NATURE | NEWS FEATURE

Scientific method: Statistical errors

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

Regina Nuzzo

12 February 2014

DALE EDWIN MURRAY

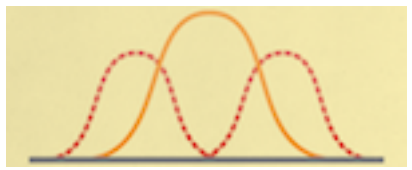
For a brief moment in 2010, Matt Motyl was on the brink of scientific glory: he had discovered that extremists quite literally see the world in black and white.

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.



Statisticians issue warning over misuse of *P* values

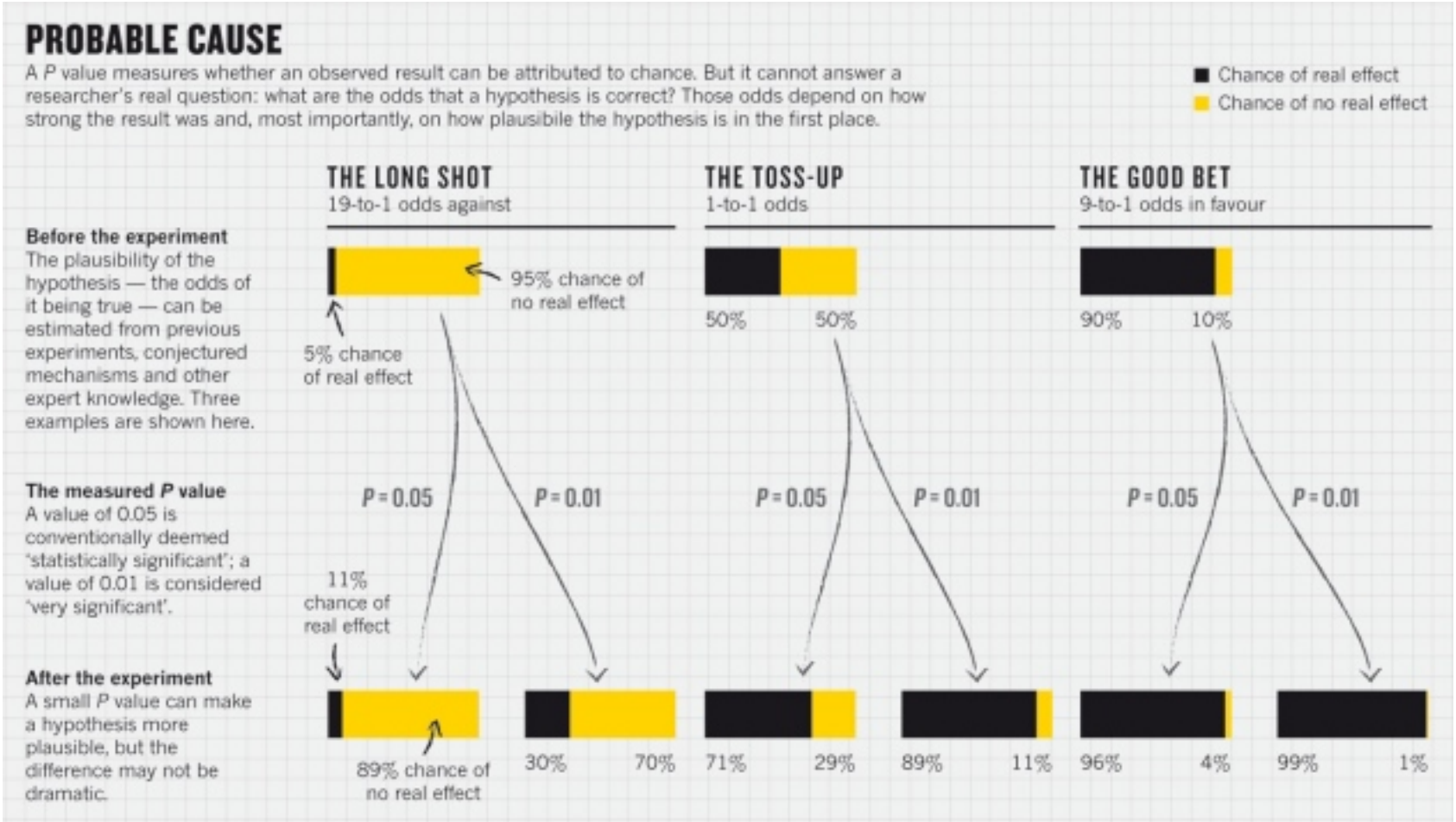
For many scientists, this is especially worrying in light of the reproducibility concerns. In 2005, epidemiologist John Ioannidis of Stanford University in California suggested that most published findings are false[2]; since then, a string of high-profile replication problems has forced scientists to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

