Facts *versus* Factions: the use and abuse of subjectivity in scientific research

***Robert A.J. Matthews**, Visiting Research Fellow, Aston University, Birmingham, UK*

Based on a monograph for the European Science and Environment Forum, Cambridge, UK.
Published September 1998.

Summary

This paper explores the use and abuse of subjectivity in science, and the ways in which the scientific community has attempted to explain away its curiously persistent presence in the research process. This disingenuousness is shown to be not only unconvincing but also unnecessary, as the axioms of probability reveal subjectivity to be a mathematically ineluctable feature of the quest for knowledge. As such, concealing or explaining away its presence in research makes no more sense than concealing or explaining away uncertainty in quantum theory. The need to acknowledge the ineluctability of subjectivity transcends issues of intellectual honesty, however. It has profound implications for the assessment of new scientific claims, requiring that their inherent plausibility be taken explicitly into account. Yet as I show, the statistical methods currently used throughout the scientific community lack this crucial feature. As such, they grossly exaggerate both the size of implausible effects and their statistical significance, and lend misleading support to entirely spurious "discoveries". These fundamental flaws in conventional statistical methods have long been recognised within the statistics community, but repeated warnings about their implications have had little impact on the practices of working scientists. The result has been an ever-growing number of spurious claims in fields ranging from the paranormal to cancer epidemiology, and continuing disappointment as supposed "breakthroughs" fail to live up to expectations. The failure of the scientific community to take decisive action over the flaws in standard statistical methods, and the resulting waste of resources spent on futile attempts to replicate claims based on them, constitutes a major scientific scandal.

Introduction

There can be no doubt that science advances. Even the most casual review of the scientific literature shows that our knowledge of the universe, its contents and our place within it is greater and more reliable now than at any other time. This more or less steady progress from ignorance to insight is widely ascribed to the insistence of scientists on the dispassionate and rational assessment of quantitative facts. In other academic disciplines such convincing evidence of progress is more elusive, as fashionable ideas come and go. But in science, objectivity paves the Golden Road to knowledge.

The need to base science on objective fact rather than mere opinion, prejudice or authority is regarded as axiomatic by the scientific community. Galileo's dispute with the Vatican ultimately centred on a battle between objectivity and religious dogma. Objectivity has allowed phenomena quite beyond the bounds of human experience and common sense, from anti-matter to curved space-time, to be discovered, studied and exploited. It has cut through bitter arguments in fields as diverse as human evolution to the cause and cure of disease. Such successes have led to objectivity being regarded as a hallmark that distinguishes genuine science from pseudo-science, quackery and fraud. As the philosopher of science Imre Lakatos puts it:

> "The objective, scientific value of a theory is independent of the human mind which creates it or understands it. Its scientific value depends only on what objective support these conjectures have in facts." (Lakatos 1978 p1)

Einstein admitted that he found the objectivity of science to be one of its most powerful personal attractions:

> "A finely tempered nature longs to escape from the personal life into the world of objective perception and thought". (quoted in Hoffman 1975, p221)

Hardly surprisingly, therefore, any attempt to argue that subjectivity may still be a potent force in science tends to provoke a vociferous response from the scientific community. Those who make such claims - especially if they are themselves non-scientists - are often accused of being supporters of the so-called "anti-science" movement, in which all scientific knowledge is seen as merely a social construct, a product of the prevailing intellectual milieu (see, e.g. Theocharis & Psimopolous 1987). Sociologists and historians of science who back their claims by specific examples of the use of subjectivity in science find themselves confronted with a variety of reactions, ranging from special pleading - "Great scientists have great judgement" (cf Wolpert 1992 p 95), through complacency - "We know better now" (cf Feynman 1985 p 342) - to *ad hominem* attack: "These people are out of their depth" (cf Dunstan 1998).

Such responses hint at a more complex relationship between scientific research and subjectivity, one with which many scientists feel somewhat ill at ease. As I now show, one reason is the recognition by working scientists that they routinely rely on subjective criteria to help them in their working lives.

The use of "everyday" subjectivity in research

Despite their public image as dispassionate seekers after truth, it is common knowledge within the scientific community that subjective methods have a vital role to play in everyday research. All working scientists are constantly bombarded with new research findings and theoretical claims, put forward in seminars, conferences, pre-prints, journals and books. Many of these new claims appear at odds with current belief. If all scientists were truly objective, however, they would have no alternative but to refuse to hold any view on the correctness or otherwise of these new claims until they had first carried out their own extensive studies.

In practice, of course, they do no such thing, for it is simply impracticable. If every more or less ludicrous claim were objectively researched, scientific progress would slow to a crawl. Even so, scientists do need a way of judging which claims to take seriously and pursue, and in the absence of any hard evidence, they resort to a range of criteria which are shot through with subjectivity. These range from personal experience and knowledge about the plausibility of the claim and its consequences to more *ad hoc* criteria such as the reputation of the researchers making the claims, their academic affiliation, and the quality of the journal in which their claims appear. As even Lewis Wolpert, one of the staunchest defenders of the public image of science, has admitted: "One of the reasons for going to meetings is to meet the scientists in one's own field so that one can form an opinion of them and judge their work" (quoted in Collins 1998, p20).

