World view

By Dorothy Bishop

How scientists can stop fooling themselves

Sampling simulated data can reveal common ways in which our cognitive biases mislead us.

The past decade has seen a raft of efforts to encourage robust, credible research. Some focus on changing incentives, for example by modifying promotion and publication criteria to favour open science over sensational breakthroughs. But attention also needs to be paid to individuals. All-too-human cognitive biases can lead us to see results that aren't there. Faulty reasoning results in shoddy science, even when the intentions are good.

Researchers need to become more aware of these pitfalls. Just as lab scientists are not allowed to handle dangerous substances without safety training, researchers should not be allowed anywhere near a *P* value or similar measure of statistical probability until they have demonstrated that they understand what it means.

We all tend to overlook evidence that contradicts our views. When confronted with new data, our pre-existing ideas can cause us to see structure that isn't there. This is a form of confirmation bias, whereby we look for and recall information that fits with what we already think. It can be adaptive: humans need to be able to separate out important information and act quickly to get out of danger. But this filtering can lead to scientific error.

Physicist Robert Millikan's 1913 measurement of the charge on the electron is one example. Although he claimed that his paper included all data points from his famous oil-drop experiment, his notebooks revealed other, unreported, data points that would have changed the final value only slightly, but would have given it a larger statistical error. There has been debate over whether Millikan intended to mislead his readers. But it is not uncommon for honest individuals to suppress memories of inconvenient facts (R. C. Jennings *Sci. Eng. Ethics* **10**, 639–653; 2004).

A different type of limitation promotes misunderstandings in probability and statistics. We've long known that people have trouble grasping the uncertainty inherent in small samples (A. Tversky and D. Kahneman *Psychol. Bull.* **76**, 105–110; 1971). As a topical example, suppose 5% of the population is infected with a virus. We have 100 hospitals that each test 25 people, 100 hospitals that test 50 people and 100 that test 100 people. What percentage of hospitals will find no cases, and wrongly conclude the virus has disappeared? The answer is 28% of the hospitals testing 25 people, 8% of those testing 50 people and 1% of those testing 100. The average number of cases detected by the hospitals will be the same regardless of the number tested, but the range is much greater with a small sample.

This non-linear scaling is hard to grasp intuitively. It leads people to underestimate just how noisy small samples can be, and hence to conduct studies that lack the statistical power needed to detect an effect.

> "Researchers need to build lifelong habits to avoid being led astray by confirmation bias."

Nor do researchers appreciate that the significance of a result as expressed in a *P* value depends crucially on context. The more variables you explore, the more likely it is that you'll find a spuriously 'significant' value. For instance, if you test 14 metabolites for association with a disorder, then your probability of finding at least one *P* value below 0.05 — a commonly used threshold of statistical significance — by chance is not 1 in 20, but closer to 1 in 2.

How can we instil an understanding of this? One thing is clear: conventional training in statistics is insufficient, or even counterproductive, because it might give the user misplaced confidence. I'm experimenting with an alternative approach: generating simulated data that students can subject to various kinds of statistical analysis. I use this to teach two crucial concepts.

First, if presented with null data sets (such as random numbers), students rapidly discover how easy it is to find false results that seem statistically 'significant'. Researchers have to learn that the interpretation of a *P* value is very different when their question is "Is A associated with B?" from when it is "For variables A, B, C, D and E, are there any correlations where $P < 0.05$?" Asking whether a particular metabolite is associated with a disease is not the same as searching a set of metabolites to see whether any are associated with it. The latter requires much more stringent testing.

Simulated data also provide insights when samples come from two 'populations' with different means. Students rapidly learn that, with small sample sizes, an experiment might be useless for revealing even a moderate difference. A 30-minute session of data simulation can leave researchers stunned when they understand the implications.

Researchers need to build lifelong habits to avoid being led astray by confirmation bias. Observations that are contrary to our expectations need special attention. In 1876, Charles Darwin said that he made it a habit "whenever a published fact, a new observation or thought came across me, which was opposed to my general results, to make a memorandum of it without fail and at once: for I had found by experience that such facts and thoughts were far more apt to escape from the memory than favourable ones". I myself have experienced this. When writing literature reviews, I have been shocked to realize that I had completely forgotten to mention papers that run counter to my own instincts, even though the papers had no particular flaws. I now make an effort to list them.

We all find it difficult to see the flaws in our own work — it's a normal part of human cognition. But by understanding these blind spots, we can avoid them.

Dorothy Bishop is an experimental psychologist at the University of Oxford, UK
e-mail: dorothy.bishop@psy.ox.ac.uk

ROBERT TAYLOR PHOTOGRAPHY