# STATISTICAL ERRORS

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

BY REGINA NUZZO

For a brief moment in 2010, Matt Motyl was on the brink of scientific glory: he had discovered that extremists quite literally see the world in black and white.

The results were "plain as day", recalls Motyl, a psychology PhD student at the University of Virginia in Charlottesville. Data from a study of nearly 2,000 people seemed to show that political moderates saw shades of grey more accurately than did either left-wing or right-wing extremists. "The hypothesis was sexy," he says, "and the data provided clear support." The *P* value, a common index for the strength of evidence, was 0.01 — usually interpreted as 'very significant'. Publication in a high-impact journal seemed within Motyl's grasp.

But then reality intervened. Sensitive to controversies over reproducibility, Motyl and his adviser, Brian Nosek, decided to replicate the study. With extra data, the *P* value came out as 0.59 — not even close to the conventional level of significance, 0.05. The effect had disappeared, and with it, Motyl's dreams of youthful fame[1].

It turned out that the problem was not in the data or in Motyl's analyses. It lay in the surprisingly slippery nature of the *P* value, which is neither as reliable nor as objective as most scientists assume. "*P* values are not doing their job, because they can't," says Stephen Ziliak, an economist at Roosevelt University in Chicago, Illinois, and a frequent critic of the way statistics are used.

For many scientists, this is especially worrying in light of the reproducibility concerns. In 2005, epidemiologist John Ioannidis of Stanford University in California suggested that most published findings are false[2]; since then, a string of high-profile replication problems has forced scientists to rethink how they evaluate results.

At the same time, statisticians are looking for better ways of thinking about data, to help scientists to avoid missing important information or acting on false alarms. "Change your statistical philosophy and all of a sudden different things become important," says Steven Goodman, a physician and statistician at Stanford. "Then 'laws' handed down from God are no longer handed down from God. They're actually handed down to us by ourselves, through the methodology we adopt."

DALE EDWIN MURRAY

## OUT OF CONTEXT

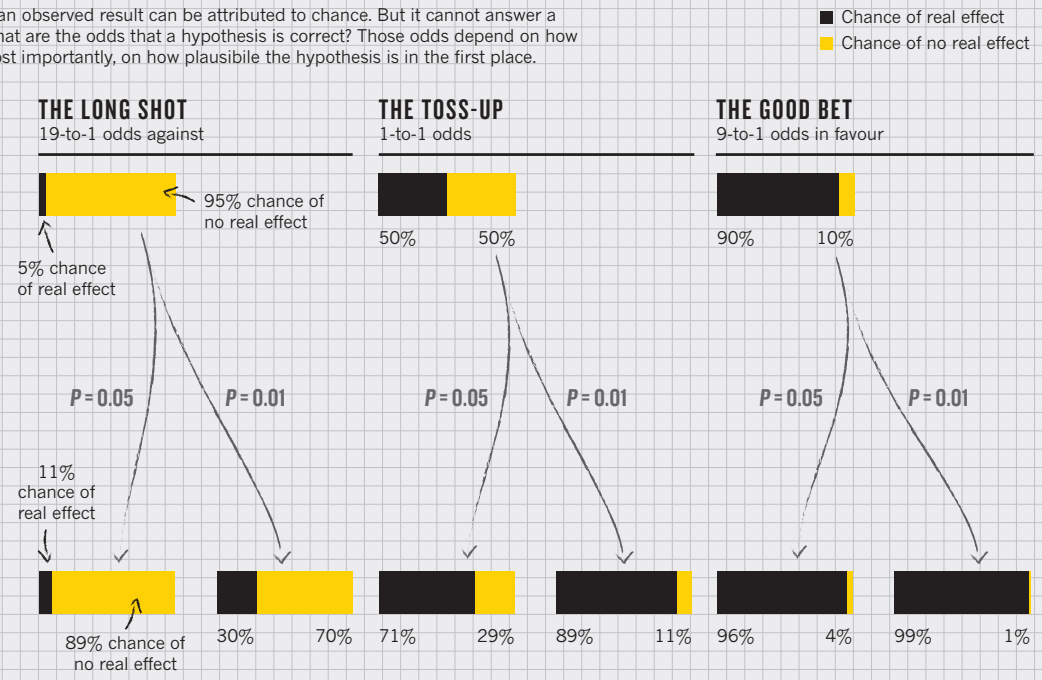
*P* values have always had critics. In their almost nine decades of existence, they have been likened to mosquitoes (annoying and impossible to swat away), the emperor's new clothes (fraught with obvious problems that everyone ignores) and the tool of a "sterile intellectual rake" who ravishes science but leaves it with no progeny[3]. One researcher suggested rechristening the methodology "statistical hypothesis inference testing"[3], presumably for the acronym it would yield.