To criticise researchers for relying on subjectivity at this level of the scientific process is clearly absurd. There is simply not enough time, resources or money to appraise objectively each new scientific claim that emerges. The fact remains, however, that while its use may be justified on the grounds of expediency, the exercise of personal judgement, no matter how "professional", is patently subjective, and has inherent dangers. Of these, the one that seems uppermost in the minds of researchers is that admitting to the presence of subjectivity in science is to play straight into the hands of their perceived enemies among post-modern philosophers and sociologists, who maintain that science is no more objective than literary criticism (Aronson 1984 p12). This fear contains a deep irony, however, and one with which I shall deal in greater detail later.

A more pragmatic concern centres on the belief that unbridled subjectivity can seriously undermine the scientific process, leading to major discoveries being overlooked, dismissed or ignored. As I now show, this concern is well-placed.

Abuses of "everyday" subjectivity

Robert Millikan is widely regarded as one of the founders of modern American science, his determination of the charge on the electron winning him the 1923 Nobel Prize for physics. In a now-famous study, the physicist and historian Gerald Holton examined the log-books for Millikan's experiments with the electron, and revealed that he repeatedly rejected data that he deemed "unacceptable" (Holton 1978). The criteria he used were blatantly subjective, as revealed by the comments in the log-books, such as "Very low - something wrong" and "This is almost exactly *right*". Throughout, Millikan appears to have been driven partly by a desire to get results that were self-consistent, broadly in agreement with other methods, and consistent with his personal view that the electron is the fundamental and indivisible unit of electric charge.

While these criteria may seem reasonable enough, they carry inherent dangers. Even today a fundamental explanation of the precise numerical value of the charge on the electron remains lacking, so Millikan was hardly in a position to decide objectively which values were high and which ones low. Previous results may have been fundamentally flawed, while the demand for self-consistent results may mask the existence of subtle but genuine properties of the electron. Millikan could also have been proved wrong in his belief that the electron was fundamental.

However, it is also clear that Millikan had another powerful motivation for using all means to obtain a convincing determination of the electronic charge: he was in a race against another researcher, Felix Ehrenhaft at the University of Vienna. Ehrenhaft had obtained similar results to those of Millikan, but they were interspersed with much lower values that suggested that the electron was not, in fact, the fundamental unit of charge. Millikan had no such doubts, published his results, and went on to win the Nobel Prize.

To many, this will seem like an egregious example of subjectivity in experimental science. Yet within the scientific community, it has been excused on the grounds that Millikan was, in the final analysis, correct: the electron is the fundamental unit of electric charge. For example, while conceding that "Millikan may have taken his judgement beyond reasonable boundaries", Wolpert argues that the episode provides an object lesson in what distinguishes great scientists from the common herd: "It is that remarkable ability not only to have the right ideas but to judge which information to accept or reject" (Wolpert 1992 p 95). This overlooks the fact that Millikan was *not* correct: fractional units of electronic charge do exist in Nature, in the form of quarks. The discovery in the 1970s of the concept of asymptotic freedom in quantum chromodynamics is now believed to prevent individual quarks from being observed; working 60 years previously, however, Millikan had no such basis for his beliefs. We can only be thankful that Millikan's "remarkable ability" to spot the truth was not available during the early days of the quark hypothesis.

Apologists for Millikan's hand-picking of data also point out that the numerical result he obtained, – 1.592 x 10^{-19} coulombs, is just 0.6 per cent below the modern value of – 1.6021892 x 10^{-19} C (Weinberg 1993 p 99). At first sight, this does indeed seem impressive. However, Millikan's stated result was based on a faulty value for the viscosity of air, which when corrected changes Millikan's result to – 1.616 x 10^{-19} C, increasing the discrepancy with the modern value by over 40 per cent. More importantly, however, it puts the latter well outside the error-bounds of Millikan's central estimate. Indeed, the discrepancy is so large that the probability of generating it by chance alone is less than 1 in 10^3. Millikan's "remarkable ability" to scent out the correct answer was clearly not as great as his apologists would have us believe. Rather more remarkable is Millikan's ability, almost half a century after his death, to evade recognition as an insouciant scientific fraudster who won the Nobel Prize by deception.