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 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.



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

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.

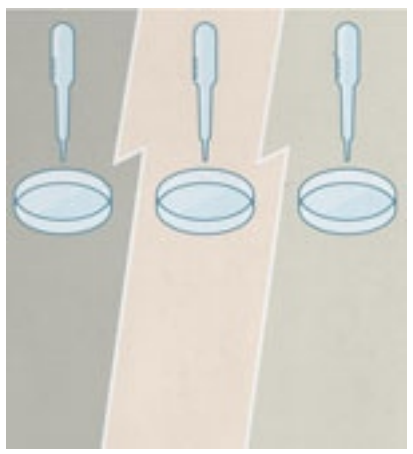
But while the rivals feuded — Neyman called some of Fisher's work mathematically “worse than useless”; Fisher called Neyman's approach “childish” and “horrifying [for] intellectual freedom in the west” — other researchers lost patience and began to write statistics manuals for working scientists. And 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.

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.



***Nature* special:** Challenges in irreproducible research

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.”

“The *P* value was never meant to be used the way it's used today.”

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.”

Statistics: *P* values are just the tip of the iceberg

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.”

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.

Related stories

- Number crunch
- Policy: NIH plans to enhance reproducibility
- Weak statistical standards implicated in scientific irreproducibility

More related stories

Related stories

- Number crunch
- Policy: NIH plans to enhance reproducibility
- Weak statistical standards implicated in scientific irreproducibility

More related stories

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.”

Nature **506,** 150–152 (13 February 2014) doi:10.1038/506150a

See Editorial *page 131*

References

1. Nosek, B. A., Spies, J. R. & Motyl, M. *Perspect. Psychol. Sci.* **7**, 615–631 (2012).
   Show context Article
   Show context

2. Ioannidis, J. P. A. *PLoS Med.* **2**, e124 (2005).
   Show context Article PubMed
   Show context

3. Lambdin, C. *Theory Psychol.* **22**, 67–90 (2012).
   Show context Article
   Show context

4. Goodman, S. N. *Ann. Internal Med.* **130**, 995–1004 (1999).
   Show context Article ISI ChemPort
   Show context

5. Goodman, S. N. *Epidemiology* **12**, 295–297 (2001).
   Show context Article PubMed ISI ChemPort
   Show context

6. Goodman, S. N. *Stat. Med.* **11**, 875–879 (1992).
   Show context Article PubMed ISI ChemPort
   Show context

7. Gorroochurn, P., Hodge, S. E., Heiman, G. A., Durner, M. & Greenberg, D. A. *Genet. Med.* **9**, 325–321 (2007).
   Show context Article PubMed ISI
   Show context

8. Cacioppo, J. T., Cacioppo, S., Gonzagab, G. C., Ogburn, E. L. & VanderWeele, T. J. *Proc. Natl Acad. Sci. USA* **110**, 10135–10140 (2013).
   Show context Article PubMed ChemPort
   Show context

9. Simmons, J. P., Nelson, L. D. & Simonsohn, U. *Psychol. Sci.* **22**, 1359–1366 (2011).
Show context Article PubMed
Show context

