictive value.

The overall model is as follows:

	positive finding	negative finding
real	fert. × true pos. rate + bias × false neg. rate	arid. × false pos. rate + bias × true neg. rate
not	arid. × true neg. rate	fert. × false neg. rate

It is simple but surprisingly effective at teasing out the implications of interactions between prior probabilities, power, bias and multiple teams / multiple testing.

The low power lottery

Seen through a statistical lens, a scientific study is a trade-off between the risk we want to run to accept fiction for truth (Type I error, false positive rate in the model above), the risk we want to run to accept truth for fiction (Type II error, false negative rate in the model above), the cost of gathering more data and and whether we're looking for something that is easy or hard to find.

Would you like to run no risk at all of committing a Type I error? That's easy enough: never claim to have discovered anything. Of course your Type II error will now be astronomically high, you will have many false negatives. Are you interested in investigating subtle, tiny effects? Get the largest sample you can and then make it twice as large, or accept that you either won't find anything or will find a lot of things that later on will turn out to be spurious correlations. Cheap, fast, good: pick two. Low Type I error, low Type II error, small samples or small expected effects: pick three.

Given this inevitable trade-off, which of these four factors do scientists tend to sacrifice?

- We have by consensus set the highest acceptable Type I error at 5% and that's going to stay where it is, so unless we personally want to adhere to an even stricter standard this factor is outside of our control.
- As science advances there is less and less how-hanging fruit, so we can't get around studying smaller effects.
- It might seem natural to just conduct larger studies, but because the variation of a sample mean around its true mean is inversely proportional to $\sqrt{n}$, we would need four times as many observations to cut the margin of error of an estimate in half, so large samples are not cost effective. In any event, most labs don't have the money for larger studies.

We sacrifice the only remaining factor, power.

Power ensures that *if* something is to be found, we will actually find it. There are currently perhaps 5,000 black rhinoceros left in Eastern and Southern Africa. That's a landmass of more than 10 million square kilometers, so you might have to walk around for a very long time before you come across a rhino and even if you don't find one, that doesn't mean you can conclude that all black rhinos are extinct. As the result of our unhappy trade-off, this undecidability is precisely the situation scientists now find themselves in.

The immediate result of underpowered studies is that in most scientific disciplines, you really cannot trust results that are not statistically significant to imply that no effect exists, because there's a lot of research going around where the chances of finding anything at all were slight to begin with. This is why scientists describe use the bizarre incantation *the null hypothesis could not be rejected* when they weren't able to find a relationship between two variables. A scientist might secretly prefer to say that they've found evidence against the existence of such a relationship, but that'd be like saying your car keys do not exist because you can't find them.

Underpowered studies also have a secondary effect, which is much more insiduous. Let's call it the low power lottery.

To obtain a statistically significant result with a small sample, the effect must be very large. Or rather it must appear to be very large, because whenever we find a large effect in a sample it can either indicate an average population effect truly of this size, or an effect that was overestimated relative to the population. If we mostly study large effects, overestimation doesn't happen as often. But if we mostly study small or uncertain effects, exaggerated becomes a bigger problem and there's even a chance of finding a statistically significant effect that's in the wrong direction. Andrew Gelman calls these Type M and Type S error, errors of magnitude and sign.

How big of a problem? Here's an example of what happens when you try to detect a difference between two groups of 0.1 standard deviation using 30 subjects in total.

The power of such a study is only 6%. To find a statistically significant result, it has to exaggerate the true population effect by a factor of 3.7. There's a 22% chance that any statistically significant effect you find will be *negative*, the opposite direction from the true effect.

n	d	$\approx r$	power	minimal Type M	Type S
10	0.1	0.05	0.05	7.29	0.34
20	0.1	0.05	0.06	4.70	0.27
30	0.1	0.05	0.06	3.74	0.22
10	0.3	0.15	0.07	2.43	0.12
20	0.3	0.15	0.10	1.57	0.05
30	0.3	0.15	0.12	1.25	0.02
10	0.5	0.24	0.11	1.46	0.04
20	0.5	0.24	0.19	0.94	0.01
30	0.5	0.24	0.26	0.75	0.00