The dangers of the injudicious use of subjective criteria is further highlighted by the aftermath of Millikan's experiments. In the decades following his work and Nobel Prize, other investigators made determinations of the electronic charge. The values they obtained show a curious trend, creeping further and further away from Millikan's "canonical" value, until finally settling down at the modern figure with which, as we have seen, it is wholly incompatible. Why was this figure not reached sooner ? The Nobel Prizewinning physicist Richard Feynman has given the answer in his own inimitable style (Feynman 1988, p 382):

> "It's apparent that people did things like this: when they got a number that was too high above Millikan's, they thought something was wrong - and they would look for and find a reason why something might be wrong. When they got a number closer to Millikan's value they didn't look so hard. And so they eliminated the numbers that were too far off"

Feynman described this example of subjective influence of personality in science as "A thing that scientists are ashamed of". Yet even Feynman, one of the most individualistic of scientists, fell back into line with the rest of the scientific community when assessing the ultimate relevance of the Millikan case for contemporary science: "We've learned those tricks nowadays", he insists, "And now we don't have that kind of disease". Such complacency is hard to reconcile with the many examples of scientific fraud by influential individuals that have come to light since the Millikan case (see, e.g. Grayson 1995, 1997 and references therein).

Experimental science is not alone in being vulnerable to abuses of subjective criteria; theoretical advances can and have been gravely affected as well. Some of the most egregious examples centre on the influence of the brilliant but notoriously arrogant theorist Wolfgang Pauli, whose dismissive opinions of the work of a number of theoreticians led to their being denied credit for major scientific discoveries in elementary particle physics. For example, the discovery of the key quantum-theoretic concept of spin is widely ascribed to Uhlenbeck and Goudsmit. However, it was first put forward by the young American theorist Ralph Kronig, who was persuaded not to publish after being ridiculed by Pauli and informed that while "very clever", the concept of spin "Of course has nothing to do with reality" (quoted in Pais 1991 p 244). Caustic *ad hominem* remarks by Pauli similarly led to the Swiss theorist Ernst Stueckelberg failing to publish his exchange model of the strong nuclear force; Yukawa subsequently published essentially identical ideas, and won the 1949 Nobel Prize for Physics. (Stueckelberg's work on renormalisation of quantum electrodynamics met a similar fate, being later duplicated by three other theorists who went on to win the 1965 Nobel Prize for physics (Crease & Mann 1996, p 142-3)). During the 1950s, Pauli together with the charismatic and influential theorist Robert Oppenheimer succeeded in stifling discussion of the de Broglie-Bohm interpretation of quantum theory by a combination of spurious arguments and subjective criticism. After being told that supposedly knock-out arguments against the de Broglie-Bohm interpretation were invalid, Oppenheimer is alleged to have remarked that "Well....we'll just have to ignore it" (quoted in Matthews 1992 p 146); ironically, Oppenheimer went on to write a book whose central thesis was the need for an open mind in science (Oppenheimer 1955).

Of all concepts in particle physics, however, none so vividly displays the presence of subjectivity within the "hard" sciences as the nature of the fundamental constituents of matter. The concept of the atom - the ultimate, indivisible particle of matter - was first raised by the Greek philosopher Leukippos in the 4th Century BC, yet even as late as 1900 the physical reality of atoms was still rejected by influential scientists, most notably the Austrian physicist Ernst Mach, and the German chemist Wilhelm Ostwald. Their refusal to countenance the existence of atoms was based largely on a Positivist agenda, in which the lack of direct evidence for atoms - and supposed impossibility of obtaining any - *ipso facto* implied their non-reality. This view led them to mount a sustained and vociferous campaign against the views of the Austrian physicist Ludwig Boltzmann, who had shown that the presumption of the physical reality of atoms led to natural explanations for the bulk properties of matter. Boltzmann and his work was successfully marginalised for many years, and by the time of his suicide in 1906, he was regarded as a scientific "dinosaur" (Greenstein 1998 p 50). Ironically, barely a year before his death, a paper appeared which ultimately established the reality of atoms. It was an analysis of the phenomenon of so-called Brownian motion, the random movement of particles in a suspension which was shown to be explicable by the existence of atoms; the author of the paper was a young patents clerk named Albert Einstein. Within two years of Boltzmann's death, experimental studies of Brownian motion had compelled even Ostwald to accept the reality of atoms.

The Boltzmann case shows how the subjective (in this case, philosophical) prejudices of a few influential individuals can prevent the acceptance and application of fundamental advances for decades. What makes the case especially interesting, however, is the way in which its principal features emerged again 60 years later, with the controversy over the concept of quarks. The claim that the neutron and proton, supposedly fundamental components of atoms, are not indivisible was first put forward in 1964 in an eight-paragraph note in *Physics Letters* by the American physicist Murray Gell-Mann (Gell-Mann 1964). Like Boltzmann, Gell-Mann based his claim on a mathematical demonstration of the explanatory power of the new concept; in this case, the ability of quarks to explain the properties of hadrons. This in turn led Gell-Mann to predict that quarks had fractional electric charges. The absence of evidence for such charges he ascribed to the permanent confinement of the quarks within their host particles.