10. Simonsohn, U., Nelson, L. D. & Simmons, J. P. *J. Exp. Psychol.* http://dx.doi.org/10.1037/a0033242 (2013).
Show context
Show context

11. Campbell, J. P. *J. Appl. Psych.* **67**, 691–700 (1982).
Show context Article
Show context

Related stories and links

From nature.com

- **Number crunch**
  12 February 2014
- **Policy: NIH plans to enhance reproducibility**
  27 January 2014
- **Weak statistical standards implicated in scientific irreproducibility**
  11 November 2013
- **Uncertainty on trial**
  02 October 2013
- **Matters of significance**
  29 August 2013
- **Announcement: Reducing our irreproducibility**
  24 April 2013
- **Replication studies: Bad copy**
  16 May 2012
- **Blog post: Let's give statistics the attention it deserves in biological research**
- **Blog post: Statistics is the sexy in science**
- ***Nature* special: Challenges in irreproducible research**

From elsewhere

- ***Psychological Science* tutorial on alternatives to the *P* value**
- **The BUGS (Bayesian inference Using Gibbs Sampling) Project**
- **Bayesian Cognitive Modelling: A Practical Course**

Author information

Affiliations

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

37 comments

Subscribe to comments

Guest • 2014-03-06 07:31 PM

Can anyone help me understand the "probable cause" picture of the paper? I admit that I am lost. 1. What is the meaning of "odds of hypothesis"? A hypothesis can be Right, or wrong. What is odds of it mean? If we know the odds, do we still need to know pValue? 2. How can I get the number in the picture: 【with 1 to 19 odds of hypothesis】 + 【pValue = 0.05】 --> odds become 11%vs. 89%. Thanks.

Charles Green • 2014-02-21 04:21 PM

In statistics P values although called confidence values are not measures of accuracy. They are a statement of what the distribution of results that can be made when the test is replicated. If one uses them correctly they are a way to select future projects. A test with a very high P could be replicated with a smaller sample thus saving in the cost of replication. If the replication reults fails then another test is needed with a larger sample. Only after many tests can conclusions be made, Use of different tests does not improve the situation. I tested magazine vs TV spending with three tests all with P = 98 to 99.4. All were wrong due to an error in the underling data. I assume continuous data when is was discrete data. More than three quarters of my education in statistics dealt with errors in design and consideration of the underling data.

David Clarke • 2014-02-18 05:40 PM

I'm probably stupid but how am I supposed to assess the probability of my hypothesis into a "long shot" at 19 to 1;" a toss up" at 1 to 1; a "good bet" at 9 to1?

Barry Cohen • 2014-02-15 10:37 PM

Regina: In your figure labeled Probable Cause, you cite Sellke et al. (2001) as the source. I read that article, and could not see how you derived your figure from that article. Specifically, Sellke et al. (2001) needed to posit a value for xi, which as they defined it corresponds to what psychologists usually call delta (the basis for determining the power of a t test), in order to find the odds for different p values (see their Figure 2). It is not clear what value for xi you are using, though from your results, I would guess it is about .75, which would lead to an average power for real effects of about .12 for a ,05, two-tailed significance test. More generally, your figure relies on the chance of a "real effect" in each case, but are you defining a real effect as anything other than exactly zero? Doesn't it matter how large, on average, these real effects are? In other words: We know the probability of obtaining a p value of .05 or smaller when there is no real effect, but doesn't the probability of obtaining a p value of .05 or smaller when there IS a real effect depend on the size of the real effect (for a given sample size)? What assumption are you making there?

deborah mayo • 2014-02-15 03:53 AM

Regina: There is a citation from Neyman in this article but I don’t see the reference. I’d be grateful if you provided it. I'm fairly sure it's entirely taken out of context. "worse than useless" is a technical term. Poor Motyl “was on the brink of scientific glory” by means of shoddy statistics! Glory, I tell you, glory. Maybe he should be given a medal for not rushing into print as imagined by those who view science as an unthinking screening effort. Of course, it can’t be that he’s fallen into the dumbest of dumb misuses of p-values. It cannot be that he’s exploiting fraudulent uses of statistics. No, this author blames the statistical tools for his highly questionable exploitation of p-values. The truth is that the only shades of grey here is the fact that misuse of statistics differs only in degree from out and out fraud. Any inference is questionable if the researcher cannot show that flaws in his or her analysis would have been detected with high probability. Fields (like this one) that regularly spin out results without showing they have worked hard—or have even tried—to subject their own analysis to severe scrutiny are pseudoscientific. Pseudoscientists are frauds and should be treated as such. Science writers who exploit the fashion of dumping on p-values only give them excuses.

