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

ESEF Working Paper 2/98

This paper is being issued as a working paper intended for discussion among interested parties. Please refer your comments to Robert Matthews.

rajm@compuserve.com

The European Science and Environment Forum
4 Church Lane, Barton, Cambridge, CB3 7BE
Tel: 01223 264643 Fax: 01223 264645
E-mail: enquiries@scienceforum.net

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

Robert A. J. Matthews
E-mail: *rajm@compuserve.com*

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 p 15).

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, -1.592×10^{-19} coulombs, is just 0.6 per cent below the modern value of $-1.6021892 \times 10^{-19}$ 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 -1.616×10^{-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 one in a thousand. 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[1].

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 Prize-winning 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*

[1] Millikan's cavalier attitude towards scientific research is further evidenced by his dealings with his young assistant Harvey Fletcher over authorship of the key papers on the properties of the electron (Fletcher 1982) and his role in early cosmic ray studies (Crease & Mann 1996, p 150-155)

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.
2. Execution of the experiment (for example, as a double-blind randomised case-control clinical trial), and conversion of results into a so-called test-statistic that captures both the size and variation of the effect under study.
3. 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.
4. 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 homeopathic 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, homeopathy 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 homeopathy, 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 statistically 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*.

Bayesian inference

The classical frequentist techniques of inference are not, in fact, classical at all, but relative newcomers in the long history of statistical inference. Before the 1920s, another approach to statistical inference was in general use, based on a result that flows directly from the axioms of probability. As such, this approach has solid theoretical foundations, produces intuitive, readily-understood measures of significance, and remains as valid today as it did before it was eclipsed by the flawed attempts of Fisher *et al.* to create an objective theory of statistical inference. It is known as Bayesian inference, after the 18th Century English cleric Thomas Bayes who first published the key theorem behind it: Bayes's theorem.

The power and importance of this theorem is immediately apparent in its solution to one of the central problems of standard statistical inference. As we have seen, frequentist methods do not tell us Prob(theory | data); that is, they do not tell us what our belief in a theory should be, given the data we actually saw. To

answer that question, we must turn to the axioms of probability theory, from which we find that (see, e.g. Feller 1968 Ch 5):

$$\text{Prob}(A \mid B) = \text{Prob}(B \mid A).\text{Prob}(A)/\text{Prob}(B) \quad (1)$$

This is Bayes's theorem, which becomes the basis of Bayesian inference when A is the event of a specific hypothesis being true, and B as the event of observing specific data. Bayesian inference was the standard means of performing statistical inference prior to Fisher's work in the 1920s, and it allows us to calculate a clear and unambiguous measure of support for a theory, Prob(theory | data) directly from experimental results via the relationship:

$$\text{Prob}(\text{theory} \mid \text{data}) = \text{Prob}(\text{data} \mid \text{theory}).\ \text{Prob}(\text{theory})/\text{Prob}(\text{data}) \quad (2)$$

This formulation of Bayes's theorem shows clearly that while we can calculate the quantity we are interested in, namely Prob(theory | data), this is not equivalent to Prob(data | theory), much less to a P-value. However, the formula also highlights the key stumbling-block to the application of Bayesian inference. To work out the value of Prob(theory | data), we must first establish Prob(theory); that is, we must be able to put some prior probability on the theory we are testing. As I shall show later, setting this prior probability is often far less problematic than some critics claim: it is rare that there are absolutely no previous findings or plausibility arguments available to constrain our estimate. It remains true, nevertheless, that in those cases where there is a complete absence of any previous results or insight, the prior probability of the correctness of the hypothesis will be based largely on opinion. In short, it will be *subjective*.

It is this unequivocal use of subjectivity that has made Bayesian inference so controversial, and has led to such determined attempts to find alternatives. As we have seen, working scientists may routinely use subjectivity when it suits them, but the idea of explicitly incorporating it into the very heart of data analysis remains anathema. But this attitude overlooks a striking fact about the scientific process: that all attempts to rid it of subjectivity have failed. By the usual standards of scientific research, the repeated failure of these attempts would be taken to imply that the basic thesis was flawed. And from (2) we now see that this would, indeed, be the correct conclusion to draw. For the axioms of probability, via Bayes's theorem, show that subjectivity cannot be wrung out of the scientific process for the simple reason that it is mathematically *ineluctable*. Much as we might want to, it is *impossible* to obtain the value of Prob(theory | data) without having some value for the prior probability Prob(theory).

The plain fact is that subjectivity in statistical inference is as unavoidable as uncertainty in quantum mechanics. Yet while we have all grown accustomed to the latter – not least because of the welter of theoretical and empirical support for its existence – there remains a deep-seated reluctance to embrace the presence of subjectivity in scientific research.

We have seen that this reluctance stems in part from concern about playing into the hands of the enemies of science, and also from past abuses in the application of subjectivity. Further barriers exist to the adoption of Bayesian methods in data analysis, however. Some of these are entirely pragmatic: it is undoubtedly harder to boil down Bayesian inference to the same 'cook-book' approach used in standard frequentist methods. Except in simple cases, Bayesian inference is also more mathematically and computationally demanding than frequentist methods. The dearth of textbooks and software suitable for the non-specialist wanting to carry out real-life data analysis does nothing to help (see, however, O'Hagan 1997).

None of this would matter, however, were the working scientist convinced that the effort involved in getting to grips with Bayesian methods was worthwhile. This leads one to suspect that there are other, more fundamental reasons for the failure of Bayesian inference to regain its primacy over frequentist methods.

First, advocates of Bayesian inference have failed to tackle the widely held belief that Bayesian prior probabilities are never more than wholly subjective guesses, 'plucked out of the air' to suit some or other prejudice or preconception. It cannot be stressed too highly that only rarely will there be *absolutely nothing* on which to base a reasonable prior. In many cases, there will be sources of evidence on which to base a sensible prior probability: for example, results from previous studies of similar drugs and plausibility arguments concerning, say, cancer risks from radiation based on insights from physics. Even if there really is little solid evidence on which to base a prior probability, Bayesian inference can still provide insight by allowing one to study the effect of different levels of prior belief (see, e.g. Spiegelhalter *et al.* 1994). It is

also possible to invert Bayes's theorem, and estimate what prior belief is needed for data to reach a given level of plausibility; I give examples of such 'inverse Bayesian inference' below.

The second key feature of Bayesian inference that is not sufficiently appreciated is that initial prior beliefs in a specific hypothesis become progressively less important as data accumulate. It can be shown mathematically (see, e.g. O'Hagan 1994 p 74 et seq.) that whatever prior probability is used at the outset, Bayes's theorem ensures that everyone is driven towards the same conclusion as the data accumulate. Unless one's prior is precisely zero (which is not a rational stance), the only long-term effect of the prior belief is that a sceptic starting from a low prior probability will require more data to reach the same level of belief as an enthusiast for the theory – which is hardly an egregious feature of a theory of inference. Indeed, it is striking that this mathematical feature of Bayesian inference mirrors so well how science actually operates. Starting from a wide variety of opinions about, say, the link between some chemical and cases of cancer, the accumulation of experimental and epidemiological evidence drives the scientific community toward the same conclusion about the reality or otherwise of the link, with sceptics merely taking longer to be convinced.