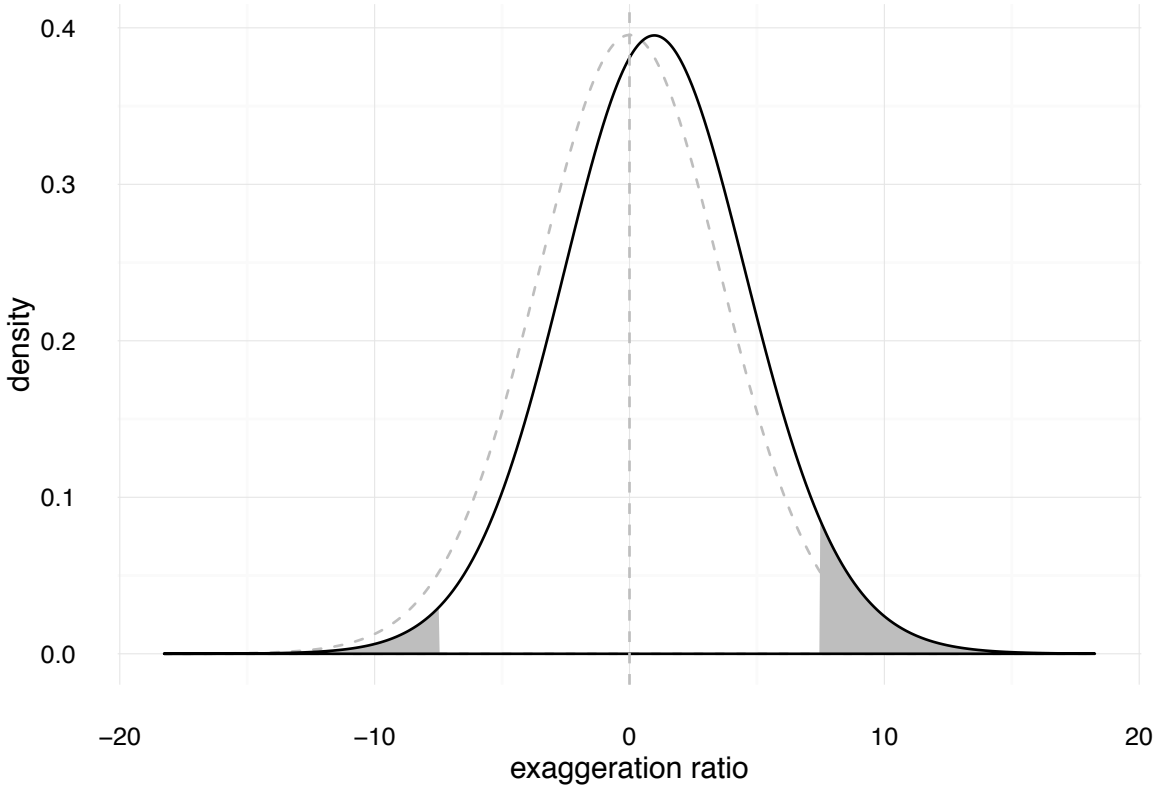


Figure 3: Type M and S error for a two-sample t-test, d=0.1 and n1=n2=15

In and of itself, scientific research that has low power, small expected effects and small sample sizes is not a cause for alarm. It is a waste of time and taxpayer's money and it reflects poorly on the primary investigator, but underpowered studies pose no threat to the underlying statistics and how we interpret them: if 17 studies find no effect or a tiny one, but a 18th fluke study gets (un)lucky and finds an oversized effect through an unrepresentative sample in which the magnitude of the real effect is exaggerated, then few people would find the evidence from this 18th study to be very convincing. Neither would a formal statistical meta-analysis. The very essence of statistics is that sometimes, strange things do happen for no reason. Large effect sizes are often adjusted downwards by follow-up research (J. P. a Ioannidis 2008) and this is annoying but one could equally argue that these adjustments are actually an example of the wonderful self-correcting nature of science: in the end, we get where we need to be.

But what if those first 17 studies never make it to publication and the 18th study does – it is after all the only study to report a significant effect? 100% of the available evidence will then point to a very large imaginary effect.

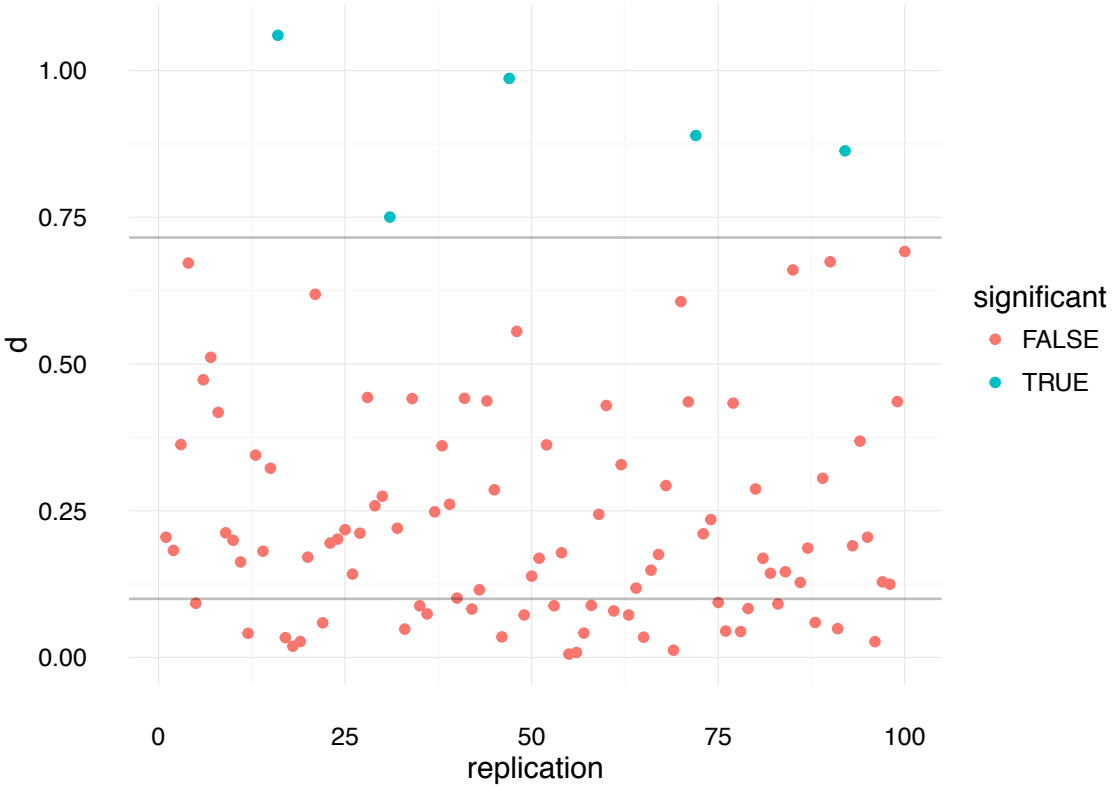


Figure 4: Nonsignificant results stay under the radar, significant ones are inflated and get published.

The resulting bias is known as *publication bias* and also sometimes called the file drawer effect, because researchers tend to ditch studies that did not result in statistically significant findings in a proverbial file drawer. Type I error is then no longer equal to α but to $1 - (1 - alpha)^n$,

not because of multiple testing within a single study but because of multiple *teams* all probing the same correlations.