The irony is that when UK statistician Ronald Fisher introduced the *P* value in the 1920s, he did not mean it to be a definitive test. He intended it simply as an informal way to judge whether evidence was significant in the

R. NUZZO; SOURCE: T. SELLKE *ET AL. AM. STAT.* **55**, 62–71 (2001)

## PROBABLE CAUSE

A *P* value measures whether an observed result can be attributed to chance. But it cannot answer a researcher's real question: what are the odds that a hypothesis is correct? Those odds depend on how strong the result was and, most importantly, on how plausibile the hypothesis is in the first place.




old-fashioned sense: worthy of a second look. The idea was to run an experiment, then see if the results were consistent with what random chance might produce. Researchers would first set up a 'null hypothesis' that they wanted to disprove, such as there being no correlation or no difference between two groups. Next, they would play the devil's advocate and, assuming that this null hypothesis was in fact true, calculate the chances of getting results at least as extreme as what was actually observed. This probability was the *P* value. The smaller it was, suggested Fisher, the greater the likelihood that the straw-man null hypothesis was false.

For all the *P* value's apparent precision, Fisher intended it to be just one part of a fluid, non-numerical process that blended data and background knowledge to lead to scientific conclusions. But it soon got swept into a movement to make evidence-based decision-making as rigorous and objective as possible. This movement was spearheaded in the late 1920s by Fisher's bitter rivals, Polish mathematician Jerzy Neyman and UK statistician Egon Pearson, who introduced an alternative framework for data analysis that included statistical power, false positives, false negatives and many other concepts now familiar from introductory statistics classes. They pointedly left out the *P* value.

But while the rivals feuded — Neyman called some of Fisher's work mathematically "worse than useless"; Fisher called Neyman's approach "childish" and "horrifying [for] intellectual freedom in the west" — other researchers lost patience and began to write statistics manuals for working scientists. And because many of the authors were non-statisticians without a thorough understanding of either approach, they created a hybrid system that crammed Fisher's easy-to-calculate *P* value into Neyman and Pearson's reassuringly rigorous rule-based system. This is when a *P* value of 0.05 became enshrined as 'statistically significant', for example. "The *P* value was never meant to be used the way it's used today," says Goodman.

### WHAT DOES IT ALL MEAN?

One result is an abundance of confusion about what the *P* value means[4]. Consider Motyl's study about political extremists. Most scientists would look at his original *P* value of 0.01 and say that there was just a 1% chance of his result being a false alarm. But they would be wrong. The *P* value cannot say this: all it can do is summarize the data assuming a specific null hypothesis. It cannot work backwards and make statements about the underlying reality. That requires another piece of information: the odds that a real effect was there in the first place. To ignore this would be like waking up with a headache and concluding that you have a rare brain tumour — possible, but so unlikely that it requires a lot more evidence to supersede an everyday explanation such as an allergic reaction. The more implausible the hypothesis — telepathy, aliens, homeopathy — the greater the chance that an exciting finding is a false alarm, no matter what the *P* value is.

**NATURE.COM**
For more on statistics, see:
go.nature.com/xlj9lr

These are sticky concepts, but some statisticians have tried to provide general rule-of-thumb conversions (see 'Probable cause'). According to one widely used calculation[5], a *P* value of 0.01 corresponds to a false-alarm probability of at least 11%, depending on the underlying probability that there is a true effect; a *P* value of 0.05 raises that chance to at least 29%. So Motyl's finding had a greater than one in ten chance of being a false alarm. Likewise, the probability of replicating his original result was not 99%, as most would assume, but something closer to 73% — or only 50%, if he wanted another 'very significant' result[6,7]. In other words, his inability to replicate the result was about as surprising as if he had called heads on a coin toss and it had come up tails.