In short, Bayesian inference provides a coherent, comprehensive and strikingly intuitive alternative to the flawed frequentist methods of statistical inference. It leads to results that are more easily interpreted, more useful, and which more accurately reflect the way science actually proceeds. It is, moreover, unique in its ability to deal explicitly and reliably with the provably ineluctable presence of subjectivity in science.

These features alone should motivate many working scientists to find out more about applying Bayesian inference in their own research. For those who still need to be convinced, however, I now demonstrate perhaps the most impressive reason for using Bayesian inference: its ability to provide a far greater level of protection than frequentist methods against seeing significance in entirely spurious research findings. For as we shall see, while frequentist methods are still widely used within the scientific community, they routinely exaggerate the real significance of implausible data, with results that can and do bring the scientific process into disrepute.

How P-values exaggerate significance

As we have seen, frequentist methods of inference provide measures of significance that are neither objective nor intuitive. More importantly, however, they give a fundamentally misleading view of the significance of data. To see this, take the simple case in which a hypothesis is to be tested via measurements of a specific parameter, θ; for example, the hypothesis may be that a toxin is linked to some disorder in children, so that θ is the level of this toxin in children suffering from the disorder. Such an investigation would then consist of measuring values of θ in a group of affected children, θ_i, computing the data mean and variance, and comparing it with θ_0, the value of θ found among normal children. We would then test the null hypothesis that any difference we find is merely the result of chance by setting up a test-statistic, z, which takes into account the sample size, its mean and variance, and compares it to θ_0, the value expected if the null hypothesis is correct. Following the frequentist approach, one would typically convert this z-score to a P-value, the probability of obtaining at least as large a value of z, *assuming* the null hypothesis that chance alone is the cause. According to convention, if the P-value is less than 0.05, then the data are taken to be significant.

However, as we have seen, a much more meaningful measure of significance is Prob(Null hypothesis | data), the probability that the difference in θ *really is* the product of chance alone. Just how big is the disparity between this measure of significance and the frequentist P-value? To find out, we can use Bayes's theorem (2), which with a little algebra becomes

$$\text{Prob(Null hypothesis | data)} = \left(1 + \frac{1 - \text{Prob(Null)}}{\text{Prob(Null).BF}} \right)^{-1} \qquad (3)$$

where Prob(Null) is the prior probability for the null hypothesis that there is no real difference in the toxin level in the children, and BF is the so-called Bayes Factor, which measures how much we should alter our prior belief about the null hypothesis in the light of the new data, as captured by z. For the value of the Bayes Factor, one can show (see, e.g. Lee p131) that under very general conditions BF has a *lower* limit of

$$BF \geq \exp(-z^2/2) \qquad (4)$$

As an example, suppose that past evidence concerning the toxin leads us to an agnostic view of the possibility that there are higher levels of the toxin in the children with the disorder; this is equivalent to setting Prob(Null) = 0.5. Inserting this and (4) into (3) we find that, for a given value of z, our initial agnosticism leads us to a probability that the null hypothesis of no real difference is indeed correct of *at least*

$$\text{Prob(Null | data)} \geq \{1 + \exp(z^2/2)\}^{-1} \qquad (5)$$

Suppose, for example, that the measurements of the toxin levels in the two groups revealed a difference with a z-value of 2.0. On the frequentist viewpoint, standard statistical tables shows that this implies a P-value of 0.044; as this is less than 0.05, the difference is deemed significant at the P = 0.05 level. As we have stressed, however, this does *not* mean that the probability that the difference *really is* a fluke is also 0.044; we can only calculate this latter probability via Bayes's theorem. Plugging in z = 2 into (5), we find that our data actually imply that Prob(null | data), the probability the difference is just a fluke, is *at least* 0.12. In other words, while the frequentist methods led us to conclude that the difference was significant, the Bayesian calculation pointed to a much higher probability of the finding being a mere fluke.

This conclusion, moreover, was based on an agnostic prior of Prob(Null) = 0.5. If there are no strong grounds for believing that the effect is genuine, then – in contrast to frequentist methods – Bayesian inference allows us to factor in this lack of plausibility explicitly into our analysis. This can have particularly dramatic effects in the assessment of "anomalous" phenomena (Matthews 1998), as the following example shows (Nelson 1997).

For over 250 years, Princeton students have attended Commencement on a Tuesday in late May or early June, an outdoor event for which good weather is vital. According to local folklore, good weather does usually prevail, prompting claims that those attending may 'wish' good weather into existence. By analysing local weather records spanning many decades, Nelson found that Princeton's weather was generally no different from that of its surroundings. However, he did find some evidence that the town was less likely to be rained on during the outdoor events. The phenomenon gave z-scores as high as 1.996, which on a frequentist basis gives a significant P-value of 0.046. Properly mindful of the implausibility of the phenomenon, however, Nelson was reluctant to take this objective finding at face value, and instead reached a more subjective conclusion: "These intriguing results certainly aren't strong enough to compel belief, but the case presents a very challenging possibility".

A Bayesian analysis allows a far more concrete assessment of plausibility to be made. Clearly, with such a bizarre claim, there is little one can say about the precise value of a sensible prior probability for the null hypothesis of no real effect, other than to say that the probability is likely to be pretty high. In such cases, Bayesian inference still gives valuable insight, as it allows one to estimate the level of prior probability necessary to sustain a belief that the effect is illusory, even in the light of Nelson's data. Using (4) and (3) and z = 1.996, this inverse Bayesian inference shows that Prob(Null | data) > 0.5 for all Pr(Null) > 0.88 In other words, for anyone whose prior scepticism about the effectiveness of wishful thinking exceeds 90 per cent, the balance of probabilities is that the effect is illusory, despite Nelson's data.

As this example shows, frequentist methods greatly exaggerate the significance of intrinsically implausible data. However, as we shall now see, frequentist methods can also seriously exaggerate both the size and significance of effects in much more important mainstream areas of research, such as clinical trials.

Misleading significance of clinical trial results

Misleading P-values

The classic method for investigating the efficacy of a new drug or therapy, or the impact of exposure to some risk-factor, is the so-called randomised case-control clinical trial. In such trials, a group of people given the new treatment or exposed to the risk-factor are compared with an unexposed control group. One common frequentist method of analysing the outcome is to reduce the results to a test-statistic (such as χ^2), which is then turned into a P-value; as before, if this is less than 0.05, then the difference between the two groups is deemed to be significant. Again, however, a Bayesian analysis reveals that the real significance of such a finding is typically much less impressive than the P-values imply.