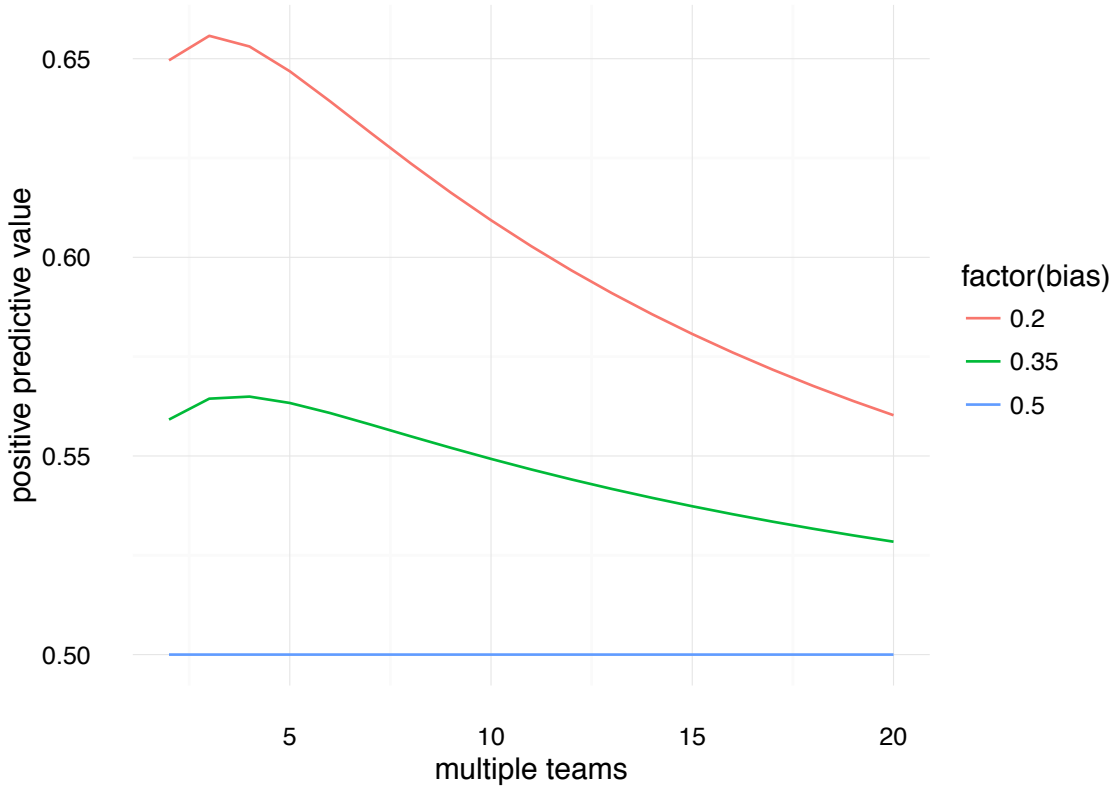


Figure 5: Positive predictive value given medium power (0.5), different levels of bias (0.2, 0.35, 0.5) and teams probing the same associations

The file drawer effect might sound bizarre and unrealistic: surely if different labs study the same thing, they'll keep each other in the loop? But the scientific community is a big place and the main fora for sharing scientific knowledge, academic journals and conferences, are usually not interested in null findings. Scientists themselves usually are not, either: experiments that don't result in significant findings can be seen as failings, not something you would widely share with colleagues at different institutions. Furthermore, biased statistical results usually are the result of a combination of publication bias, liberal statistical analyses, post-hoc hypothesizing and so on all in a single package, so it does not literally require ten or twenty studies to win the low power lottery, just a couple.

Publication bias can easily be visualized using a funnel plot. A funnel plot is a plot of different studies that share the same intervention and outcome, with the point estimate on the x-axis and the sample size on the y-axis. Other funnel plots might put the 95% confidence interval on the x-axis and 1 over the standard error on the x-axis, which amounts to the same thing. If no bias were present, you should see a nicely symmetrical pyramid shape: estimates roughly centered around the true effect size, but spread out more at lower sample sizes. Instead, funnel plots of actual published research tend to be curiously lopsided, show a gap in the

middle or too many extreme estimates, with few small studies reporting small effects.

One particularly vivid example is included in a meta-analysis from 2009 that reviewed whether setting a minimum wage leads to fewer available jobs, which shows evidence of both publication bias *and* ideological bias, where a positive effect of minimum wage on job creation is less likely to be reported. (Doucouliagos and Stanley 2009)