Like Boltzmann, Gell-Mann found considerable resistance to his proposal within the physics community, stemming from two subjective prejudices. The first was a throw-back to the days of Millikan, and the insistence that electronic charge was indivisible. The second was an echo of the Positivist arguments against Boltzmann. Gell-Mann put forward the concept of confinement - and thus the impossibility of the direct observation of individual quarks - to avoid the philosophical wrangling that had dogged Boltzmann; (Gell-Mann 1994 p182). Ironically, his rather opaque statement that quarks "exist but are not 'real' " had precisely the opposite result: according to Gell-Mann, quarks "went over like a lead balloon", with colleagues refusing point-blank to take them seriously, ridiculing the concept in the professional literature (Crease & Mann 1996 p 283-5). This was, however, a relatively mild reaction compared to those encountered by George Zweig, a young American theorist who proposed essentially the same explanation (based around "Aces" rather than quarks) in 1964, but emphasised their physical reality. His papers were summarily rejected, and his appointment to a position at a major university blocked by the head of department on the grounds that he was a "charlatan" (Crease & Mann 1996 p 285). Even so, in yet another parallel with the Boltzmann case, within five years experiments at the Stanford Linear Accelerator had demonstrated the reality of quarks within hadrons. They are now at the heart of quantum chromodynamics, the most successful theory for the strong nuclear force.

It is not difficult to find examples of where subjective prejudice has seriously delayed progress in many other fields:

- Semmelweiss's long and unsuccessful struggle during the 1840s to introduce antiseptic practices into hospitals (Asimov 1975 p 348). Despite the existence of a dramatic fall in the numbers of cases of childbed fever produced by the use of antiseptics, the practice was rejected because of resentment by the doctors that they could be causing so many deaths, nationalistic prejudice against a Hungarian working in a Viennese hospital, and annoyance at the way the antiseptics eliminated the "professional odour" on their hands after returning to the wards from working in the mortuary.
- The refusal of the astronomical community to accept reports of "stones falling from the sky", as had been long reported by many ordinary people, until investigations by Biot in the early 19th Century (Milton 1994, p 3-4). This refusal seems to have had stemmed from a combination of disdain for the claims of non-scientific outsiders, and a prejudice against the notion that the Earth could be subject to potentially serious bombardment.
- The rejection and ridiculing of Francis Peyton Rous's evidence for the existence of viruses capable of transmitting cancer (Williams 1994, p422). First put forward in 1911, Rous's evidence came at a time when the existence of viruses was still controversial - they were beyond the reach of contemporary microscopy - and when cancer was thought to be caused by "tissue irritation". Rous's claim was finally vindicated 25 years later. In 1966 he was awarded the Nobel Prize - at the age of 87.
- The vociferous response of geologists to the proposal by Alfred Wegener, a German astronomer and meteorologist, that the continents moved across the face of the Earth. Having found considerable evidence for the phenomenon, but unable to propose a physical mechanism for it, Wegener's proposal was dismissed as a "fairy tale", the product of "auto-intoxication in which the subjective idea comes to be considered as an objective fact" (Hellman 1998 p150). His claims were subsequently vindicated in the 1960s, 50 years after he first proposed them, and 30 years after his death.
- In the early 1980s, the Australian physician Barry Marshall encountered derision and hostility for his claim that a previously unknown bacterium, *Helicobacter pylori*, was responsible for stomach ulcers. Marshall's evidence went against the prevailing view that bacteria were incapable of thriving within the acidic conditions of the stomach. *H. pylori* is now accepted as the principal cause of stomach ulcers, and has also been implicated in gastric cancer.

Together with the battles faced by advocates of the atomic and quark concepts, these examples hardly support the complacent view that the scientific community has "learned its lesson", and now "knows better" how to recognise when professional judgement slips into subjective prejudice. Indeed, it is clear from these examples that subjectivity has played, and continues to play, a considerable role in the development of science. My principal aim in choosing these specific examples is not, however, to suggest that subjectivity is a uniquely evil force in science. Rather, it has been to show that the "official" responses to such examples - that they have only short-term effects, or are confined to less quantitative sciences, or are "all behind us now" - are not sustainable.

A rather more cogent response is that which many working scientists give, at least when out of earshot of the guardians of the public image of science: that while regrettable, the cases cited above represent a "price worth paying" for retaining subjective criteria to separate the scientific wheat from the chaff.

I shall now show that this "pragmatic" view is not only supported in practice, but also has a firm theoretical basis in the mathematics of scientific inference. In short, the presence and use of subjectivity in science *need not* be glossed over, explained away or concealed. Indeed, I shall demonstrate that subjectivity *must* not be treated in this way. For as we shall see, the continuing and misguided attempts to portray scientific research as a wholly objective pursuit has led to practices which threaten its reputation as a source of reliable knowledge.