Critics also bemoan the way that *P* values can encourage muddled thinking. A prime example is their tendency to deflect attention from the actual size of an effect. Last year, for example, a study of more than 19,000 people showed[8] that those who meet their spouses online are less likely to divorce ($p < 0.002$) and more likely to have high marital satisfaction ($p < 0.001$) than those who meet offline (see *Nature* http://doi.org/rcg; 2013). That might have sounded impressive, but the effects were actually tiny: meeting online nudged the divorce rate from 7.67% down to 5.96%, and barely budged happiness from 5.48 to 5.64 on a 7-point scale. To pounce on tiny *P* values and ignore the larger question is to fall prey to the "seductive certainty of significance", says Geoff Cumming, an emeritus psychologist at La Trobe University in Melbourne, Australia. But significance is no indicator of practical relevance, he says: "We should be asking,

'How much of an effect is there?', not 'Is there an effect?'"

Perhaps the worst fallacy is the kind of self-deception for which psychologist Uri Simonsohn of the University of Pennsylvania and his colleagues have popularized the term *P*-hacking; it is also known as data-dredging, snooping, fishing, significance-chasing and double-dipping. "*P*-hacking," says Simonsohn, "is trying multiple things until you get the desired result" — even unconsciously. It may be the first statistical term to rate a definition in the online Urban Dictionary, where the usage examples are telling: "That finding seems to have been obtained through *p*-hacking, the authors dropped one of the conditions so that the overall *p*-value would be less than .05", and "She is a *p*-hacker, she always monitors data while it is being collected."

Such practices have the effect of turning discoveries from exploratory studies — which should be treated with scepticism — into what look like sound confirmations but vanish on replication. Simonsohn's simulations have shown[9] that changes in a few data-analysis decisions can increase the false-positive rate in a single study to 60%. *P*-hacking is especially likely, he says, in today's environment of studies that chase small effects hidden in noisy data. It is tough to pin down how widespread the problem is, but Simonsohn has the sense that it is serious. In an analysis[10], he found evidence that many published psychology papers report *P* values that cluster suspiciously around 0.05, just as would be expected if researchers fished for significant *P* values until they found one.

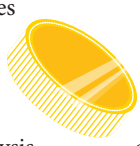
## NUMBERS GAME

Despite the criticisms, reform has been slow. "The basic framework of statistics has been virtually unchanged since Fisher, Neyman and Pearson introduced it," says Goodman. John Campbell, a psychologist now at the University of Minnesota in Minneapolis, bemoaned the issue in 1982, when he was editor of the *Journal of Applied Psychology*: "It is almost impossible to drag authors away from their *p*-values, and the more zeroes after the decimal point, the harder people cling to them"[11]. In 1989, when Kenneth Rothman of Boston University in Massachusetts started the journal *Epidemiology*, he did his best to discourage *P* values in its pages. But he left the journal in 2001, and *P* values have since made a resurgence.

Ioannidis is currently mining the PubMed database for insights into how authors across many fields are using *P* values and other statistical evidence. "A cursory look at a sample of recently published papers," he says, "is convincing that *P* values are still very, very popular."

Any reform would need to sweep through an entrenched culture. It would have to change how statistics is taught, how data analysis is done and how results are reported and interpreted. But at least researchers are admitting that they have a problem, says Goodman. "The wake-up call is that so many of our published findings are not true." Work by researchers such as Ioannidis shows the link between theoretical statistical complaints and actual difficulties, says Goodman. "The problems that statisticians have predicted are exactly what we're now seeing. We just don't yet have all the fixes."

> "THE *P* VALUE WAS NEVER MEANT TO BE USED THE WAY IT'S USED TODAY."



Statisticians have pointed to a number of measures that might help. To avoid the trap of thinking about results as significant or not significant, for example, Cumming thinks that researchers should always report effect sizes and confidence intervals. These convey what a *P* value does not: the magnitude and relative importance of an effect.