As before, I shall demonstrate this by taking a real-life case. During the early 1990s, research emerged to suggest that the risk of coronary heart disease (CHD) is associated with childhood poverty (Elford *et al.* 1991). Following the discovery that infection with the bacterium *H. pylori* is also linked to poverty, some researchers suspected that the bacterium may form the missing link between the two. Precisely how a

bacterium in the stomach might cause heart disease is less than clear – raising the key issue of plausibility, to which we shall return shortly. Nevertheless, a number of studies were undertaken to investigate the link between CHD and *H. pylori.* In one of the first such studies (Mendall *et al.* 1994), 60 per cent of patients who suffered CHD were found to be infected with *H. pylori*, compared with 39 per cent of normal controls. When the effects of age, CHD risk factors and current social class had been controlled for, the results led to a χ^2 value of 4.73. Using frequentist methods, this leads to a P-value of 0.03, implying that the rate of CHD among those infected with *H. pylori* is significantly higher than those without.

On the face of it, this finding raises the intriguing prospect of being able to tackle one of the major killers of the western world using nothing more than antibiotics. Yet while the evidence that both CHD and *H. pylori* infection are more common among the poor is suggestive of a link between the two, it is hardly unequivocal. Such scepticism is underscored by the lack of any convincing mechanism by which a gastric bacterium could trigger heart disease. The frequentist P-value, however, cannot reflect any of these justifiable qualms; sceptics of the link have no option but to say that on this occasion they are just going to ignore the supposed significance of Mendall *et al.*'s finding.

In contrast, Bayesian inference requires no such arbitrary 'moving of the goalposts': it allows explicit account to be taken of the plausibility of the findings. In the case of the supposed link between CHD to *H. pylori*, the lack of any convincing mechanism balanced against the socio-economic evidence of a link suggests that an agnostic prior probability of Prob(Null) = 0.5 would be a reasonable starting-point for assessing results like those found by Mendall *et al.* Inserting this into (3) implies that the probability of the results being due to chance, given the observed data, is

$$\text{Prob(Null | data)} = \text{BF}/(1 + \text{BF}) \qquad (6)$$

where BF is the Bayes Factor for the null hypothesis of chance effect. One can show that for in a wide range of practical situations, including this type of case-control study, the *lower* bound on BF is given by (see, e.g. Berger & Sellke 1987)

$$\text{BF} \geq \sqrt{(\chi^2)}.\exp[(1- \chi^2)/2] \qquad (7)$$

Inserting the value of $\chi^2 = 4.73$ found by Mendall *et al.* into (6) shows that the BF is *at least* 0.337. Putting this in (6) we find that Prob(Null | data), the probability that Mendall *et al.*'s results are due to nothing more than chance is *at least* 0.25. In other words, even using an agnostic prior, the frequentist P-value has over-estimated the real significance of the findings by almost an order of magnitude.

Those taking a more sceptical view of a link between a gastric bacterium and CHD would, of course, set Prob(Null) somewhat higher. Applying the concept of inverse Bayesian inference used earlier, it emerges that even a relatively modest sceptical prior of just Prob(Null) = 0.75 is enough to lead to a balance of probabilities that Mendall *et al*.'s findings are entirely illusory.

Misleading Confidence Intervals

Some defenders of frequentist methods regard criticism of P-values as an attack on a straw man, pointing out that P-values are increasingly being supplanted by 95 per cent confidence intervals (CIs), which convey more information about effect size than a single-figure P-value. Yet as we have seen, frequentist CIs share many of the same problems of interpretation as P-values. Most importantly, they also share an inability to take into account the plausibility of the hypothesis under test. As such, 95 per cent confidence intervals are also prone to exaggerate both the size and the significance of intrinsically implausible effects.

In contrast – and as one might expect by now – the Bayesian counterpart of CIs (known as Credible Intervals or Highest Density Regions), are more comprehensible, more meaningful and more reliable indicators of real significance. With frequentist CIs, the 95 per cent refers to the reliability of the statistical test; the Bayesian CI, in contrast, means precisely what it seems to mean: that there is a 95 per cent probability that the true value of the parameter lies within the stated range.

As already noted, Bayesian CIs are numerically identical to their frequentist counterpart if there is only very vague prior knowledge about plausible values of the parameter of interest (see e.g. Berger & Delampady 1987 p 328, and Appendix to this paper). However, such complete ignorance about the likely size of the effect under study is rarely defensible, and in general frequentist and Bayesian CIs will not coincide. In such

cases, a Bayesian CI is always a more reliable guide to the true significance of a finding than its frequentist counterpart.

Again, let us illustrate this through a real-life example. In the early 1990s, the Grampian region early anistreplase trial study (GREAT Group, 1992) generated considerable interest in the medical community, as it seemed to show that heart-attack victims given this clot-busting drug at home had a 50 per cent higher chance of survival than those given the drug once they arrived in hospital. While there were good reasons for expecting that early intervention with the drug would produce some improvement, the size of the claimed benefit surprised many. Nevertheless, frequentist measures of significance appeared to give objective support to the finding: the team found a relative risk (RR) of death for those given the drug early of 0.52 – i.e. a 48 per cent risk reduction – with a 95 per cent CI of (0.23 0.97). As this excludes an RR of 1, this surprising result is also significant in frequentist terms, the equivalent P-value being 0.04.

However, as was pointed out shortly after the publication of the GREAT results (Pocock and Spiegelhalter 1992), a considerable amount of prior information existed with which to assess the plausibility of the GREAT finding; for example, a much larger European study involving the same drug pointed to a much smaller benefit. Drawing on this existing knowledge, Pocock and Spiegelhalter carried out a Bayesian re-assessment of the GREAT results; an outline of how such an analysis can be performed is given in the Appendix to this paper. The prior information was captured through a probability distribution which peaked at an RR of 0.83 while giving low probabilities to RRs greater than 1.0 (no benefit) or less than 0.6 (dramatic improvement). When combined with the GREAT data, the resulting ('posterior') probability distribution peaked at an RR of around 0.75, with a 95 per cent Bayesian CI of (0.57 1.0). While still pointing to a more impressive effect than that suggested by previous studies, the GREAT results emerge from the analysis as markedly less impressive than suggested by the frequentist methods.

At this point, it is natural to ask whether this Bayesian analysis really did give a more accurate picture of reality than the frequentist methods. The simple answer is yes. Six years after the publication of the GREAT findings, the overall picture emerging from international studies is that early use of clot-busters like anistreplase does indeed confer extra benefit, with RRs of around 0.75 to 0.8 (Fox, quoted in Matthews 1997). This is only half the improvement suggested by the frequentist analysis of the GREAT data, but in impressive agreement with Pocock and Spiegelhalter's Bayesian analysis.