David Lovell • 2014-02-15 02:52 AM

Thanks for raising awareness on this Regina. I would be very interested to see a followup article or comments about False Discovery Rate (FDR) procedures used in situations where multiple comparisons are made. I'm not a statistician; my intuition is that FDR further masks the shortcomings of Null Hypothesis Significance Testing. As far as I understand it, FDR amounts to setting a more stringent p-value at which one regards data to be statistically significant. FDR procedures abound in bioinformatics and other areas of modern quantitative bioscience where measurements are plentiful. Is my skepticism of FDR warranted?

Ben Wise • 2014-02-14 03:11 PM

Morris DeGroot at CMU made this point way back in the 1980's. P-Values and "significance" measure the probability of data given the hypothesis, not the probability of the hypothesis given the data. This is exactly backwards, as (to quote DeGroot) "I already know the probability of the data: 1, because I just observed it!". It is easy to find actual cases where P(D|H) = 0.99 and P(-D|-H) = 0.99, but P(H|D)=0.01. In English, the hypothesis has a high P-Value and is extremely "significant" but is almost certainly wrong. No wonder so many medical studies are overturned by later studies: they were highly significant, but not very probable. A comment was made below that Bayesian methods must be used carefully. I just repeat DeGroot's response (which I heard when he was confronted with the same criticism): it is better to do the right calculation carefully than to do the wrong one easily.

deborah mayo • 2014-02-15 03:42 AM

p-values are NOT likelihoods, however, they permit computations that Bayesian likelihoods alone cannot. They allow evaluating the probability that the testing procedure would have resulted in a less impressive departure (from the null) under the assumption the null is true, and also under the assumption of varying discrepancies from the null. It's a small part of the panoply of methods that use error probabilities. Guess what? Bayesians are the ones who only use likelihoods conditional on the observed value! So no error probabilistic assessments are possible. Oh, but there's a prior you say? No error control there either--just what someone believes, and very few scientists want to mix their prior beliefs into the study. The point of the research is to test claims--not beg the question by imputing prior beliefs!

Huw Llewelyn • 2014-02-14 12:59 PM

The common sense question faced by those who interpret data in a public way (e.g. doctors, engineers and research scientists) is “What do I predict from this observation?” The answer can be that (1) the observation will probably not be replicated and is probably spurious (2) that it suggests a simple prediction or a prediction linked to possible narrative or mathematical models (e.g. a diagnosis, a working engineering model, a general scientific hypothesis, theory or law) that will in turn make many other useful predictions. The first hurdle to overcome is whether or not the observation will probably be replicated. This can done by showing that all the possible reasons for non-replication are improbable. One of these causes of non-replication is that the number of observations is too low (if the ‘observation’ is made up of a number of different observations). This is where statistical significance testing comes in (successfully repeating the entire set of observations independently makes the probability of further non-replication due to this reason very low of course). There are many other reasons for non-replication to be considered eg. poor documentation or vague writing, dishonesty, poor methodology, contradictory observations by someone

else, etc, etc. In order for the probability of replication to be high, the probabilities of all these causes of non-replication also have to be low. The reasoning process inevitably has to be subjective but there is a formal basis for it in probability theory (that incorporates Bayes rule) to guide us (see also Llewelyn H. Reasoning in medicine and science. OUP blog, September 2013).