Subjectivity in the testing of theories

The value of any scientific theory, no matter how theoretically elegant or plausible, is ultimately tested by experiment. Conventionally, this crucial element of the scientific process involves extracting a clear and unequivocal prediction from the theory, investigating this prediction experimentally, and assessing the outcome objectively. Exactly how this comparison is performed, and what conclusions are drawn, has long been a subject of debate among scientists and philosophers. Many scientists consider themselves to be followers of Karl Popper and the concept of falsifiability (Popper 1963): that to be considered scientific, a theory must be capable of being proved wrong. On this view, the experiment and the analysis of data should be performed to discover if the theory is falsified, and if it is, it must be abandoned. As such, theories are never proved "correct": they merely survive until the next experimental attempt at falsification.

There are a great many fundamental problems with Popper's widely-held - and admittedly appealing - view of the scientific process (see especially Howson & Urbach 1993). Put simply, these problems boil down to the fact that the concept of falsification is supported neither in principle nor in practice. Over 90 years ago the French physicist and philosopher Pierre Duhem pointed out that the testable consequences of scientific theories are not a pure reflection of the theory itself, but are based on many extra assumptions. As a result, if an experiment appears to falsify a theory, this does not automatically imply that the theory must be false: it is always possible to blame one of the auxiliary assumptions.

It should be stressed that this is not merely a philosophical objection to the concept of falsifiability: there are many cases of now well-attested theories being "falsified", from the Standard Model of elementary particle physics (Crease & Mann 1996 pp 383-390) through to the concept of cancer viruses (Wolpert & Richards 1989 Ch 12). Even Einstein's special theory of relativity was "falsified" barely a year after its publication. In what appears to be the very first published response citing Einstein's famous paper, Walter Kaufmann at the University of Gottingen reported that two rival theories gave a better fit to data from studies of beta particles than relativity. Einstein conceded that Kaufmann's work was carefully executed, based on solid theory, and that the results showed a better fit with rival theories. Even so, he bluntly refused to concede defeat, arguing on the entirely subjective grounds that the rival theories seemed to him inherently less plausible. It took another decade for Einstein's view to be vindicated (Pais 1982 p159).

Once again we see a major disparity between the way science is said to operate and how it actually does. We again see scientists applying subjective criteria for essentially pragmatic reasons: it simply makes no sense to take seriously every apparent "falsification" of a plausible theory, any more than it makes sense to take seriously every new scientific idea. Judgements based on considerations ranging from the reputation of the experimentalists to a hunch about the correctness of a theory may not be utterly reliable, but they appear to work pretty well most of the time.

Yet, once again, there is a reluctance by the scientific community to admit to what every working scientist knows: that, for all its faults, subjectivity plays a key role in setting "objective" experimental findings in their proper context. The need to accept this fact transcends the demands of intellectual honesty, however. For as I shall now show, past attempts to sweep subjectivity "under the carpet" have led to the adoption of apparently "objective" methods for analysing experimental data that are neither objective nor reliable.

The standard theory of statistical inference

The Popperian image of an experiment is one of clear-cut falsification. Yet, as ever, working scientists readily admit that such black and white, pass/fail outcomes are rarely possible (e.g. Medawar 1979 Ch 9). This raises another major objection to the Popperian scheme: for if falsification cannot be clear-cut, what criteria should be used to decide whether a theory has been at least partly falsified ? This problem is most acute where data are *statistical* in nature - the common outcome of experimental investigations in fields from particle physics to psychiatry. Faced with a set of results from, say, a group of depressives where 79 per cent of those given cognitive therapy improved, compared to 68 per cent of those given tricyclics, how is one to decide when the difference between the two groups is "significant" ?

Clearly, there is considerable scope for subjective criteria to be applied here: psychopharmacologists sceptical of "talk therapy" may well demand more impressive findings than their cognitive therapist colleagues. However, the standard techniques for gauging the statistical "significance" of an experimental result seem to eliminate such vexations. These textbook methods of apparently wholly objective statistical inference were developed largely by Ronald Fisher, Jerzy Neyman and Karl Pearson during the 1920s and 1930s. Their aim was to provide objective mathematical tests capable of falsifying theories, and to this end they developed the methods still widely used by the scientific community.

One key feature of these statistical tests is that they appear to require no skill or training in statistics, and seem to lead to a single, objective and easily-understood result. They typically appear in the form of a kind of cook-book recipe, as follows:

1. Specify the hypothesis under test. This is usually the "null hypothesis" of no real difference; for example, that the difference in the proportions of patients benefiting in both the treatment and the control groups is no greater than that due to mere chance. The "alternative" hypothesis would then be that there is an improvement in the treated group that cannot be ascribed to fluke alone.
1. Execution of the experiment (for example, as a double-blind randomised clinical trial), and conversion of results into a so-called test-statistic that captures both the size and variation of the effect under study.
1. Determination of the so-called P-value of the test statistic, that is, the probability of obtaining a test-statistic at least as large as that actually observed, on the assumption that the null hypothesis is actually true.
1. If the P-value is less than a certain cut-off figure (the "level of significance", usually denoted by α), the null hypothesis is held to be "rejected", and the experimental result is deemed "significant at the α level".