In a similar vein, the current consensus concerning the supposed *H. pylori*-CHD link is that a plausible mechanism relating the two is lacking, and that a causal link remains dubious (Danesh *et al*. 1997). This suggests that the basis of the above Bayesian analysis of the supposed link remains valid – a conclusion supported by a recent large-scale study that failed to find any convincing evidence for an association (Wald *et al*. 1997).

These cases are hardly the only examples of the tendency of frequentist methods to exaggerate both effect size and significance of clinical findings. Undoubtedly the most disturbing evidence comes from the continuing failure of many impressive drug trial results to produce similarly impressive results once approved for general release. It is widely recognised that most new therapies for cancer and heart disease have proved far less effective than initially believed (e.g. Fayers 1994, Yusuf *et al.* 1984). Very recently, a UK study uncovered evidence that the use of "clinically proved" drugs for myocardial infarction since the early 1980s has had no effect on mortality, with death-rates on the wards at least double those found in trials (Brown *et al* 1997).

Such a finding would come as no surprise to those familiar with the inherent ability of frequentist methods to exaggerate both effect sizes and significance. It is of course perfectly possible that at least part of the explanation for such disappointing findings lies elsewhere: the greater care taken of all patients in clinical trials, for example, and the fact that trials tend to be conducted in centres of excellence. Brown *et al.* suggest that their disappointing findings may be due to a failure to optimise the use of the available treatments for myocardial infarction. This highlights another factor in the continuing failure of Bayesian methods to supplant frequentist methods: the existence of many other apparently plausible explanations capable of masking the failings of frequentist methods.

'Explaining away' frequentist failures

The most common explanation for studies whose spuriously significant findings fail to be confirmed is that the sample size was too small. This seems plausible enough: after all, everyone knows that the smaller a sample, the less reliable its conclusions. Yet the argument overlooks two key facts. First, the calculation of a

P-value takes full account of sample size. On the frequentist viewpoint, we must regard a P-value of 0.03 as significant whether it is based on a sample of 10 or 10,000 people; larger samples are just more likely to detect significance in smaller effects. And this is related to the second flaw in the sample size defence of frequentist failures. Small samples are indeed more susceptible to statistical noise than large ones, but only in the sense that their lack of statistical power makes them more prone to missing real effects. For a given P-value, both small and large studies of the same quality are equally likely to see significance in results that are really due to chance. As such, blaming the failure of large studies to replicate significant positive findings from smaller studies purely on sample size is simply fallacious.

A more sophisticated, and plausible, defence of frequentist failures is that the original studies were undermined by biasing and confounding factors. Bias undermines the separation of subjects into cases and controls, due to, say, misdiagnosis of the disease whose cause is under investigation. Confounding undermines attempts to link a cause to its effects; for example, failure to take into account dietary differences can undermine attempts to link carcinogens to observed cases of cancer.

Both bias and confounding are exceptionally difficult to deal with, and undoubtedly explain many failures to replicate results. For example, when Mendall *et al.* applied further controls for the confounding effect of overcrowding and hot water supplies in childhood risk-factors for infection by *H. pylori*, the link between the bacterium and CHD remained, but its P-value was no longer significant.

The undoubted power of bias and confounding to undermine clinical research findings has provided defenders of frequentist methods with a further reason for shunning Bayesian inference. The argument is that while Bayesian methods may indeed deal more effectively with the risk of seeing significance in fluke results, it is no better at dealing with bias and confounding than the standard frequentist methods, and these are typically far more important.

This is also incorrect. Even relatively simple Bayesian analysis does allow concern about bias and confounding to be taken into account, via the form of the prior probability distribution, in the assessment of the posterior probability. Similar remarks apply to the supposed inability of Bayesian methods to take into account the many other potential influences on trial outcome, from poor randomisation to the better care received by patients in clinical trials. All these can be captured by a prior reflecting past real-life experience of just how successful drugs usually turn out to be.

Ultimately, however, all these supposed objections to the use of Bayesian methods serve only to conceal the key advantage of Bayesian inference: that it offers far greater protection against seeing significance in implausible results. The importance of this can best be seen through another real-life example, and one of great contemporary interest: the assessment of the risk of lung cancer faced by passive smoking of environmental tobacco smoke (ETS). The strongest evidence for this risk is generally held to be a recent meta-analysis of 37 published studies (Hackshaw *et al.*, 1997). This found a relative risk (RR) for lung cancer among life-long non-smokers living with smokers of 1.24 with a 95 per cent CI of (1.13, 1.36). A detailed assessment of both bias and confounding was carried out, but the central estimate for the RR remained essentially unchanged at 1.26 with a 95 per cent CI of (1.07, 1.47). On the basis of standard inference methods, this implies a highly significant link between passive smoking and lung cancer ($P < 0.006$). To underline the credibility of their results, Hackshaw *et al.* performed an informal plausibility assessment of their findings, using indirect measures of the likely intake of ETS by passive smokers. These suggest that passive smokers have about 1 per cent the exposure to cigarettes of their smoking partners. Assuming smokers typically consume 25 cigarettes a day, face an RR of 20 and that there is a linear dose-risk relation, Hackshaw *et al.* reached an estimate of RR ~ 1.19 for passive smokers.

While broadly similar to the RR found by the meta-analysis, this plausibility argument has itself been criticised as implausible (Lee 1998, Nilsson 1998 p 20). However, both Hackshaw *et al*. and their critics underestimate the crucial importance of a much more rigorous assessment of the plausibility of such weak results. Hackshaw *et al*. devoted about 10 times more of their paper to the assessment of bias and confounding than to plausibility; as I now show, however, a Bayesian analysis reveals that plausibility has a far more dramatic effect on the significance of the results.

Of the many criticisms that can be levelled at Hackshaw *et al.*'s plausibility argument, the most serious is their reliance on markers of ETS exposure which are both indirect and not linked to carcinogenicity. The use of such markers is especially hard to justify in the face of evidence from *direct* studies of ETS exposure that consistently point to much lower levels of exposure. An ongoing series of such studies (see e.g. Phillips *et al.*

1994, Phillips *et al.* 1998 and references therein) has found median exposures figures of ~ 0.02 cigarettes a day for the most exposed passive smokers. Even adopting the same linear dose-risk relation as Hackshaw *et al*. (which again is questionable, Nilsson 1998 pp21-22) this suggests a plausible RR for passive smoking of around 1.02, an excess risk 10 times lower than that estimated by Hackshaw *et al.* Only the top 10 per cent of the most exposed passive smokers in the studies by Phillips *et al.* were found to face anything like the risk predicted by Hackshaw *et al.*

Incorporating these results into a plausibility argument via a Bayesian prior distribution leads to an altogether different view of the risks of passive smoking. Specifically, it suggest that the excess lung-cancer risk is both 11 times smaller, with a 95% CI of (1.00. 1.04) than that given by Hackshaw *et al.,* and statistically non-significant. Bayesian inference thus strongly suggests that the growing consensus that ETS is a proven and major health risk is misplaced. Whether or not the outcome of this Bayesian analysis will be borne out is as yet unclear. What is clear is that there is a very real danger of the frequentist evidence for a 'significant extra' risk from ETS becoming canonical. This, in turn, raises the possibility that Hackshaw *et al.*'s risk figure will be used routinely to subtract out the confounding effect of passive smoking in future studies of the causes of cancer. If this risk figure has been substantially over-estimated – as the above Bayesian analysis strongly suggests it has – attempts to assess the true risk posed by many other health hazards will be seriously undermined (Nilsson 1997 p140).