Mark Brewer • 2014-02-13 04:51 PM

I'm glad this article has provoked discussion. What I find surprising is the fact that the "Probable Cause" infographic presents a beautiful argument for a Bayesian approach, without actually saying so, or even realising it is doing so.

H T • 2014-02-13 08:30 PM

I don't think that replacing frequentist with the Bayesian approach is the answer, nor is that the message of the article. Bayesian statistics can demonstrate the shortcomings of frequentist statistics very well in some situations (like in the infographic), but also requires great care to handle. It would be naive to think that researchers who abuse p-values would not do the same to priors and model specification.

deborah mayo • 2014-02-15 03:48 AM

They'd necessarily do worse and fraud-busting would be dead. Why? All criticisms turn on being able to evaluate error probabilities (even if only informally), e.g., showing the study like Motyl's has done practically nothing to prevent the worst kind of abuse and fraudulent use of statistics. I agree it's a nice picture, but the article is misleading in dozens of ways. Simonsohn is interviewed but the author doesn't bother to mention that he points out how Bayesian statistics only introduces more flexibility into the analysis. It's quite a biased article, which really defeats the purpose.

Paul Hayes • 2014-02-15 07:11 AM

Simonsohn was wrong on that point: http://doingbayesiandataanalysis.blogspot.co.uk/2011/10/false-conclusions-in-false-positive.html

Bob O Hara • 2014-02-14 10:10 AM

Indeed. In fact, we could just replace the abuse of p-values with the abuse of Bayes factors and Bayesian p-values.

John Vidale • 2014-02-13 04:50 PM

Very good article, but it misses the mark in two ways, IMO. This is really a primer for the public to take science headlines with a grain of salt. First, the underlying reason for misuse of statistics is the natural optimism of scientists - we think we will find what no one has previously found, and that our experiment was the one sensible way to explore the problem. That is, we overestimate the a priori likelihood that our solution was right, and we underestimate the amount of fiddling we (and others) have done leading up to our latest result. Second, scientists vary greatly in their familiarity with statistics and basic common sense - they always have and always will. Requiring tedious publication of all data, studies through sequential publications, application of multiple statistical tests in each study may ameliorate some problems, but will impede many others. As has been true forever, scientists looking at data need to understand statistical tools to use them right, as asserted. I also doubt it is a new phenomenon that scientists recognize the fallibility of the latest, hottest study. I recall several of their editors telling me that many (most) Science and Nature papers are incorrect.

Allen Bryant • 2014-02-13 04:04 PM

Having taught Statistics for a number of years, the issue of what the value of the P-Value is, isn't really that important. What is important is a properly structured hypothesis test. The P-value is a measure of what degree of confidence we wish to know something might be true should the hypothesis test prove that we can reject our null hypothesis in favor of the alternative hypothesis. If results are not reproducible, perhaps your hypothesis can't be rejected and you need to completely reconsider your hypothesis.

Ben Wise • 2014-02-14 03:23 PM

Morris DeGroot had a comment on this line of reasoning back in the 1980's.The only way to judge when it is "properly structured" is by comparing it to Bayesian reasoning, that is, to make sure a high P(D|H) occurs only when P(H|D) is high. But if you have to do the exact Bayesian analysis in order to make sure the p-value heuristic (as he called it) is doing the right thing, why not just keep the Bayesian analysis and skip the heuristic? This is the approach I taught my graduate students; I'd recommend DeGroot's work as a very balanced approach that combines both practical commonsense (you need p-values to get published, and they are easier to calculate) and theoretical rigor (when can they be relied upon).

