DALE EDWIN MURRAY

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.

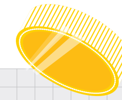
At the same time, statisticians are looking for better ways of thinking about data, to help scientists to avoid missing important information or acting on false alarms. "Change your statistical philosophy and all of a sudden different things become important," says Steven Goodman, a physician and statistician 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."

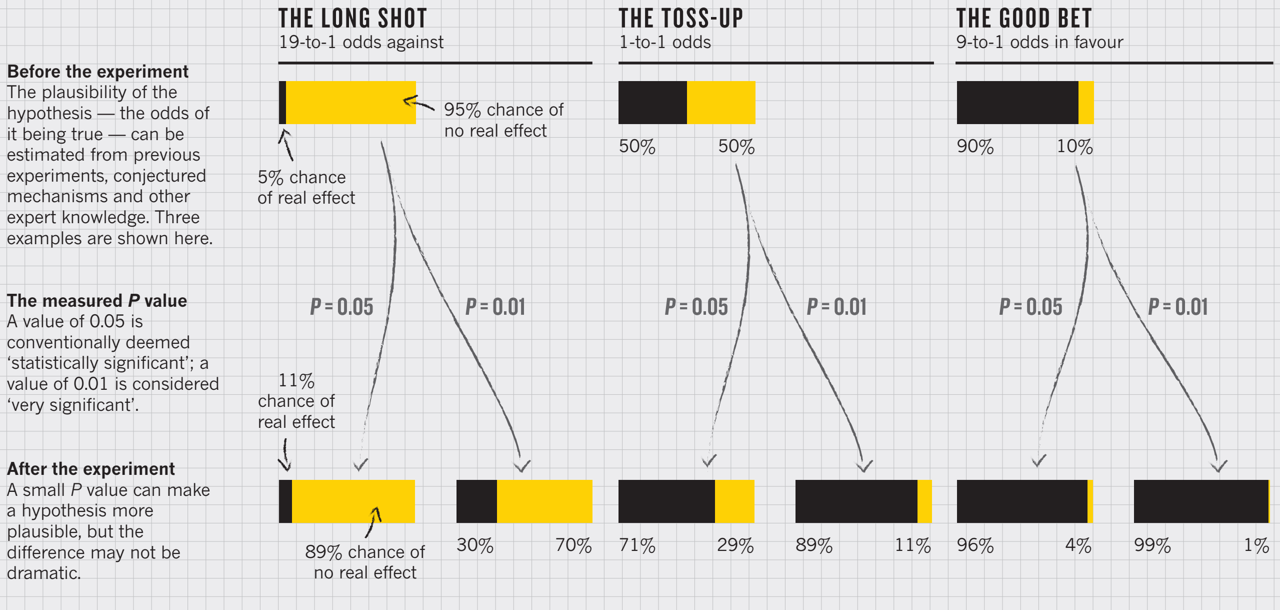
OUT OF CONTEXT

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

© 2014 Macmillan Publishers Limited. All rights reserved





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

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,

© 2014 Macmillan Publishers Limited. All rights reserved

'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).

© 2014 Macmillan Publishers Limited. All rights reserved

THIS WEEK

EDITORIALS

ECOLOGY The troubled tale of the wolf, the moose and the ice **p.132**

WORLD VIEW German court verdict ends scientist's surreal nightmare **p.133**

PESTICIDES Evidence of chemical damage to bumblebees **p.134**

Suicide watch

Despite a high death toll, public-health efforts to combat suicide lag far behind those focused on preventing accidents and diseases such as cancer. A US initiative aims to redress the balance.

Some 38,000 people killed themselves in the United States in 2010. That's more than were killed in traffic accidents (34,000) or by prostate cancer (29,000), and more than twice the number murdered (16,000). Shocking though that is, many other countries monitored by the Organisation for Economic Co-operation and Development have even higher suicide rates. So why do public-health authorities put less effort into preventing death from suicide than they do death from accidents or diseases such as prostate cancer?

One institution that has started to take the matter very seriously is the US army. Since 2008, the suicide rate among soldiers has exceeded that of the general population, and in the past few years the army has lost more soldiers to suicide than to combat. In 2009, the army launched a US$65-million, six-year project called Army STARRS to collect genomic, medical, psychological and lifestyle data from more than 100,000 soldiers to try to identify suicide risk factors and prevention measures, as well as biomarkers of resilience such as epigenetics or brain connectivity. In 2010, it co-launched the National Action Alliance for Suicide Prevention, a public–private partnership, which last week released a pioneering 172-page report on suicide and how it might be tackled.