This example of passive smoking and lung cancer provides the final strand in the case for the widespread and routine use of Bayesian inference in the analysis of data. This can be summed up as follows:

- It allows both previous knowledge and the inherent plausibility of a hypothesis to be explicitly taken into account;
- It gives measures of significance that are more meaningful than those generated by frequentist methods;
- These measures have more intuitive and straightforward definitions than their frequentist counterparts, and are thus much less prone to misinterpretation;
- Bayesian inference is less likely to see significance in entirely spurious findings, especially in poorly-motivated research of low inherent plausibility. As such, it provides more protection against seriously – even dangerously – misleading findings whose attempted replication or extension will ultimately prove futile.

Conclusions

In this paper, I have shown that the scientific community has a deeply ambiguous attitude towards the presence of subjectivity in research. While both desiring and proclaiming objectivity, working scientists routinely use subjective criteria in their everyday research. The justification is pragmatic, and entirely reasonable: it is impossible for working scientists to deal with the plethora of new results and theories that constantly present themselves in any other way. However, mindful of past abuses in the history of science, the scientific community remains committed to keeping the presence of subjectivity in the research enterprise to a minimum.

This commitment has led to the widespread adoption of techniques for statistical inference that appear to be objective. Known as frequentist methods, they have become central to the research enterprise, with their outcomes – P-values and 95 per cent confidence intervals – becoming a *sine qua non* for acceptance by leading science journals. As I have shown, however, these textbook methods are neither objective nor reliable indicators of either effect size or statistical significance of research findings. By failing to take into account the intrinsic plausibility of the hypothesis under test, frequentist methods are capable of greatly exaggerating both the size and the significance of effects which are in reality the product of mere chance.

The implicit recognition of these failings by the scientific community is evidenced by the way in which essentially identical results from the supposedly objective frequentist methods are interpreted in entirely different ways, according to the subjective belief of researchers. Thus, a large and highly statistically significant result in parapsychology will be ignored, while a small and statistically non-significant link between passive smoking and cancer will be deemed to add considerably to the case against environmental tobacco smoke.

The persistent failure of scientists to rid the research process of subjectivity, and the failings of frequentist techniques, can both be traced to the same fundamental source: the axioms of probability. These show that in the assessment of hypotheses, subjectivity is mathematically *ineluctable.* All attempts to banish subjectivity

from the research process are thus ultimately futile, and are at best no more than exercises in sweeping subjectivity 'under the carpet'.

The vexed problem of subjectivity in science has its solution in those same axioms, however. Bayes's theorem provides the underpinning for an entire theory of statistical inference which takes explicit account of plausibility, and supplies measures of statistical significance that are more relevant, more comprehensible and more reliable than those of frequentist methods. As such, the wider adoption of Bayesian inference will undoubtedly save substantial amounts of time, resources and public money currently spent on futile attempts to replicate significant support for intrinsically implausible hypotheses.

Some idea of the extent of this waste can be obtained by noting that each month journals covering disciplines from sociology and psychology to geology and genetics carry many papers claiming to have results significant at the 0.05 level with P-values in the range $0.01 < P \leq 0.05$. Even assuming that these claims are all sufficiently well-motivated to merit an agnostic prior, it can be shown that (6) and (7) point to *at least* a quarter of such claims are meaningless flukes (Matthews 1998). For research meriting even a very moderate level of scepticism, this proportion rapidly rises to over 50 per cent. This is a finding that should worry anyone concerned with the reliability and funding of scientific research.

The fact that just two independent clinical trials with results significant at the 0.05 level are sufficient for new therapies to win approval from national regulatory bodies is hardly less worrying. As so often with frequentist concepts, this P-value standard can and is misinterpreted as implying that the probability of the therapy being ineffective is less than 1 in 400 (see, e.g. Buyse 1994). The true proportion will be far higher, especially among therapies whose claims of efficacy are poorly motivated – a fact reflected in the many cases of where initial euphoria over some new breakthrough turns into disappointment (Yusuf *et al.* 1984, Pocock & Spiegelhalter 1992, Fayers 1994, Brown *et al.* 1997). Bayesian assessment of trial results give regulatory bodies a formal means of incorporating this crucial 'reality check' into their deliberations. In contrast, the frequentist methods currently used by regulatory bodies have no means of incorporating such key knowledge: given the same raw data, they cannot distinguish between streptokinase or snake-oil. A number of regulatory bodies will accept Bayesian assessments of drug trials; in the light of the above, the use of such methods should not be optional but mandatory.

Lack of theoretical underpinning has an especially large impact on areas of research such as parapsychology and alternative medicine. Bayesian inference applied here would certainly cast grave doubt on claims that appear impressive from a frequentist viewpoint. It is important to stress, however, that this does not imply that all research into alternative or anomalous fields should be abandoned. Bayesian inference merely implies that the standard frequentist criteria for judging statistical significance in these areas are especially inadequate. It can be shown that in such fields of research, there are few grounds for viewing as significant any result whose two-tailed P-value exceeds 0.003 (Matthews 1998). This value assumes an agnostic prior of Pr(Null) = 0.5, which is undoubtedly generous for most claims for the existence of anomalous phenomena; even so, the resulting P-value is 17 times more demanding than the conventional 0.05 criterion used for gauging significance, and it is clear that many current claims for anomalous phenomena fail to meet it.

Reputable researchers would no doubt feel more confident defending evidence for an anomalous phenomenon by applying at least a mild level of scepticism in their assessment of significance. In this case, a P-value of no more than around 2×10^{-4} is appropriate, a value 250 times more demanding than the conventional 0.05 criterion. These technical results can be stated much more succinctly, however: extraordinary claims require extraordinary evidence. This is a well-attested and widely-accepted principle, yet it is noticeable by its absence in the mathematics of frequentist inference.

It must also be emphasised that many of the concerns about frequentist inference expressed here have been recognised by leading statisticians for decades (see e.g. Jeffreys 1961, Edwards *et al*. 1963, Lindley 1970). This inevitably raises the question of why Bayesian inference is still failing to (re)gain its central role in the scientific enterprise. This is, I believe, due largely to the failure of its advocates to convey three key facts to working scientists:

- That while subjectivity may be an unwelcome feature of the scientific process, the axioms of probability show that it is unavoidable, and that Bayes's theorem is the correct way to deal with it;
- That while Bayesian inference does allow subjective prior knowledge to be incorporated into the assessment of data, such knowledge is not 'plucked out of thin air'. Rather, it allows an entirely reasonable yet crucial assessment of plausibility to be factored into the analysis.