Thomas Dent • 2014-02-13 01:43 PM

The author should take her own advice and show some valid statistics to back up sweeping, off-the-cuff claims about what 'most scientists' or 'most researchers' might or might not do. "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." Define 'most scientists'. What evidence is there for this claim? What is your sample size: how many scientists have been objectively tested on what they would say in these circumstances? What is the effect size: how many did in fact say the thing you claimed? Is it a fair sample of all scientists, or are some disciplines or some levels of seniority or some nationalities over- or under-represented? How can we be sure the author doesn't cherry-pick conversations where someone appears not to understand p-values? This is not a joke; it's a very serious point. There are scientists who do understand p-values and put considerable effort into using them correctly; the particle physics community is one example. Articles like this one which blame 'the p-value' for everything people do wrong with statistics, as if the method itself - rather than the uses to which it is put - was somehow the root of all evil, amount to unfairly smearing results obtained by a correct and rigorous usage of p-values, i.e. with blind data analysis and honestly-accounted-for trials factors (the 'look-elsewhere effect'). To say that muddled thinking and self-deception are *caused* by use of p-values is absurd; people who are prone to muddled thinking and self-deception will carry on being so regardless of the statistical framework. You might as well claim that abuse of significant figures is caused by the use of the decimal point.

Ben Wise • 2014-02-14 03:16 PM

I think the driving factor is not "most scientists" but the reviewers of major journals. It is essentially impossible to get a paper published without a high p-value, which drives everyone else to design their work to generate high p-values, whether they agree with them or not. "Publish or perish".

Abhay Sharma • 2014-02-13 11:01 AM

Over-selection and over-reporting of false positive results are increasingly plaguing the published research with an alarming rate (Nature 485, 149; 2012). In the current practice, such reporting is considered as honest errors not amounting to misconduct (Nature 485, 137; 2012). However, since intention is the core of misconduct, one may very well argue that reporting of results with systematic positive bias should also be placed under the ambit of misconduct. Scientific community and policy makers need to consider this tough option in the overall interest of science. [This is a part of the comments made earlier (http://www.nature.com/news/policy-nih-plans-to-enhance-reproducibility-1.14586)].

Mark Alexander • 2014-02-13 10:34 AM

In spite of the article's comment that p-values from research into phenomena like telepathy are likely to be "false alarms", in point of fact, some of the most significant p-values in any area of research come from precisely this direction. I think of the ganzfeld studies, which have produced mind-boggling probability values on the order of 10^-18. Goodman's formula doesn't do such a value any damage. The comment in the article – especially insofar as it groups such research with "homeopathy and aliens" as a label of derision – reflects a widespread but regrettable lack of knowledge about what has been achieved in this area.

Mark Brewer • 2014-02-13 05:09 PM

I'm afraid that p-values are always going to be flawed (quoting numbers of the order of 10^-18 just smacks of desperation) when the basic underlying "science" is flawed.

Mark Alexander • 2014-02-14 07:08 AM

Nevertheless, those p-values are objectively present. On the one hand, findings in this area are routinely dismissed because 'extraordinary claims require extraordinary evidence'. But then, when p-values such as these are presented, it only 'smacks of desperation'. What kind of "science" is that?

Paul Hayes • 2014-02-15 02:12 AM

It's good science. Interpreting a probability of only 10^-18 that your telepathy experiment results were caused by 'chance' as evidence that they were caused by telepathy, given what is already known from relevant previous results[1], is bad science. Very bad science. [1] http://blogs.discovermagazine.com/cosmicvariance/2008/02/18/telekinesis-and-quantum-field-theory/#.Uv7L8XWbBiB http://www-biba.inrialpes.fr/Jaynes/cc05e.pdf

Mark Alexander • 2014-02-18 09:43 AM

Actually acquaint yourself with the literature before you pass judgement. Or at least a good chunk of it. In spite of making reference to the 'quality of previous results', it's clear that you haven't seen them -- and that's precisely the point I'm making. Your first link rules out these phenomena as a matter of principle. Your second link doesn't impact on the ganzfeld results I'm citing in any way whatsoever. It's simply a secondhand discussion of why all such results must surely be mistaken. In other words, all you've done is cite two

claims why such results should be dismissed out of hand. And no, that's not good science.