Many statisticians also advocate replacing the *P* value with methods that take advantage of Bayes' rule: an eighteenth-century theorem that describes how to think about probability as the plausibility of an outcome, rather than as the potential frequency of that outcome. This entails a certain subjectivity — something that the statistical pioneers were trying to avoid. But the Bayesian framework makes it comparatively easy for observers to incorporate what they know about the world into their conclusions, and to calculate how probabilities change as new evidence arises.

Others argue for a more ecumenical approach, encouraging researchers to try multiple methods on the same data set. Stephen Senn, a statistician at the Centre for Public Health Research in Luxembourg City, likens this to using a floor-cleaning robot that cannot find its own way out of a corner: any data-analysis method will eventually hit a wall, and some common sense will be needed to get the process moving again. If the various methods come up with different answers, he says, "that's a suggestion to be more creative and try to find out why", which should lead to a better understanding of the underlying reality.

Simonsohn argues that one of the strongest protections for scientists is to admit everything. He encourages authors to brand their papers '*P*-certified, not *P*-hacked' by including the words: "We report how we determined our sample size, all data exclusions (if any), all manipulations and all measures in the study." This disclosure will, he hopes, discourage *P*-hacking, or at least alert readers to any shenanigans and allow them to judge accordingly.

A related idea that is garnering attention is two-stage analysis, or 'preregistered replication', says political scientist and statistician Andrew Gelman of Columbia University in New York City. In this approach, exploratory and confirmatory analyses are approached differently and clearly labelled. Instead of doing four separate small studies and reporting the results in one paper, for instance, researchers would first do two small exploratory studies and gather potentially interesting findings without worrying too much about false alarms. Then, on the basis of these results, the authors would decide exactly how they planned to confirm the findings, and would publicly preregister their intentions in a database such as the Open Science Framework (https://osf.io). They would then conduct the replication studies and publish the results alongside those of the exploratory studies. This approach allows for freedom and flexibility in analyses, says Gelman, while providing enough rigour to reduce the number of false alarms being published.

More broadly, researchers need to realize the limits of conventional statistics, Goodman says. They should instead bring into their analysis elements of scientific judgement about the plausibility of a hypothesis and study limitations that are normally banished to the discussion section: results of identical or similar experiments, proposed mechanisms, clinical knowledge and so on. Statistician Richard Royall of Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland, said that there are three questions a scientist might want to ask after a study: 'What is the evidence?' 'What should I believe?' and 'What should I do?' One method cannot answer all these questions, Goodman says: "The numbers are where the scientific discussion should start, not end." ■ SEE EDITORIAL P. 131

**Regina Nuzzo** *is a freelance writer and an associate professor of statistics at Gallaudet University in Washington DC.*


1. Nosek, B. A., Spies, J. R. & Motyl, M. *Perspect. Psychol. Sci.* **7,** 615–631 (2012).
2. Ioannidis, J. P. A. *PLoS Med.* **2,** e124 (2005).
3. Lambdin, C. *Theory Psychol.* **22,** 67–90 (2012).
4. Goodman, S. N. *Ann. Internal Med.* **130,** 995–1004 (1999).
5. Goodman, S. N. *Epidemiology* **12,** 295–297 (2001).
6. Goodman, S. N. *Stat. Med.* **11,** 875–879 (1992).
7. Gorroochurn, P., Hodge, S. E., Heiman, G. A., Durner, M. & Greenberg, D. A. *Genet. Med.* **9,** 325–321 (2007).
8. Cacioppo, J. T., Cacioppo, S., Gonzagab, G. C., Ogburn, E. L. & VanderWeele, T. J. *Proc. Natl Acad. Sci. USA* **110,** 10135–10140 (2013).
9. Simmons, J. P., Nelson, L. D. & Simonsohn, U. *Psychol. Sci.* **22,** 1359–1366 (2011).
10. Simonsohn, U., Nelson, L. D. & Simmons, J. P. *J. Exp. Psychol.* http://dx.doi.org/10.1037/a0033242 (2013).
11. Campbell, J. P. *J. Appl. Psych.* **67,** 691–700 (1982).