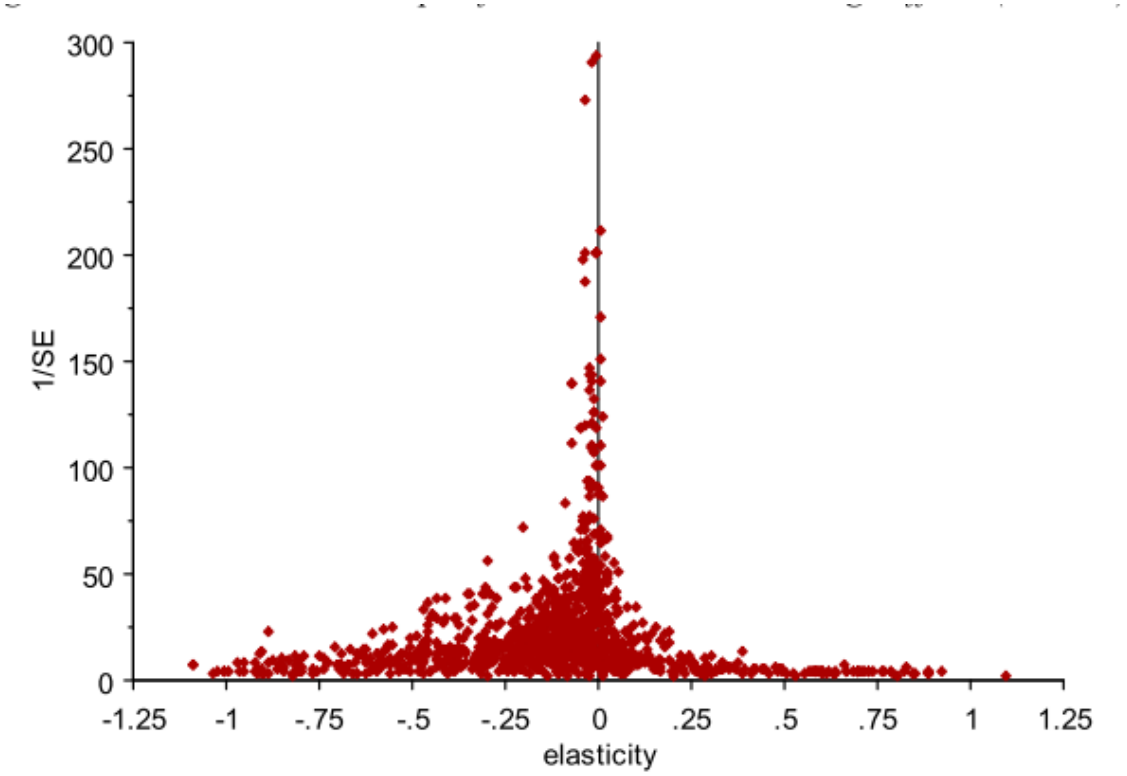


Figure 6: Funnel graph of estimated minimum-wage effects, n=1424, showing publication bias and ideological bias; excerpted from Doucouliagos and Stanley 2009

Perverse incentives

Usually, when talking of **conflicts of interest** we think of pharmaceutical companies sponsoring research or conducting it in house, or of think thanks masquerading opinions as research, and these are indeed good examples of situations in which scientists are subtly or not so subtly pushed towards the outcome that is to the financial or ideological advantage of the sponsor.

Medical journals are increasingly endorsing the CONSORT standards which require publication of a protocol before conducting the study and transparency about who funded the research and other potential conflicts of interest. One of the meta-analyses that motivated

the CONSORT standards calculated that industry-sponsored research has four times the odds of reaching the conclusion most favorable to the sponsoring party (Lexchin et al. 2003). It's an uphill battle and a particular problem for the trustworthiness of medical research, but it's also a fairly obvious problem so we won't discuss it in detail.

The most common form of bias is much subtler and not due to an explicit agenda or conflict of interest but due to perverse incentives and wishful thinking.

Every scientist hopes they will one day discover something hitherto unknown or construct a theory that finally makes sense of unexplained observations, and while this desire is the fuel that keeps them going, it also runs counter to the scientific imperative to be our own devil's advocate, to triple check our results and to not make any claims without strong evidence. When an experiment returns a statistically significant result, it is tempting to shut off our brains for a minute and conclude *well, the stats work out, this is a bona fide scientific discovery* even though in reality inferential statistics are closer to a sanity check and just one small part of what makes a scientific finding credible. When an experiment does not return a statistically significant result, it is tempting to go and hunt for small changes to the analysis that might still vindicate your hypothesis – get rid of an outlier, dichotomize a variable, change from a parametric to a nonparametric test.

Bias also manifests itself in **the choice of what to study and how**. Particularly in psychology, journal editors put a premium on results that are counterintuitive or unexpected; people (and scientists are people!) enjoy reading narratives that proclaim that *everything you thought you knew is wrong*. One well-known example is a study by Daniel Oppenheimer and colleagues that was published in 2010 with the particularly catchy title "Fortune favors the bold (and the italicized)". The study purported to show that text printed in small and grayed out fonts is remembered more easily than the same material presented in more legible script, presumably because readers have to put in that extra bit of effort to decipher everything and the effect encourages their minds stay sharp. It's a wonderfully counterintuitive finding. Two years later Hannah Haysom at the University of Queensland tried to replicate the effect but couldn't, and more recent attempts haven't been able to either (Meyer et al. 2015).

Overturning conventional wisdom is a noble undertaking, but counterintuitive hypotheses are by definition less likely to be true. Given that many of these results come from underpowered studies, it is not surprising that these "cute" research hypotheses rarely survive replication.

Scientists also have **reasons to stick to small and underpowered research** despite its severe shortcomings. The old Upton Sinclair quip is very apropos here: "it is difficult to get a man to understand something, when his salary depends upon his not understanding it!" Given a choice between directing the fixed budget of a research grant towards data collection or towards salaries, scientists have a vested interest in choosing the latter. Papers in scientific journals often conclude with the cliché that more research is needed, but it isn't so hard to argue that in fact *less* research is needed, as Trish Greenhalgh at the London School of Medicine insists (Greenhalgh 2012). Rodger Kessler and Russell Glasgow even go so far as to suggest a moratorium on randomized controlled trials in healthcare (not pharmacology) until we have figured out how to establish a clear pathway from trial to practice and policy (Kessler and Glasgow 2011), as so much medical research is lost in translation. But unavoidably fewer studies means less work for scientists.

The **statistical analysis** of experimental and observational data provides another locus for bias.