The report, produced and published in partnership with the US National Institute of Mental Health in Bethesda, Maryland, outlines a strategy to reduce suicide rates in the general population by 20% over the next five years. It also makes shockingly clear how little is known about suicide. There is no standard way to define and so recognize what it means to be suicidal. Relevant statistics are not routinely collected, which makes it hard to know, for example, the effect of round-the-clock crisis teams, and good follow-up care for those who attempt suicide.

Cases of suicide linked to cyber-bullying in young people feature prominently in the media, but few studies have addressed how social media might increase suicide risk through bullying or contagion (prompting copycat suicides). In any case, people over the age of 65 kill themselves much more frequently than do young people.

Two things we do know. First, a high number of people with a psychiatric disorder such as schizophrenia, depression or substance abuse kill themselves — somewhere between 50% and 90% of all suicides are thought to be associated with mental illness. Second, stressful life events, particularly during childhood, greatly increase suicide risk. However, most people who are under stress or mentally ill do not kill themselves. And even as the use of psychiatric medications has soared in the past two decades, suicide rates in the United States and most other countries have remained stable. So what is going on? And what might help?

It will never be possible to eliminate suicide, but it should be possible to reduce rates in different risk groups by attacking the problem from many sides. Biological approaches could identify and help the vulnerable, and sociological interventions could reduce stress in societies.

More long-term studies such as Army STARRS are required to shed light on the biology. And clinical trials can identify therapies that target personality traits or feelings likely to lead to suicide — impulsivity and helplessness, perhaps. One large clinical trial that directly addressed suicide and psychiatric disease indicated that the antipsychotic drug clozapine could help to cut suicide rates in people with schizophrenia (L. Alphs *et al. Schizophr. Bull.* **30,** 577–586; 2004). And small trials have hinted that lithium may do the same for those with depression.

There are no good animal models for suicide risk at present, so biological investigations will have to rely on work with humans. But much can already be done to reduce suicide numbers, even in the absence of biomarkers. One powerful option, on which the report's strategy for reducing suicides by 20% strongly depends, would be to reduce people's access to means of suicide.

"There is no standard way to define and so recognize what it means to be suicidal."

Surprisingly, many people intent on suicide abandon their plan if their chosen means is not available. Firearms account for about half of US suicide deaths, and modelling work carried out for the new report shows that almost 10% of all suicides could be prevented by restricting access to guns. In 2010, 735 people in the United States killed themselves with carbon monoxide from car exhausts; the report suggests that 600 of those deaths might have been prevented if car manufactures were required to install a sensor inside the vehicle that turns off the engine when carbon monoxide builds up.

The report's 20% target will probably not be achieved in the desired five years, but it opens a useful debate that will help more people to understand that the action of committing suicide needs to be considered in the same way as a disorder — as something that can be addressed, not an unavoidable product of circumstance. ■

Number crunch

The correct use of statistics is not just good for science — it is essential.

In the fragmented media marketplace, it is a brave publisher that dismisses the professional competence of most of its readers. So sensitive subscribers might want to avoid page 150 of this week's *Nature*.

The criticism in question appears in a News Feature on the thorny issue of statistics. When it comes to statistical analysis of experimental data, the piece says, most scientists would look at a P value of 0.01 and "say that there was just a 1% chance" of the result being a false alarm. "But they would be wrong." In other words, most researchers do not understand the basis for a term many use every day. Worse, scientists misuse it. In doing so, they help to bury scientific truth beneath an avalanche of false findings that fail to survive replication.

© 2014 Macmillan Publishers Limited. All rights reserved

As the News Feature explains, rather than being convenient shorthand for significance, the *P* value is a specific measure developed to test whether results touted as evidence for an effect are likely to be observed if the effect is not real. It says nothing about the likelihood of the effect in the first place. You knew that already, right? Of course: just as the roads are filled with bad drivers, yet no-one will admit to driving badly themselves, so bad statistics are a well-known problem in science, but one that usually undermines someone else's findings.

The first step towards solving a problem is to acknowledge it. In this spirit, *Nature* urges all scientists to read the News Feature and its summary of the problems of the *P* value, if only to refresh their memories.

The second step is more difficult, because it involves finding a solution. Too many researchers have an incomplete or outdated sense of what is necessary in statistics; this is a broader problem than misuse of the *P* value. Among the most common fundamental mistakes in research papers submitted to *Nature*, for instance, is the failure to understand the statistical difference between technical replications and independent experiments.