While such a recipe is certainly easy to execute, it undoubtedly contains many perplexing features. Most obvious among them is the strangely convoluted definition of the key determinant of "falsifiability", the P-value. This is said to give the probability of obtaining results *at least* as impressive as those actually observed *on the assumption* that the null hypothesis is true. Put another - hardly more illuminating - way, *assuming* the null hypothesis is true, if the same experiment were repeated many times, the frequency with which we would obtain data at least as impressive as those obtained is equal to the P-value (this latter definition leads to these conventional text-book methods being called "frequentist").

Those who bother to analyse either of these convoluted definitions are apt to ask themselves why they should care about a probability involving results never actually obtained, and calculated assuming the very hypothesis under test. Why is the measure of the "significance" of the results not simply the probability of the hypothesis under test being true ?

A little more reflection suggests that these cook-book recipes are not, in fact, truly objective. For example, what objective principle underpins the choice of α , the cut-off level for "significance", or the preference of one frequentist method over another ?

Many of those coming to significance testing for the first time find these issues confusing, and somewhat disturbing (see, e.g. Sivia 1996 p *vi*; Lee 1997 p *ix*). Yet the widespread use of frequentist methods suggests that most statistical neophytes decide that their qualms must stem from some minor philosophical or mathematical misapprehension of little consequence.

It is one of the most disturbing yet poorly-recognised facts of contemporary science that such qualms are far from misplaced. There are indeed fundamental problems with the standard methods of statistical inference, and warnings about their impact on scientific research have been repeatedly pointed out for over 30 years in mathematical research papers (e.g. Edwards *et al.* 1963, Berger & Sellke 1987), textbooks (e.g. Jeffreys 1961, Lindley 1970, Howson & Urbach 1993, O'Hagan 1994, Lee 1997) and even general science publications (e.g. Berger & Berry 1988, Matthews 1997). All these authors have pointed to the conceptual flaws in the standard methods of statistical inference, and the logical and practical dangers they present to the scientific enterprise. So far, however, these warnings have had virtually no effect beyond the community of mathematical statisticians. The bulk of the scientific community still uses the standard techniques, at best only vaguely aware of some apparently esoteric concern over their reliability. As we shall see, this concern could hardly be more serious.

Flaws and failings of standard statistical inference

The failure to provide "objectivity"

The most obvious failing in the standard textbook methods of statistical inference is that they are not objective. This is most clearly apparent in their requirement for a value of α , the cut-off level for "significant" P-values. Textbooks on classical inference typically introduce a value for $\alpha = 0.05$, stating blandly that it is "conventionally used", "widely used", or "accepted" as the value below which a P-value is deemed "significant". Similarly, values of $\alpha = 0.01$ are quoted as being the "standard" cut-off for "highly significant" P-values, and $\alpha = 0.001$ for "very highly significant" results. Yet these same textbooks typically give no clue to the objective underpinnings of these choices. The disturbing truth is that these ubiquitous standards of significance, by which research findings are held to stand or fall, have their origins in nothing more objective or statistically defensible than a coincidence. Through a mathematical quirk of the Normal distribution, 95 per cent of the area under this distribution is enclosed within almost exactly two standard deviations of the mean value. It was this juxtaposition of an integer value for the ordinate and a seemingly convenient 95 per cent probability led Fisher to set $\alpha = 0.05$ as the cut-off for judging significance (Fisher quoted in Jeffreys 1961, p 388-9). As we shall see, it was both an indefensible and unhappy choice.

Altogether more subtle are the logical fallacies lurking in the definitions of frequentist measures of significance. The strangely convoluted definition of the P-value, for example, stems from the fact that it is calculated from an integral, that is, the area under a probability curve such as the familiar bell-shaped normal distribution. This curve is calculated on the *assumption* of the null hypothesis; the fact that the required probability is given by the area under this curve forces the inclusion of entirely hypothetical data points that were, in fact, never observed.

All this is reflected in the more formal mathematical definition of the P-value of Prob(data | null hypothesis). In other words, the P-value is the probability of getting at least as impressive data from an experiment *given* the null hypothesis. While this explains the far-from-intuitive nature of the P-value, it is still far from clear why anyone should be interested in the final result. Working scientists typically want something far more straightforward: the probability that the null hypothesis *really is* correct, *given* the data they observed, that is, Prob(null hypothesis | data).

The difference between this and a P-value seems to be nothing more that switching the order of "null hypothesis" and "outcome". Indeed, the two are often taken to be equivalent even by the authors of some standard statistics texts (see, e.g. Bourke *et al.* 1985 p71, Heyes *et al.* 1993 p116). This is, however, a fundamental and potentially disastrous fallacy known as "transposition of conditioning": the fallacy of taking Prob(A | B) to be always identical to Prob(B | A).