- That, in any case, the effect of the choice of prior becomes increasingly irrelevant as data accumulates, with the only persistent effect of priors being the entirely natural one that sceptics of a specific claim require stronger evidence to reach the same level of belief than its advocates.

There is a dangerous irony in the continuing reluctance of the scientific community to adopt Bayesian inference. For this reluctance stems largely from a deep-rooted fear that adopting methods that embrace subjectivity is tantamount to conceding that the scientific enterprise really is a social construct, as claimed by the post-modern advocates of the 'anti-science' movement. The central lesson of Bayes's theorem is, however, quite the opposite. It shows, with full mathematical rigour, that while evidence for a specific theory may indeed start out vague and subjective, the accumulation of data progressively drives the evidence towards a single, objective reality about which all can agree.

It is ironic indeed that by failing to recognise this, the scientific community continues to use techniques of inference whose unreliability undermines confidence in the scientific process, and which thus threatens to deliver science into the hands of its enemies.

Appendix: Bayesian inference using confidence intervals

A growing proportion of research findings are reported via confidence intervals, in which a central parameter value, M, is accompanied by a range of values of the form (L , U), which form the so-called 95 per cent confidence interval (CI) for the results. As discussed in the main article, the frequentist interpretation of a CI is not as straightforward as it may appear: the 95 per cent figure refers to the reliability of the statistical test applied, and not to the probability that the true parameter value lies in the stated range. In contrast, a Bayesian 95 per cent CI (often also called a Credible Interval) means precisely what it seems to mean: there is a 95 per cent probability that the true value lies within the stated range.

We now outline the procedure for calculating Bayesian CIs for a given set of data. For both frequentist and Bayesian CIs, the range (L, U) is calculated from the mean parameter value, M and its standard deviation SD, via the formulas

$$L = M - 1.96.SD \qquad A1$$

$$U = M + 1.96.SD \qquad A2$$

In the textbook frequentist approach, M and SD are calculated directly from the raw data. In the Bayesian approach, however, the M and SD are the so-called 'posterior' mean and standard deviation, formed by combining the raw values extracted from the data with prior values based on extant knowledge and insight about the effect under study. The resulting posterior mean and standard deviation thus sets the new findings into their proper context, taking explicit account of their intrinsic plausibility.

The first step in a Bayesian analysis is thus to capture this prior knowledge and insight. In many real-life cases, this can be achieved by specifying a Normal distribution which peaks at the most plausible value for the parameter of interest, M_o, and whose 95 per cent 'tails' (L_o , U_o) reflect the plausible range of that parameter. The standard deviation of this prior distribution, SD_o can then be calculated from (A1), (A2):

$$SD_o = (U_o - L_o)/3.92 \qquad A3$$

The next step is to combine this prior distribution with the experimental data, whose mean is M_d and standard deviation is SD_d; the resulting posterior distribution will have a mean M_p and standard deviation SD_p. It can be shown that Bayes's theorem leads to a posterior distribution with parameters given by (see e.g. Lee 1997 Ch 2)

$$SD_p = 1/\sqrt{[1/SD_o^2 + 1/SD_d^2]} \qquad A4$$

$$M_p = (SD_p)^2 . [(M_o/SD_o^2) + (M_d/SD_d^2)] \qquad A5$$

The Bayesian 95 per cent CI then follows putting A4 and A5 into A1 and A2; the result is a range of values for the parameter of value in which the true value will lie with 95 per cent probability. Two key implications of equations A4 and A5 should be noted. First, they show that the frequentist and Bayesian definitions of the CI are equivalent only when SD_o is infinite, corresponding to a stance of complete ignorance about the plausible range of values for the parameter of interest. This is rarely justifiable, and in general the frequentist and Bayesian CIs will not coincide. Equations A4 and A5 also show that the inclusion of prior information has the effect of moving the posterior probability distribution in the direction of the prior. Thus if results from, say, a clinical trial are strikingly more impressive than seems plausible, failure to account for this lack of plausibility via a prior distribution will exaggerate both the size of the effect, and its statistical significance. As we have seen, frequentist methods cannot explicitly incorporate such plausibility arguments, and are thus especially prone to lend unjustified credibility to remarkable data.

The growing tendency to state results in terms of frequentist 95 per cent CIs does at least summarise results in a form that can easily be combined with prior knowledge using the techniques given above, as I now show.

Example: In their analysis of the GREAT study, Pocock and Spiegelhalter captured the implications of previous studies via a prior relative risk (RR) of death of 0.825, with a 95 per cent CI of (0.6, 1.0). To apply the above formulas, a logarithmic transformation has to be applied to the central estimate and range (strictly speaking, RRs should also first be transformed into a so-called Odds Ratio, but in many cases – including this one, and epidemiological studies of rare diseases such as lung cancer – the difference is immaterial). Thus we take the prior distribution to be Normal, with a peak at $\ln(RR_o)$, with its standard deviation SD_o

being calculated from A3 using the natural logarithms of the upper and lower ranges of the CI, $\ln(U_o)$ and $\ln(L_o)$. This leads to a prior distribution that peaks at $M_o = -0.19$, with a standard deviation of 0.13.

To calculate M_d and SD_d, we note that the GREAT study found a mean RR of 0.515, with a (frequentist) 95 per cent CI of (0.23, 0.97). We can convert this into the mean and standard deviation required by A4 and A5 by taking natural logarithms and using A2: this gives $M_d = -0.664$, and $SD_d = 0.367$. Using A4 and A5 we can now work out the posterior probability distribution; we find $M_p = -0.245$ and $SD_p = 0.123$.

Using A1 and A2 and then transforming back out of natural logarithms, we finally arrive at a posterior RR figure of 0.78 with a Bayesian 95 per cent CI of (0.6, 1.0). This central risk figure is substantially less impressive than the value that emerges from the raw data; this reflects the impact of the inclusion of a prior reflecting the implausibility of gaining so large a risk reduction. Furthermore, the Bayesian 95 per cent CI encompasses an RR of 1.0, which implies that the possibility that there is no benefit is not entirely ruled out by this (small) study. As discussed in the main article, the results of Pocock and Spiegelhalter's Bayesian analysis ultimately proved more realistic than those suggested by the raw GREAT data alone.