In an ethnography of cognitive psychology labs, sociologist David Peterson describes how teams of researchers over time start to rationalize questionable research practices and adopt a *bend-but-don't-break philosophy*: they would never fake data but might ignore sloppy data coding, they would never just run a battery of alternative statistical procedures and pick the most flattering but might massage the data under the well-meaning guise of cleaning it. Peterson includes one particularly pithy quote from a researcher: "You want to know how it works? We have a bunch of half-baked ideas. We run a bunch of experiments. Whatever data we get, we pretend that's what we were looking for." (Peterson 2015, 6)

Even scrupulous scientists are inevitably drawn towards this moral gray zone: scientific findings published in high-profile scientific journals are the primary form of scientific currency and publications determine who gets tenure and grants. Conscientious scientists will on average have fewer publications and thus are selected against and slowly driven out of academia (Smaldino and Mcelreath 2016). Focusing on quantity over quality, on conducting as many underpowered studies as you can in the shortest possible amount of time, "often represents a more efficient research strategy (in terms of finding $p < .05$)" (Bakker, Dijk, and Wicherts 2012, 543).

The shift happens so gradually and come to seem so normal that few scientists are aware that these practices jeopardize their work; they either don't see the harm (John, Loewenstein, and Prelec 2012) or are not aware that their actions bias the outcome (Gelman and Loken 2013). As a result, the desperate pleas see to increase the quality of research by the likes of Douglas Altman and Tom Lang in medicine (Altman 1994, Lang (2004)), Paul Romer in economics (Romer 2015), Tal Yarkoni in neuroscience (Yarkoni 2009), Marjan Bakker and Jelle Wicherts in psychology (Bakker and Wicherts 2011) and their exhortations to please consider larger samples, good experimental design, valid proxies, clean statistics... these recommendations all fall on deaf ears because readers and listeners nod in agreement, remark to themselves "yeah, some researchers really are clueless", and it never occurs to anyone that the criticism might apply to them personally.

If scientists don't realize that criticism of questionable research practices applies to them, it makes no sense to try and improve statistical education, as it will fall on deaf ears. Instead, it seems wiser to fight against the perverse incentives that cause scientists to adopt these questionable practices in the first place (Nosek, Spies, and Motyl 2012).

A perfect storm

Science is hard. Biologists estimate that over one in four cell lines used in drug research are mislabeled or contaminated with other cells (Lorsch, Collins, and Lippincott-schwartz 2014). One in five published genomics papers that use Excel were recently found to suffer from the software's automatic conversion of gene indicators into dates, e.g. turning SEPT2 into the 2nd of September (Ziemann et al. 2016). The majority of published research papers contain statistical errors (Strasak et al. 2007), often tiny ones like off-by-one errors in a test's

degrees of freedom (Bakker and Wicherts 2011) but inappropriate models and tests are not rare. Decades of evidence pointing to the beneficial effect of moderate alcohol consumption on heart health was overturned in the 2000s because few of the original studies had properly accounted for the fact that, as a category, non-drinkers include former alcoholics and those who abstain for medical reasons, making drinkers look healthy in comparison (Fillmore et al. 2007).

Philosopher José Ortega y Gasset once wrote that "experimental science has progressed thanks in great part to the work of men astoundingly mediocre, and even less than mediocre. That is to say, modern science, the root and symbol of our actual civilization, finds a place for the intellectually commonplace man and allows him to work therein with success." (Ortega y Gassett 1930, 110 as popularized by J. R. Cole and Cole (1972)).

Lousy research is refuted by other scientists, meta-analyses distill the truth from noisy studies, replications determine whether earlier research generalizes beyond its original setting and independent re-analysis provides for independent verification. Science can survive mistakes, incompetence, bias and even fraud: it is self-correcting.

The most dangerous kinds of scientific practices, then, are those which corrode science's self-correcting mechanisms. This is what makes publication bias the ultimate threat to scientific progress. Today's scientific weather is a perfect storm of antiquated statistical standards, an overabundance of promising but insufficiently understood techniques and methods, perverse incentives that lead to questionable research practices and investigations into intrinsically unlikely hypotheses using underpowered studies, with publication bias as the ultimate catalyst, the self-imposed veil that causes "null findings" to be swept under the rug. Together, they create a corrosive mix that severely impairs science's ability to self-correct.

Higher standards

The knee jerk reaction of those who first learn of the replication crisis is to demand more stringent standards. If we require P-values to be at or below 5% to declare a statistically significant finding, perhaps we should henceforth require at or below 1%? But as we saw earlier when discussing the low power lottery, statistical standards are not really like bars to jump over, they're akin to communicating vessels that maintain equilibrium. Shrink Type I error and a heap of Type II error will take its place. Demand higher sample sizes and scientists will use it as an excuse to hunt for patterns that are harder to detect. Ask to see power calculations before data is collected, and wishful thinking will lead the investigator to plug in a high expected effect size which in turn keeps the minimum required sample size small.

Given that studies in psychology, for example, even now rarely attain 50% power, more stringent standards in that field without concomitant changes to research practices would further increase Type M error.

The replication police

Scientists like Seth Roberts contend that the core mission of science is to keep discovering new things about human behavior and the natural world, not to be right all the time. False discoveries will be filtered out eventually anyway. In fact, before his untimely death Roberts was an advocate for large-scale citizen science and self-experimentation: the idea that we can learn a lot from noisy, inexpertly measured and unblinded data collection such as people experimenting at home with different diets, exercise regimens and so on (Roberts 2004).

All the time spent on replication, according to this argument, is time not spent on *real science*, not too different from medieval scholars who would spend their entire lives ruminating on the meaning of a particular paragraph of Galen of Pergamum's writing.

John Ioannidis himself has warned scientists that it's not easy to find a solution to the replication crisis without collateral damage: we can throw away all research with even minimal bias, but then we'd never learn anything (J. P. a. Ioannidis 2014).