Statistics can be a difficult discipline to master, particularly because there has been a historical failure to properly teach the design of experiments and the statistics that are relevant to basic research. Attitudes are also part of the problem. Too often, statistics is seen as a service to call on where necessary — and usually too late — when, in fact, statisticians should be involved in the early stages of experiment design, as well as in teaching. Department heads, lab chiefs and senior scientists need to upgrade a good working knowledge of statistics from the 'desirable' column in job specifications to 'essential'. But that, in turn, requires universities and funders to recognize the importance of statistics and provide for it. *Nature* is trying to do its bit and to acknowledge its own shortcomings. Better use of statistics is a central plank of a reproducibility initiative that aims to boost the reliability of the research that we publish (see *Nature* **496,** 398; 2013). We are actively recruiting statisticians to help to evaluate some papers in parallel with standard peer review — and can always do with more help. (It has been hard to find people with the right expertise, so do please get in touch.) Our sister journal *Nature Methods* has published a series of well-received columns, Points of Significance, on statistics and how to use them.

"Too many researchers have an incomplete or outdated sense of what is necessary."

Some researchers already do better than others. In the big-data era, statistics has changed from a way to assess science to a way of doing science — and some fields have embraced this. From genomics to astronomy, important discoveries emerge from a mass of information only when they are viewed through the correct statistical prism. Collaboration between astronomers and statisticians has spawned the discipline of astrostatistics. (This union is particularly apposite, because it mirrors the nineteenth-century development of statistical techniques such as least squares regression to solve problems in celestial mechanics.)

Among themselves, statisticians sometimes view their contribution to research in terms of a paraphrase of chemical giant BASF's classic advertising tag line: "We don't make the products. We make them better." In doing so, they sell themselves short. Good statistics can no longer be seen as something that makes science better — it is a fundamental requirement, and one that can only grow in importance as funding cuts bite and competition for resources intensifies.

Most scientists use statistics. Most scientists think they do it pretty well. Are most scientists mistaken about that? In the News Feature, *Nature* says so. Go on, prove us wrong. ■

Lone wolves

A declining island wolf population underlines the influence that humans have on nature.

Ecologists have studied the wolves and moose on Isle Royale, a remote island in Lake Superior, for more than 50 years. As we report on page 140, after decades of isolation and inbreeding, the wolf population may be on the verge of dying out.

The US National Park Service, which manages the island, is moving slowly in deciding how to proceed. It has three options: total non-intervention; reintroduction of wolves only after the current population has hit zero; or pre-emptive genetic rescue by bringing in wolves from the mainland to diversify the gene pool. Arguments for non-intervention tend to rely on the perceived need to let nature take its course. This is nonsense. The whole system is highly artificial: wolves and moose have been on the island for less than 100 years, and human activity has been key to the wolves' decline. A previous wolf-population crash in the 1980s was caused by a disease transmitted by a domestic dog. Anthropogenic climate change is almost certainly reducing how often ice bridges form to the mainland, which makes it hard for new wolves to come to the island. Some even think that humans put moose on Isle Royale in the first place.

Arguments are more convincing for reintroducing wolves only if the current population dies out: waiting and watching may yield some useful insights into how highly inbred populations function. But the ecologists who run the island's predator–prey observation study warn that, as the wolves die out, the moose will gorge unchecked on their key food plant, balsam fir, preventing the plant from regenerating. The researchers think that by the time the old wolf population has died out and a new one is established, the ecosystem may have become dominated by pine or spruce, without enough firs to support a moose population that can in turn feed a viable wolf population. If the wolves die out, they could become nearly impossible to reintroduce.

And that might be fine, except that tourists and locals love the wolves of Isle Royale, and the National Park Service was founded with an obligation to protect "the enjoyment of future generations". Furthermore, the predator–prey study — the world's longest — would have to end. That would be a shame: it would be difficult to find another place where none of the predators, herbivores or trees are routinely exploited by humans.

The study's lead ecologists are in favour of genetic rescue. This fairly cheap intervention would allow the project to continue, and would stabilize an ecosystem with which many people feel a deep connection. Some researchers have suggested that any data on reintroduced wolves would have to be treated with caution. Certainly, the influence of the reintroduction would be acknowledged and studied. But the introduced population would not be any more artificial than the population that survived disease, or that which could suffer the effects of climate change.

Isle Royale data help ecology to approach one of its grandest questions. As study leader John Vucetich puts it: "Are ecosystems like other physical systems, governed by law-like patterns and processes, or are they more like human history, where we see one contingency after the next?" The early years of the study seemed to support predictions that in a closed system, predator and prey populations would follow law-like mirror-image cycles, driven by predation pressure. But the data never fitted the theoretical curves that well. And since then, factors from disease to fir abundance, weather, moose ticks and wolf inbreeding have taken turns as the key driver in shaping the populations.

The driver that will shape the future of Isle Royale is now the decision on whether to stage a rescue. Thus of the story of all Earth's systems is writ small on a wooded isle in a frozen lake: the course of human history is no longer merely analogous to the course of ecology. Ecology depends on human history. ■

➲ NATURE.COM
To comment online, click on Editorials at:
go.nature.com/xhunqv

© 2014 Macmillan Publishers Limited. All rights reserved