References

1. Aronson, J. L. 1984 *A realist philosophy of science* (London: Macmillan)
2. Asimov, I. 1975 *Asimov's Biographical Encyclopaedia of Science and Technology* (London : Pan Books)
3. Barber, B. 1961 Resistance by Scientists to Scientific Discovery *Science* **134** 596
4. Berger, J. & Berry, D. Statistical Analysis and the Illusion of Objectivity 1988 *American Scientist* **76** 159
5. Berger, J. & Delampady, M. 1987 Testing Precise Hypotheses *Stat. Sci.* **2** 317
6. Berger, J. & Sellke, T. Testing a point null hypothesis: the irreconcilability of P-values and evidence *J. Amer. Statist. Ass.* **82** 112 (1987)
7. Bofetta, P., Brennan, P., Lea, S. Ferro, G. 1997 Lung cancer and exposure to environmental tobacco smoke. *Biennial Report 1996/7* (Lyon: IARC/WHO)
8. Bourke, G. J., Daly, L.E. McGilvray, J. 1985 *Interpretation and uses of Medical Statistics* (3rd Edn). (St Louis : Mosby)
9. Breslow, N., Day, N. E. 1980 Statistical methods in cancer research vol. 1: The analysis of case-control studies *IARC Scientific Publication No. 32* (Lyon : IARC)
10. Brown, N., Young, T., Gray, D., Skene, A., Hampton, J.R. 1997 Inpatient deaths from acute myocardial infarction, 1982-92: analysis of data in the Nottingham heart attack register *Brit Med J* **315** 159
11. Buyse, M.E. 1994. Remarks in response to Spiegelhalter *et al*. 1994 (below); p 399
12. Collins, H. 1998 What's Wrong with Relativism? *Physics World* **11** (4) 19
13. Cooper, C. Dennison, E. 1998 Do silicone breast implants cause connective tissue disorder ? *Brit Med J* **316** 403
14. Crease, R. P., Mann, C.C. 1996 *The Second Creation* (London: Quartet)
15. Danesh, J. Collins, R. Peto, R. 1997 Chronic infections and coronary heart disease: is there a link ? *Lancet* **350** 430
16. Dunstan, D. 1998 Letter *Physics World* **11** (6) 15
17. Edwards, W. Lindman, H. & Savage, L. J. 1963 Bayesian statistical inference for psychological research. *Psychol. Rev.* **70** 193
18. Elford, J. Whincup, P. Shaper, A.G. 1991 Early life experience and adult cardiovascular disease *Intl J Epid* **20** 833
19. Fayers, P. 1994 Remarks in response to Spiegelhalter *et al*. 1994 (below); p 402
20. Feller, W. 1968 *An Introduction to probability theory and its applications* 3rd Edn. (New York: Wiley)
21. Feynman, R. P. 1985 *Surely you're joking Mr Feynman* (London: Unwin)
22. Fletcher, H. 1982 *Physics Today* June 43
23. Freedman, D. Pisani, R. & Purves, R. 1998 *Statistics* (3rd Edn.) (New York : Norton)
24. Gell-Mann, M. 1964 A schematic model of baryons and mesons *Physics Letters* **3** 214
25. Gell-Mann, M. 1994 *The Quark and the Jaguar* (London: Little, Brown)
26. Grayson, L. 1995 *Scientific Deception* (London: The British Library)
27. Grayson, L. 1997 *Scientific Deception – An Update* (London: The British Library)
28. GREAT Group 1992 Feasibility, safety and efficacy of domiciliary thrombolysis by general practitioners: Grampian region early anistreplase trial *BMJ* **305** 548
29. Greenstein, G. 1998 *Portraits of Discovery* (New York: Wiley).
30. Hackshaw, A.K. Law, M.R., Wald, N.J. 1997 The accumulated evidence on lung cancer and environmental tobacco smoke *Brit Med J* **315** 980
31. Hoffmann, B. 1975 *Albert Einstein* (London: Paladin)
32. Holton, G. 1978 Sub-electrons, Presuppositions and the Millikan-Ehrenhaft Dispute *Historical Studies in the Physical Sciences* **9** 161
33. Hellman, H. 1998 *Great Feuds in Science* (New York: Wiley)
34. Heyes, S., Hardy, M., Humphreys, P., Rookes, P. 1993 *Starting Statistics in Psychology and Education* 2nd Edn (London : Weidenfeld & Nicolson)
35. Howson, C. Urbach, P. 1993 *Scientific Reasoning* 2nd Edn (Chicago: Open Court)
36. Jeffreys, H. *Theory of Probability* 1961 (3rd Edn), (Oxford : University Press)
37. Kempthorne, O. 1971 "Probability, statistics and the knowledge business" *in Foundations of Statistical Inference* (Ed. Godambe & Sprott) (Toronto: Holt, Rinehart & Winston)
38. Lakatos, I. 1978 *Philosophical Papers* (Worrall & Currie, eds.) vol. 1 (Cambridge: the University Press) .
39. Lee, P. N. 1997 *Bayesian Statistics: An Introduction* 2nd Ed. (London : Arnold)
40. Lee, P.N. 1998 Difficulties in assessing the relationship between passive smoking and lung cancer *Stat Meth Med Res* **7** 137
41. Lindley, D. V. *Introduction to Probability & Statistics Part 2: Inference* 1970 (Cambridge: University Press)
42. Linet, M, *et al.* 1997 Residential exposure to magnetic fields and acute lymphoblastic leukaemia in children *New Eng. J Med* **337** 1