Risk of false interpretation

To see the dangers inherent in this fallacy, suppose a patient walks into a doctor's surgery covered with spots. The doctor knows that the probability of getting spots *given* a measles infection is very close to certainty, i.e. Prob(spots | measles) $\simeq$ 1. However, it clearly does not follow that the probability that the patient really *has* got measles is also close to 1, i.e. that Prob(measles | spots) $\simeq$ 1: there is a vast number of other diseases apart from measles that produce spots. Deciding which the patient has will involve taking into account other sources of information, such as whether there is chicken pox in the family, and whether the patient has recently travelled abroad.

Clearly, mistaking Prob(spots | measles) for Prob(measles | spots) could lead to a doctor being struck off. Yet the standard methods of statistical inference can and do prompt working scientists to fall into precisely the same trap: P-values are all too easily taken to be identical to Prob(null hypothesis | data), so that a low P-value is taken to imply that the probability that chance alone explains the data is similarly low. There is no simple relationship between P-values and the probability working scientists actually want, and as I shall show shortly, confusing the two can and does lead to meaningless fluke results being regarded as "significant".

There is a further serious logical fallacy lurking in the interpretation of a P-value: simply because a result has a low probability on the basis of the null hypothesis, this does not imply that a specific alternative hypothesis is confirmed to a corresponding degree. For example, suppose that a case-control trial shows that a higher proportion of patients on the drug benefited relative to the control group, with a P-value of 0.02. In conventional parlance, as the P-value is below 0.05, this is a "significant" result. As we have seen, however, this does *not* imply that the probability P of the results being a fluke is 1 in 50. Still less does it imply that the probability of the drug being efficacious are 49/50: Prob(efficacy | outcome) does not equal 1-P, and in any case the efficacy of the drug is just one out of a host of possible explanations for a positive result.

It must be said that the existence of these problems has been acknowledged by some advocates of standard inference, who have put forward a number of rejoinders. For example, some concede that P-values may not be particularly relevant, but insist that they are still a simple and convenient way of summarising a research finding. This is hardly convincing. Any summary of data worthy of the name must not mislead those without access to the full results - and as we have seen, P-values are all too likely to mislead. Arguing that they are a "convenient summary" is equivalent to claiming that "A patient with glandular fever has a high probability of swollen glands" is a convenient summary of a diagnosis of the Black Death.

In an attempt to rid frequentist methods of some of their subjectivity, some authors recommend that the P-value alone should be stated, without comparison to the entirely subjective standard cut-off levels for significance (see e.g. Freedman *et al.* 1998 pp 547-8). It is usually conceded, however, that this does nothing to prevent others - especially editors and referees of journals - from making the comparison themselves, and acting accordingly.

Yet others eschew use of P-values altogether, arguing instead for so-called estimation methods and the use of "confidence intervals" (CIs). Rather than using just a single figure, confidence intervals summarise a finding as a central figure, plus a range of values for a parameter of interest, e.g. the relative risk of contracting cancer from some carcinogen. If this range excludes the value corresponding to no additional risk, then the results are deemed to be "significant".

Conscious of the criticisms of P-values, many medical journals now ask for results to be quoted in terms of CIs. Despite appearances, however, CIs still fail to resolve the key issue of the interpretation of the outcome of conventional statistical tests. At first sight, a 95 per cent CI *seems* to imply that there is a 95 per cent probability that the true value of the parameter of interest will lie within the stated bounds. Its correct interpretation, however, is just as convoluted as that of the P-value: the 95 per cent actually refers to the frequency with which the statistical test used will generate bounds capturing the true figure. That is, the "95 per cent confidence" refers to the reliability of the *test*, not to the *parameter*. Indeed, so subtle is this distinction that 95 per cent CIs are arguably even more confusing than P-values. Defenders of their use typically respond that - unlike P-values - the distinction between the perceived and correct meanings of 95 per cent CIs can often be ignored. However, as we shall see, this is true only when there is no prior reason for suspecting that the true value of a parameter lies within a well-defined range of values. It is rare that a claim of such complete ignorance can be justified. In any case, the choice of the value of 95 per cent for the CI is entirely arbitrary and subjective, so that in the end a 95 per cent CI is no more "objective" a measure of significance than a P-value.

Nothing so clearly illustrates the many flaws of frequentist inference than the way in which the scientific community feels able - indeed, sometimes obliged - to decide on entirely subjective grounds which "objectively significant" results they are going to take seriously, and which they will reject.

Subjective interpretations of study outcomes

If scientists and their statistical methods were truly objective, then the research enterprise would be relatively simple. When a carefully designed study finds a sizeable effect with a P-value of less than 0.05 (or, equivalently, a 95 per cent CI that excludes no effect), then everyone would agree that a "significant" effect potentially worthy of further investigation had been found. If, on the other hand, a large study failed to reveal a significant outcome despite having the statistical power to do so, then researchers would know to start to looking elsewhere.

This is, of course, not at all how scientists respond to research findings. Large and "objectively significant" effects found in some fields of research are repeatedly ignored by the scientific community, while small and non-significant effects found in other fields are deemed to be impressive.