New avenues for research have more than once started with small samples, quirky statistics and dubious interpretations. Ronald Fisher, for example, makes a strong case that Gregor Mendel's experiments on inheritance in garden peas were fabricated (Fisher 1936), but it nonetheless helped launch modern genetics.

One case worth thinking about is an experiment by chemist Georg Wittig from 1960 that did not survive replication and was formally retracted by Wittig in 1964, only to be resurrected in 2015 by another team led by Peter Chen who found out that Georg Wittig was right all along (Künzi et al. 2016), leading the science blog *Retraction Watch* to ask "50 years later, is it time to retract a retraction by a Nobel prize-winning author?" (Perkel 2015)

Georg Wittig's original experiment described a novel method for cyclopropanation, which is a chemical process that shapes hydrocarbons into ring forms by introducing an additional carbon element and letting it steal two electrons from a triple bond or from two double bonds. These ring molecules are used to create antibiotics as well as cyclopropane, which used to be an important anesthetic and is where the process got its name. Making these ring molecules is not an easy process, so the new method was an incredible advance for chemistry. Except that when other chemists tried to replicate the experiment in 1964, they could not, leading to Wittig's retraction. But in 2015 the aforementioned team at ETH Zurich figured out what the problem was: to work, the process needs a small amount of nickel as a catalyst. Labs back then did not have the sophisticated machinery we do now to avoid contamination by trace elements, so nickel contamination is probably what led to Wittig's original discovery.

Strictly speaking Wittig's retraction still stands: without nickel, the reaction does not work as stated. But imagine a different world, where the failed replication in 1964 had not been interpreted as relegating Wittig's technique to the graveyard of unproductive scientific ideas, but instead as a prompt to try harder, and to try and explain why these two results disagreed. In such a world we might not have had to wait until 2015 to rediscover what we already knew in 1960.

This concern is echoed by a small but vocal clique of present-day scientists. As a biologist, Mina Bissell worries that replications might lead us to prematurely give up on promising

lines of research and that they exalt the null results of mediocre scientists at the expense of top scientists who *did* manage to come up with important new findings (Bissell 2013). Daniel Gilbert talks of a "replication police" in psychology and how they are "shameless little bullies" (Gilbert 2014). They worry that replications can be underpowered, run by scientists with a vindictive agenda and poorly executed.

The critique of Bissell and Gilbert can be interpreted as nothing more than a useful reminder that replications need to ensure that they've taken every possible care to ensure that the effect *can* be found if indeed it's there.

But their critique contains a more harmful message as well, suggesting that only scientists who have been able to find similar effects in the past should be allowed to conduct a replication, and that failing to replicate previous studies just makes you a schmuck. In this form, the critique is really not so very different from demanding that the efficacy of homeopathic medicine only be verified or rejected by homeopaths. One scientist disparagingly calls this the *Harry Potter Theory* of replication (Neuroskeptic 2014): running an experiment is like casting a magic spell, wizards can do it but muggles can't. Therefore, failing to replicate a study just shows you lack the magic touch.

Scientists are a smart bunch, so it is no surprise that they can find creative ways to defend their theories in the face of contradicting evidence, forgetting that science, unlike law, is at least in principle a collaborative search for the truth and not an adversarial process.

Paul Meehl blamed the slow progress in much of psychology on this tendency of researchers to get defensive about their pet theory: "There is a period of enthusiasm about a new theory, a period of attempted application to several fact domains, a period of disillusionment as the negative data come in, a growing bafflement about inconsistent and unreplicable empirical results, multiple resort to ad hoc excuses, and then finally people just sort of lose interest in the thing and pursue other endeavors." (Meehl 1978, 807)

Perverse incentives

Attempts at setting higher statistical standards are likely to backfire, but we can set higher standards in other ways. We have already mentioned how the CONSORT guidelines encourage or require full disclosure of any conflicts of interest and preregistration of the protocol. The Panton Principles (Murray-Rust et al. 2010) and Science Code Manifesto (Barnes 2011) are making scientists aware of the need to freely share code with colleagues to allow for their analyses to be verified independently. Fields like neuroscience are slowly converging around a couple of accepted techniques for analyzing fMRI data, all of which keep the false discovery rate low. The journal *Basic and Applied Social Psychology* no longer requires null hypothesis significance testing (circumventing the low power lottery) and PLOS ONE published a collection of papers in 2015 dedicated to "Negative, Null and Inconclusive Results". *Scientific Data* and dozens of other scientific journals now exist dedicated solely to the publication of data so researchers can get credit for good experimental design and not just for significant findings.

With all of these high profile examples, though, it is easy to forget that today's scientific institutions and practices are overwhelmingly the same they were 10 or 20 years ago. It is easy to be lulled into complacency by thinking of the replication crisis as something akin to growing pains. Growing pains, after all, disappear without much active intervention. But consider that much of what John Ioannidis described in 2005 in large part echoes work by Paul Meehl in the 1960s. Fifty years is a long time and if anything the pressure to *publish or perish* has become worse, with more perverse incentives than ever to conduct and publish shoddy science. The replication crisis will not abate until these perverse incentives are removed: when publications stop judging the quality of scientific work on the basis of P-values, when tenure committees look beyond publications and impact factors, and when universities and governments realize there's more important things in the world than the latest Shanghai Ranking.

