FEATURES - SCIENCE & HEALTH: Errors behind fluke results

By Robert Matthews, Financial Times
Published: Jul 09, 2004

Can we really believe what we read in the press - not the tabloids, but scientific papers in world-renowned research journals? This disturbing question has been raised by the discovery of several statistical blunders in two such journals, Nature and the British Medical Journal (BMJ), by researchers at the University of Girona, Spain.

Emili Garca-Berthou and Carles Alcaraz took a sample of papers from the two journals and checked the calculations in them - specifically, those used to measure "statistical significance", a widely used criterion for deciding which findings are worth taking seriously.

The researchers found at least one error in one-quarter of the papers in the BMJ and more than one-third of those in Nature. Most were trivial slips, but the researchers suspect that about 4 per cent could have a big impact on the conclusions drawn.

The results, published in the online journal BMC Medical Research Methodology, have sparked controversy; The Economist last month condemned "sloppy stats which shame science".

Many scientists will see this as an over-reaction. Yet both responses are misplaced. The reality is that these occasional slips are just a small part of a scientific scandal that has been rumbling on for years.

It centres not on computational blunders, but on the routine abuse by scientists of the very concept of statis tical significance. Its impact raises questions about far more than the 4 per cent of results that triggered the recent hand-wringing.

Introduced in the 1930s by the eminent British statistician R.A. Fisher (later Sir Ronald Fisher), "significance testing" has become a ritual for scientists trying to back up claims that they have made an interesting discovery. Experimental data - say, the cure-rates of two competing drugs - are fed into a formula, which spits out a number called a P-value. If this is less than 0.05, the result is said to be "statistically significant".

The technique was seized on by researchers looking for a hard-and-fast rule for making sense of their data. The P-value has since become a standard feature of research papers in many fields, especially softer sciences such as psychology.

Yet almost as soon as it was introduced, significance testing caused alarm among statisticians. Their concern stemmed from the fact that the whole concept of the P-value rests on an assumption of which many scientists seemed unaware.

At first sight, a P-value appears to be the probability that a finding is just a fluke; with a "statistically significant" P-value of 0.05 implying a 95 per cent chance that the finding is not a fluke, and thus worth taking seriously.

Statisticians warned, however, that this was a dangerous misconception. The theory behind P-values, they pointed out, assumes as a precondition that every finding is a fluke. It then asks what is the probability of observing a result at least as extreme as that seen, given that the finding is a fluke.

This probability cannot measure the chances of the finding being a fluke in the first place, as that assumption has already been made. Yet this is precisely what scientists began to use P-values for - with potentially disastrous consequences for the reliability of research.

As long ago as 1963, a team of statisticians at the University of Michigan warned that P-values were "startlingly prone" to see significance in fluke results. Such warnings have been repeated many times since. During the 1980s, James Berger, a professor at Purdue University, and colleagues published a series of papers showing that P-values exaggerate the true significance of implausible findings by a factor of 10 or more - implying that vast numbers of "statistically significant" results in the scientific literature are actually meaningless flukes.

Despite this, scientists still routinely use P-values to assess new findings - not least because getting "statistically significant" results is virtually a *sine qua non* of having papers accepted by leading journals.

My own study of all the papers in a recent volume of the leading journal Nature Medicine shows that almost two-thirds cite P-values. Of these, more than 30 per cent show clear signs that the paper's authors do not understand their meaning.

Nor is this atypical: in a study to be published this summer, Gerd Gigerenzer and colleagues at the Max Planck Institute for Human Development in Berlin describe a survey showing that, even among academics teaching statistics in six German universities, 80 per cent had a flawed understanding of significance testing.

The impact of this on the practice and reliability of scientific research is disturbing. By exaggerating the real "significance" of findings, P-values have led to a host of spurious assertions gaining false credibility, from health scares to claims for paranormal phenomena (see right). Attempts to replicate such findings have wasted untold time, effort and money.

Prof Gigerenzer and his colleagues will call for a radical overhaul in statistics education to wean scientists off reliance on P-values. Yet experience suggests it is the editors of leading journals that hold the key to bringing about change. In 1986, Prof Kenneth Rothman of Boston University, editor of the American Journal of Public Health declared he would no longer accept results based on P-values. His stance led to changes in statistical courses at leading public health schools, which taught their students more sophisticated statistical methods.

Medical journals have also begun to require that authors do more than calculate P-values to back up their claims. Dr Juan Carlos Lopez, chief editor of Nature Medicine, says that while the journal has no plans to eliminate P-values, it is carrying out an investigation into the scale of the problem before deciding on action.

After decades of warnings, scientists may finally be waking up to the dangers of P-values. Their ability to exaggerate significance may have led to many headline-grabbing findings - but the price has been to make academic journals barely more credible than the tabloids.

The writer is visiting reader in science at Aston University, Birmingham

BIZARRE CLAIMS BASED ON FLUKE RESULTS The ability of P-values to exaggerate the real "significance" of meaningless fluke results has led to a host of implausible discoveries entering the scientific literature. Many headline-grabbing health stories are based on evidence backed by P-values: last month, Japanese researchers used P-values to claim that women who lose their teeth in later life are at

higher risk of heart disease. Concern about use of P-values to back implausible claims is mounting. In March, a team from the US National Cancer Institute, Bethesda, Maryland, warned: "Too many reports of associations between genetic variants and common cancer sites and other complex diseases are false positives". It added: "A major reason for this unfortunate situation is the strategy of declaring statistical significance based on a P-value alone." The use of P-values has proved valuable to those seeking scientific backing for such flaky notions as the existence of bio-rhythms and the effectiveness of wishing for good weather. One of the most bizarre examples centres on a study by Leonard Leibovici of the Rabin Medical Centre, Israel, purporting to show the effectiveness of "retroactive prayer". Some early research has hinted that patients may benefit from being prayed for. According to Prof Leibovici's study, published in the British Medical Journal in 2001, prayers even helped patients who had already recovered. The findings, whose supposed significance was demonstrated using P-values, sparked calls for a complete overhaul in notions of space and time. To statisticians, however, the results are just further proof of the dangers of misunderstanding P-values.