For example, researchers at a number of respected academic institutions have investigated the concept of telepathy, the transmission of information from one person to another by extrasensory means. The most highly-regarded studies centre on the so-called autoganzfeld technique (see e.g. Radin 1997 Ch 5), in which subjects have to identify one of four images which a "sender" attempts to transmit to them by telepathic means. The null hypothesis of no telepathy suggests a random hit rate of 0.25; a recent meta-analysis of over 2,500 sessions (Radin 1997 p87) showed an average hit-rate of 0.332, with an extraordinarily "significant" P-value of less than 10^{-15}. By the usual criteria of "objective" statistical inference, such a finding should convince even the most sceptical of the existence of telepathy. Yet many if not most scientists continue to reject the existence of telepathy out of hand, often citing past examples of fraud and incompetence in parapsychology to support their stance (Radin 1997 Ch 13). Similarly, recent trials of a number of homoeopathic treatments have been found to produce large and highly significant effects for some ailments, such as migraine and allergy (for a review, see Vallance 1998). Even so, homoeopathy is still regarded with suspicion by much of the medical profession (see e.g. Vandenbroucke 1997).

Both these examples are clear cases of the use of "double standards". Many scientists feel entirely comfortable about their stance, however, citing the lack of any mechanism to explain telepathy or homoeopathy, and past evidence of fraud and incompetence by researchers in these areas.

Given the lack of clear mechanisms for the action of many drugs, and the cases of fraud and incompetence in entirely "conventional" fields of research, this defence of the use of double standards is hardly convincing. It seems particularly disingenuous when one considers the response of the scientific community to findings in other, more conventional areas of research. Now results that are both minor and statistically non-significant are said to constitute substantial support for the prevailing wisdom. For example, the World Health Organisation (WHO) and International Agency for Research on Cancer (IARC) recently conducted the largest case-control study of the effects of passive smoking ever performed in Europe (Bofetta *et al.* 1997). The aim was to establish, as unequivocally as possible, the extra risk of lung cancer faced by non-smokers who live with smokers. This extra risk is typically quantified by the so-called Odds Ratio (OR), in which an OR greater than 1 constitutes an additional risk.

The WHO/IARC study found only a small and non-significant Odds Ratio (OR) for lung cancer for spouses exposed to environmental tobacco smoke (ETS) of 1.16 with a 95 per cent CI of (0.93 1.44). As well as being statistical non-significant, so small an effect size lies within the range at which the IARC itself concedes that unequivocal results may be forever unachievable (Breslow & Day 1980). Yet following the publication of a negative interpretation of their results in the media (Macdonald 1998), the WHO/IARC team publicly insisted that their findings "add substantially" to previous evidence for the link between ETS and lung cancer. The WHO went on to issue a press release clearly implying that the results proved a link between passive smoking and lung cancer.

No competent statistician would agree that the WHO/IARC results "add substantially" to the case against ETS, much less that they "prove" the existence of a link with lung cancer. Moreover, the WHO's interpretation of such weak evidence is in striking contrast to the "official" interpretation of very similar findings in studies of other supposed health risks, in which the "politically correct" line is one of considerable scepticism. For example, a recent major study of the supposed link between electric power lines and childhood leukaemias (Linet *et al.* 1997) produced an OR of 1.24, with a 95 per cent CI of (0.86 1.79). This result is very similar to that obtained by the WHO/IARC passive smoking study; this time, however, the researchers concluded that so small and non-significant effect provided "little evidence" of a link between power lines and leukaemia. The team's funding organisation, the US National Cancer Institute, went further, declaring that the study showed magnetic fields "do not raise children's leukaemia risk".

Similarly, a recent study of women with breast implants (Nyren *et al.* 1998) found an OR for hospitalisation for connective tissue disorders of 1.3, with a non-significant 95 per cent CI of (0.7, 2.2). This is again similar to the WHO/IARC study findings, but again the lack of significance was held to "add weight" to the conclusion that silicone breast implants "are *not* associated with a meaningful excess risk of connective tissue disorder" (Cooper & Dennison 1998, emphasis added).

There are many other examples of where the results of supposedly "objective" statistical methods are interpreted according to the prevailing subjective opinion of the scientific community. Together, they provide further evidence of the gulf between how scientists are supposed to conduct even quantitative research, and how they actually go about it. The insouciance with which subjectivity is used in the assessment of scientific claims suggests that many working scientists accept - consciously or otherwise - that a key feature is missing from conventional statistical methods: specifically, an explicit means of taking into account the *plausibility* of the claim under study. Indeed, as one leading advocate of frequentist inference has noted, it is "curious that personal views intrude always" (Kempthorne 1971 p 480).

This "curious" fact, combined with the many problems and pitfalls associated with frequentist measures of "significance", raises an obvious question: is there a better way? As I now show, the answer is *yes*.

END OF PART 1. To jump to Part 2, click HERE