Appendix: calculating posterior predictive value

```
# Operating characteristics of scientific publications for a given
# nominal Type I error, Type II error, fertility of a field of study,
# bias in favor of significant findings and multiple teams testing the
# same associations. A reformulation of the 2005 Ioannidis model.
roc <- function(alpha, beta, prior=0.5, bias=0, n=1) {
  fertility <- prior
  aridity <- 1 - fertility
  false_negative_rate <- beta^n
  true_negative_rate <- (1 - alpha)^n
  false_positive_rate <- 1 - true_negative_rate
  true_positive_rate <- 1 - false_negative_rate
  true_positives <- fertility * true_positive_rate +
    bias * false_negative_rate
  false_positives <- aridity * false_positive_rate +
    bias * true_negative_rate
  true_negatives <- aridity * true_negative_rate
  false_negatives <- fertility * false_negative_rate
  all_positives <- true_positives + false_positives
  all_negatives <- true_negatives + false_negatives
  odds <- true_positives / false_positives
  list(
    true_positives=true_positives,
    false_positives=false_positives,
    true_negatives=true_negatives,
    false_negatives=false_negatives,
    sensitivity=true_positives / fertility,
    miss_rate=false_negatives / fertility,
    specificity=true_negatives / aridity,
    fall_out=false_positives / aridity,
    positive_likelihood_ratio=true_positives / false_positives,
    negative_likelihood_ratio=false_negatives / true_negatives,
    diagnostic_odds_ratio=(true_positives * true_negatives) /
      (false_positives * false_negatives),
    positive_predictive_value=true_positives / all_positives,
    false_omission_rate=false_negatives / all_negatives,
    false_discovery_rate=false_positives / all_positives,
    negative_predictive_value=true_negatives / all_negatives,
    accuracy=true_positives + true_negatives
  )
}
```

References

Altman, D G. 1994. "The scandal of poor medical research." *BMJ* 308 (January): 283–84. doi:10.1136/bmj.308.6941.1438b.

Bakker, Marjan, and Jelte M Wicherts. 2011. "The (mis)reporting of statistical results in psychology journals." *Behavior Research Methods* 43 (3): 666–78. doi:10.3758/s13428-011-0089-5.

Bakker, Marjan, Annette van Dijk, and Jelte M Wicherts. 2012. "The Rules of the Game Called Psychological Science." *Perspectives on Psychological Science* 7 (6): 543–54. doi:10.1177/1745691612459060.

Barnes, Nick. 2011. "Science Code Manifesto." http://sciencecodemanifesto.org/.

Bennett, Craig M, Abigail a Baird, Michael B Miller, and George L Wolford. 2011. "Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction." doi:10.1016/S1053-8119(09)71202-9.

Bissell, Mina. 2013. "Reproducibility: The risks of the replication drive." *Nature* 503 (7476): 333–4. doi:10.1038/nature08217.

Chang, Andrew C, and Phillip Li. 2015. "Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say 'Usually Not'." *Finance and Economics Discussion Series*, Finance and economics discussion series, 2015 (83). Washington: Board of Governors of the Federal Reserve System: 1–26. doi:10.17016/FEDS.2015.083.

Cole, Jonathan R., and Stephen Cole. 1972. "The Ortega Hypothesis." *Science* 178 (4059): 368–75.

Dewald, William G., Jerry G. Thursby, and Richard G. Anderson. 1986. "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project." *The American Economic Review* 76 (4): 587–603. doi:10.2307/1806061.

Doucouliagos, Hristos, and Td Stanley. 2009. "Publication Selection Bias in Minimum Wage Research? A Meta-Regression Analysis." Melbourne: Deakin University Australia. http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8543.2009.00723.x/full.

Fillmore, Kaye Middleton, Tim Stockwell, Tanya Chikritzhs, Alan Bostrom, and William Kerr. 2007. "Moderate Alcohol Use and Reduced Mortality Risk: Systematic Error in Prospective Studies and New Hypotheses." *Annals of Epidemiology* 17: S16–23. doi:10.1016/j.annepidem.2007.01.005.

Fisher, R.a. 1936. "Has Mendel's work been rediscovered?" *Annals of Science* 1 (2): 115–37. doi:10.1080/00033793600200111.

Freedman, David H. 2010. "Lies, Damned Lies, and Medical Science." http://www.theatlantic.com/magazine/archive/2010/11/lies-damned-lies-and-medical-science/308269/.

Gelman, Andrew, and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. doi:10.2277/0521867061.

Gelman, Andrew, and Eric Loken. 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." http://www.stat.columbia.edu/\protect\T1\textbraceleft~\protect\T1\textbracerightgelman/research/unpublished/p\protect\T1\textbraceleft/_\protect\T1\textbracerighthacking.pdf.

———. 2014. "The Statistical Crisis in Science." *American Scientist*, 460–65.

Gilbert, Daniel. 2014. "Psychology's replication police prove to be shameless little bullies." https://twitter.com/dantgilbert/status/470199929626193921.

Gómez, Omar, Natalia Juristo, and Sira Vegas. 2010. "Replication , Reproduction and Re-analysis : Three ways for verifying experimental findings." *Reproduction* 39 (6): 2010–12. doi:10.1080/00207540010028119.

Greenhalgh, Trish. 2012. "Less research is needed." http://blogs.plos.org/speakingofmedicine/2012/06/25/less-research-is-needed/.

Henrich, Joseph, Steven J Heine, and Ara Norenzayan. 2010. "The weirdest people in the world?" *The Behavioral and Brain Sciences* 33: 61–83; discussion 83–135. doi:10.1017/S0140525X0999152X.

Ioannidis, John P a. 2008. "Why most discovered true associations are inflated." *Epidemiology* 19 (5): 640–48. doi:10.1097/EDE.0b013e31818131e7.

Ioannidis, John P. a. 2014. "How to Make More Published Research True." *PLoS Medicine* 11 (10): e1001747. doi:10.1371/journal.pmed.1001747.

John, L. K., G. Loewenstein, and D. Prelec. 2012. "Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling." *Psychological Science* 23 (5): 524–32. doi:10.1177/0956797611430953.