Chandrika B-Rao • 2014-02-13 09:15 AM

Misinterpretation of p-value seldom comes from biostatisticians. For the biologist, after having done many months of work generating the data, the final statistical analysis seems to be a minor matter, not deserving much attention. Many biologists prefer to do their own statistical analysis rather than involving another person for this 'minor' work. The cautious or conservative interpretation of data provided by the statistician sometimes doesn't go down well with biologists wanting definitive conclusions from their hard work. Many non-statistical journals publish p-values without any associated information like : what was the sample size, what statistical test was done, what hypothesis was being tested, what was the power of the test etc., resulting in sloppy statistical analysis and sloppier reporting, making true the saying, "there are lies, damn lies and statistics". Non-statistician referees seldom ask adequate questions about the statistical methods used and analysis done. When an article with substandard statistical work gets accepted for publication in a good biology journal, the biologist no longer feels the need to talk with a statistician. Journals with word limits also unconsciously encourage cutting corners on statistical analysis reporting. I am very surprised by the heat of "We must collaborate with statisticians, not let them decide what is good for patients." coming from Giovanni Codacci-Pisanelli. Are statisticians irrational, unforgiving ogres, not caring for the good of the patients??

Peter Gerard Beninger • 2014-02-13 09:13 AM

I'm happy to see that this issue is beginning to emerge on the radar of more scientists, notably the readers of Nature and Science. However, the ones who persist in the worst, and most common, misuse of frequentist statistics rarely, if ever, read these journals, and seem equally oblivious to the vast number of publications, in all fields, which make the same points. Their papers constitute the majority of many mainstream specialty journals. I have taken this up directly with senior editors, who simply reply that they do their jobs by relying on reviewers for quality control. My suggestion that each reputable journal should have a full-time statistician on board to review the procedures used in all 'provisionally-accepted' papers, as well as for all statistically-contested papers (as is the pracice in the best medical journals), has so far fallen on deaf ears for all of the journals in my own research field. The situation will only improve if we push the publishers hard enough.

Jane Public • 2014-02-13 08:06 AM

Brian: I think I can answer at least part of this for you. I had a discussion about this with someone just the other day. Although I don't think he got the point. Anyway, let's use a purely hypothetical example to illustrate the point. Let's say someone decides to study the IQs of the students at universities, and test

correlations between IQ and various other factors. IQ tests are administered to 10,000 students, and the results more-or-less follow the expected normal distribution, with a measurement error of +/- 2 points. So now they start comparing with other measured factors. And they find something very surprising: there is a very strong inverse correlation (P = 0.001... they're VERY sure of this), between nipple size and IQ! (Hey... I've seen much sillier things in studies before.) So... they go on the evening news with their startling discovery. But what does this mean? Well, if you were to look at the effect size, it turns out that people with aureolae that measured 1.5cm across had an IQ that was 0.02 points higher than students whose aureolae were 6cm across. So the effect -- 0.02 IQ points -- is very, very small. Even though there is strong statistical evidence of a correlation, the actual effect is so small as not to really matter. Even worse; the effect is smaller than the measurement error for their IQ tests. (Pretty much invalidating their work, if anybody bothered to check.) So this "statistical significance", while very strong, has about zero "significance" in the real world. Although this is a somewhat exaggerated example, this kind of thing is not that unusual. As I say, I was trying to explain this to someone the other day, about exactly this kind of announcement: a reported strong correlation, but the effect size was tiny, and essentially buried in the small print.

Brian crawford • 2014-02-12 10:21 PM

I liked the article but have a quick question. When the author says "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?'”" How do you decide what level of effect is appropriate to report? Is it just subjective decision? For example, would a enrichment of particular set of genes of 54% in one sample compared with 46% in another be enough of an effect? Even if they are very significant?

Bob O Hara • 2014-02-13 09:08 AM

You're asking the right question, but I (as a statistician) can't answer it for you: it's biological judgement. And this is a good thing. After all, you are doing science, not statistics, so your judgement of what is 'significant' should be based on science. I think if we all used effect sizes and confidence intervals, a hidden benefit would be that it would make us think more about the actual scientific relevance of our results.

Ben Wise • 2014-02-14 03:47 PM

There are (at least) two different senses of the word "significant" being mixed together. One is "having a high P(Data|Hypothesis)" and the other is "reasonable to act upon". The IQ

correlation example above is one that has high P(D|H) but has no practical implications for anything anyone would decide to do or not. For example, it does not help one decide whether to release a drug onto the market or not. The second sense of "significance" leads one directly into the realm of decision theory and the actual cost of (e.g.) Type I and Type II errors. But decisions invariably involve the weighing of costs and benefits, which fall differently on different groups, and so involve a lot of debates that classical statisticians try to avoid. Again, a nice compromise is the report actual probability values, like P(H|D),. They make a nice "decoupled interface", in that they can be taken either as a summary statement of how strongly the hypothesis is indicated, or as the start of a decision-theoretic analysis.

Luar Moreno-Alvarez • 2014-02-12 09:47 PM

Although this is a very good article, it is limited in approach and references to social and biological subjects. Perhaps, in order to achieve a more general and technical view of this important issue, a deeper review of works from Statistics journals would be desirable. The paper 'P-Value Precision and Reproducibility' of Boos & Stefanski in 'The American Statistician' (2011), for example, could be useful to the enrichment of this discussion.

Giovanni Codacci-Pisanelli • 2014-02-12 08:16 PM

This article is a long awaited reminder of what statistics can do, and of what they cannot do! The Peanuts strips about statisticians are funnier...but not as extensive. In clinical oncology p-values often are the aim of the clinical trial. Still, just using a computer spreadsheet programme it is easy to prove that if you enter enough values you p will become significant even when the difference between two means is minimal. Unfortunately "statistically significant" is considered a synonym of "true", but very often it rather seems to mean "clinically irrelevant" (a trial with a 0.7 weeks difference in progression-free survival of patients with advanced cancer had a significant p). But what is even more disturbing is the number of retracted papers (for example on gene signatures) based on the "statistical evaluation" of results...that could not be reproduced. We must collaborate with statisticians, not let them decide what is good for patients.

deborah mayo • 2014-02-19 04:10 PM

A long-awaited repeat of a cookbook article that follows the recipe of so many "front page news" stat exposes in every purported science mag for years. and just as shallow... Poor Motyl, it's so hard to do good science...

Bob O Hara • 2014-02-13 09:05 AM

Don't blame the statisticians! We've been banging on about this for years, but p-values are just so entrenched in the way a lot of scientists think science should be done. The problem is one of inertia: p-values are accepted as standard, so scientists teach their students that this is how things should be done, so that's all they learn.

Steve Schwartz • 2014-02-12 07:36 PM

A critical aspect of p-values, and hypothesis testing under a frequentist framework more generally, that is not addressed by this column, is that these techniques were developed and originally implemented in the setting where random error is the only (or at least dominant) reason why a particular study might not yield the correct answer. In the setting of non-experimental research, where the "exposure" or study condition has not been randomly assigned, whether the study yields the true relationship has far more to do with biases in measurement of key variables, in the selection of study subjects, and in accounting for confounding or similar relationships among variables. Neither p-values nor confidence intervals measure these features of a study. But because p-values and confidence intervals are easy to produce, and measures of many non-random biases are not easy to make, the statistical indices have become the coin of the scientific realm in research designs such as observational studies where they were not originally intended to be used.

H T • 2014-02-12 06:16 PM

This article serves as a welcome reminder of the many fallacies that still bely modern scientific research. However, I feel it could use a bit more balance in the context of reproducibility. In the paragraph where Motyl's p-value of 0.01 is revisited, the probability of replicating this result at a 'significant' ($p < 0.05$) level is cited as only 73%. It would have been useful to clarify that this is 'less than expected' becasuse: (1) there is a chance that the initial result is a false-positive; and (2) there is an appreciable probability that repeating the test may produce a false negative result even if the hypothesis is true. Hence, this is not just about the limits of p-value, but also about the limits of relying on a single experimental replicate.

See other News & Comment articles from *Nature*

Nature ISSN 0028-0836 EISSN 1476-4687