Kessler, Rodger, and Russell E. Glasgow. 2011. "A Proposal to Speed Translation of Healthcare Research Into Practice." *American Journal of Preventive Medicine* 40 (6). Elsevier Inc.: 637–44. doi:10.1016/j.amepre.2011.02.023.

Künzi, Stefan A., Juan Manuel Sarria Toro, Tim Den Hartog, and Peter Chen. 2016. "A Case for Mechanisms." *Israel Journal of Chemistry* 56 (1): 53–61. doi:10.1002/ijch.201500041.

Lang, Tom. 2004. "Twenty Statistical Errors Even YOU Can Find in Biomedical Research Articles." *Croatian Medical Journal* 45 (4): 361–70.

Lehrer, Jonah. 2010. "The Truth Wears Off."

Lexchin, Joel, Lisa A Bero, Benjamin Djulbegovic, and Otavio Clark. 2003. "Pharmaceutical industry sponsorship and research outcome and quality: systematic review." *BMJ* 326 (7400): 1167–70. doi:10.1136/bmj.326.7400.1167.

Lorsch, Jon R, Francis S Collins, and Jennifer Lippincott-schwartz. 2014. "Fixing problems with cell lines: Technologies and policies can improve authentication." *Science* 346: 1452–3.

Mayes, L C, R I Horwitz, and a R Feinstein. 1988. "A collection of 56 topics with contradictory results in case-control research." *International Journal of Epidemiology* 17 (3): 680–85.

Meehl, Paul E. 1978. "Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology." *Journal of Consulting and Clinical Psychology* 46 (113): 806–34.

Meyer, Andrew, Shane Frederick, Terence C. Burnham, Juan D. Guevara Pinto, Ty W. Boyer, Linden J. Ball, Gordon Pennycook, Rakefet Ackerman, Valerie A. Thompson, and Jonathon P. Schuldt. 2015. "Disfluent fonts don't help people solve math problems." *Journal of Experimental Psychology* 144 (2): e16–30. doi:10.1037/xge0000049.

Murray-Rust, Peter, Cameron Neylon, Rufus Pollock, and John Wilbanks. 2010. "Panton Principles." http://pantonprinciples.org/.

Murthy, Vivek H., Harlan M. Krumholz, and Cary P. Gross. 2004. "Participation in Cancer Clinical Trials." *The Journal of the American Medical Association* 291 (22): 2720–26.

Neuroskeptic. 2014. "The Replication Crisis: Response to Lieberman." http://blogs.discovermagazine.com/neuroskeptic/2014/08/31/replication-crisis-response-lieberman/.

Nosek, B. A., Jeffrey R. Spies, and Matt Motyl. 2012. "Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability." Perspectives on Psychological Science. doi:10.1017/CBO9781107415324.004.

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

Ortega y Gassett, José. 1930. *The Revolt of the Masses*.

Perkel, Jeffrey. 2015. "50 years later, is it time to retract a retraction by a Nobel prize-winning author?" http://retractionwatch.com/2015/09/25/five-decades-later-is-it-time-to-retract-a-nobelists-retract

Peterson, David. 2015. "The Baby Factory: Difficult research objects, disciplinary standards, and the production of statistical significance." *Socius* 2: 1–10. doi:10.1177/2378023115625071.

Prasad, Vinay, Andrae Vandross, Caitlin Toomey, Michael Cheung, Jason Rho, Steven Quinn, Satish Jacob Chacko, et al. 2013. "A decade of reversal: An analysis of 146 contradicted medical practices." *Mayo Clinic Proceedings* 88 (8). Elsevier Inc: 790–98. doi:10.1016/j.mayocp.2013.05.012.

Roberts, Seth. 2004. "Self-experimentation as a source of new ideas: ten examples about sleep, mood, health, and weight." *The Behavioral and Brain Sciences* 27 (2): 227–262; discussion 262–87. doi:10.1017/S0140525X04000068.

Romer, By Paul M. 2015. "Mathiness in the Theory of Economic Growth." *American Economic Review* 105 (5): 89–93.

Silberzahn, R., E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, et al. 2014. "Many Analysts, One Dataset." https://osf.io/gvm2z.

Smaldino, Paul E, and Richard Mcelreath. 2016. "The natural selection of bad science." University of California-Davis, Max Planck Institute for Evolutionary Anthropology.

Strasak, Alexander M., Qamruz Zaman, Karl P. Pfeiffer, Georg Göbel, and Hanno Ulmer. 2007. "Statistical errors in medical research - A review of common pitfalls." *Swiss Medical Weekly* 137: 44–49. doi:2007/03/smw-11587.

Thistlethwaite, Donald L, and Donald T Campbell. 1960. "Regression-discontinuity analysis: An alternative to the ex post facto experiment." *Journal of Educational Psychology* 51 (6): 309–17. doi:10.1037/h0044319.

Vul, Edward, Christine Harris, Piotr Winkielman, and Harold Pashler. 2009. "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition." *Perspectives on Psychological Science* 4 (3): 274–90. doi:10.1111/j.1745-6924.2009.01125.x.

Wainer, Howard. 1999. "The Most Dangerous Profession: A Note on Nonsampling Error." *Psychological Methods* 4 (3): 250–56.

Yarkoni, Tal. 2009. "Big Correlations in Little Studies." *Perspectives on Psychological Science* 4 (3): 294–98. doi:10.1111/j.1745-6924.2009.01127.x.

Ziemann, Mark, Yotam Eren, Assam El-Osta, BR Zeeberg, J Riss, DW Kane, KJ Bussey, et al. 2016. "Gene name errors are widespread in the scientific literature." *Genome Biology* 17 (1). Genome Biology: 177. doi:10.1186/s13059-016-1044-7.