43. Macdonald, V. 1998 Official: passive smoking does not cause cancer *The Sunday Telegraph* 8 March p 1
44. Matthews, R.A.J. 1992 *Unravelling the Mind of God* (London: Virgin)
45. Matthews, R.A.J. 1997 Faith, Hope and Statistics *New Scientist* **156** 36
46. Matthews, R.A.J. 1998 The statistical assessment of anomalous phenomena *J Sci Expl* (accepted)
47. Medawar, P. 1978 *Advice to a Young Scientist* (New York : Harper & Row)
48. Mendall, M.A. *et al.* 1994 Relation between *Helicobacter pylori* infection and coronary heart disease *B Heart J* **71** 437
49. Milton, R. 1994 *Forbidden Science* (London: Fourth Estate)
50. Nelson, R. 1997 Wishing for good weather: a natural experiment in group consciousness *J. Sci. Expl.* **11** 47
51. Nilsson, R. 1997 Is environmental tobacco smoke a risk factor for lung cancer ? In *What Risk: science, politics and public health ed. Bate, R* (Oxford: Butterworth Heinemann
52. Nilsson, R. 1998 *Environmental Tobacco Smoke Revisited: The reliability of the evidence for risk of lung cancer and cardiovascular disease* (Cambridge: The European Science and Environment Forum)
53. Nyren, O. *et al.* 1998 Risk of connective tissue disease and related disorders among women with breast implants: a nationwide retrospective cohort study in Sweden *Brit Med J* **316** 417
54. O'Hagan, A. 1994 *Kendall's Advanced Theory of Statistics Vol. 2B: Bayesian Inference* (London: Arnold)
55. O'Hagan, A. *FirstBayes -freeware Bayesian inference software.* Available from
56. http://www.nott.ac.uk/maths/aoh/
57. Oppenheimer, J. R. 1955 *The Open Mind* (New York : Simon & Schuster)
58. Pais, A. 1982 *Subtle is the Lord* (Oxford: University Press)
59. Pais, A. 1991 *Niels Bohr's Times* (Oxford: Clarendon Press).
60. Phillips, K. Howard, D.A., Browne, D. Lewsley, J.M. 1994 Assessment of personal exposures to environmental tobacco smoke in British non-smokers *Environment International* **20** 693
61. Phillips, K. Howard, D.A., Bentley, M. C. Alvan, G. 1998 Measured exposures by personal monitoring for respirable suspended particles and environmental tobacco smoke of housewives and office workers resident in Bremen, Germany *Int Arch Occup Environ Health* **71** 201
62. Pocock, S. J., Spiegelhalter D. J. 1992 Letter *Brit Med J* **305** 1015
63. Popper, K. 1963 *Conjectures and Refutations* (London : Routledge)
64. Radin, D. 1997 *The Conscious Universe: the scientific truth of psychic phenomena* (San Francisco: Harper)
65. Sivia, D. S. 1996 *Data Analysis: A Bayesian Tutorial* (Oxford: University Press)
66. Spiegelhalter, D. J., Freedman, L. S., Parmar, M.K.B., 1994 Bayesian approaches to randomised trials (with discussion) *J Roy Stat Soc A* **157** 357
67. Theocharis, T. & Psimopolous, M. 1987 Where science has gone wrong *Nature* **329** 595
68. Vallance, A. K. 1998 Can Biological Activity be Maintained at Ultra-High Dilution? An Overview of Homeopathy, Evidence, and Bayesian Philosophy *J Alt Comp Med* **4** 49
69. Vandenbroucke, J.P. 1997 Homeopathy trials: going nowhere *The Lancet* **350** 824
70. Wald, N. Law, M.R. Morris, J.K. Bagnall, A.M. 1997 *Helicobacter pylori* infection and mortality from ischaemic heart disease: negative result from a large prospective study *Brit Med J* **315** 1199
71. Weinberg, S. 1993 *The Discovery of Sub-atomic particles* (London: Penguin)
72. Williams, T. 1994 *Biographical Dictionary of Scientists* (London: Collins)
73. Wolpert, L. 1992 *The Unnatural Nature of Science* (London: Faber)
74. Wolpert, L., Richards, A. 1989 *A Passion for Science* (Oxford: the university press)
75. Yusuf, S., Collins, R., Peto, R. 1984 Why do we need some large, simple randomized trials ? *Statistics in Medicine* **3** 409

Acknowledgements
In arriving at the arguments presented in this Working Paper, I have benefited enormously from the assistance, advice and comments of many researchers. I should particularly like to thank David Balding of Reading University, James Berger of Purdue University, Colin Howson of the London School of Economics, Robert Nilsson of Stockholm University, and Stuart Pocock and Ian White of the London School of Hygiene and Tropical Medicine. I am also most grateful to Roger Bate of ESEF for inviting me to collect together my thoughts in this Working Paper. I would also like to thank the three anonymous referees who provided useful criticisms of the paper. My biggest debt, however, is to Dennis Lindley, from whom – like so many others working in this field – I have learned so much.

The author
Robert Matthews is Visiting Fellow in the Neural Computing Research Group at Aston University, Birmingham. A graduate in physics from Oxford University, he has published many research papers in fields ranging from astrodynamics and probability theory to the statistical analysis of anomalous phenomena and the mathematical basis of 'urban myths'. A Fellow of the Royal Statistical Society and Royal Astronomical Society, he also acts as science correspondent for *The Sunday Telegraph,* London. His website is at http://www.ncrg.aston.ac.uk/People/index.html

Academic Members of ESEF
August 1998
Prof. Tom Addiscott UK
Prof. Bruce Ames USA
Dr Sallie Baliunas USA
Dr Alan Bailey UK
Dr Robert C. Balling USA
Prof. A. G. M. Barrett UK
Dr Jack Barrett UK
Mr Roger Bate UK
Dr Sonja Boehmer-Christiansen UK
Prof. Dr Frits Böttcher The Netherlands
Prof. Norman D. Brown UK
Prof. Dr K. H. Büchel Germany
Dr John Butler UK
Mr Piers Corbyn UK
Prof. Dr. A. Cornelissen The Netherlands
Dr Barrie Craven UK
Mr Peter Dietze Germany
Dr A. J. Dobbs UK
Dr John Dowding UK
Dr John Emsley UK
Dr Patricia Fara UK
Dr Oeystein Faestoe UK
Dr Frank Fitzgerald UK
Prof Dr Hartmut Frank Germany
Dr James Franklin Belgium
Dr Alastair Gebbie UK
Dr T. R. Gerholm Sweden
Prof. Dr Gerhard Gerlich Germany
Prof. D. T. Gjessing, Norway
Dr Manoucher Golipour UK
Dr Adrian Gordon Australia
Dr Vincent R. Gray New Zealand
Dr Gordon Gribble USA
Prof Dr Hans-Eberhard Heyke Germany
Dr Vidar Hisdal Norway
Dr Jean-Louis L'hirondel France
Dr Sherwood Idso USA
Dr Antoaneta Iotova Bulgaria
Prof. Dr Zbigniew Jaworowski Poland
Dr Tim Jones UK
Prof. Dr Wibjörn Karlén Sweden
Dr Terrence Kealey UK
Prof. Dr Kirill Ya.Kondratyev Russia
Prof. Dr F. Korte Germany
Mr Johan Kuylenstierna Sweden
Dr Theodor Landscheidt Germany
Dr Alan Mann UK
Dr John McMullan UK
Prof. Dr Helmut Metzner Germany
Dr Patrick Michaels USA
Sir William Mitchell UK
Dr Paolo Mocarelli Italy
Dr Asmunn Moene Norway
Dr Brooke T. Mossman USA
Prof Dr Hans-Emil Müller Germany
Prof Dr Dr Paul Müller Germany
Dr Joan Munby UK
Mr Liam Nagle UK
Dr Genrik A. Nikolsky Russia
Dr Robert Nilsson Sweden
Prof. Dr Harry Priem The Netherlands
Dr Christoffer Rappe Sweden
Dr Ray Richards UK
Dr Michel Salomon France
Dr Tom V. Segalstad Norway
Dr S. Fred Singer USA
Dr Willie Soon USA
Dr G. N. Stewart UK
Dr Gordon Stewart UK
Dr Maria Tasheva Bulgaria
Dr Wolfgang Thüne Germany
Dr Alan Tillotson UK
Dr Brian Tucker Australia
Prof. Dr med. Karl Überla Germany
Prof. Dr H. P. van Heel The Netherlands
Dr Robin Vaughan UK
Prof. Nico Vlaar The Netherlands
Dr Horst Wachsmuth Switzerland
Dr Michael P. R. Waligórski Poland
Dr Gunnar Walinder Sweden
Dr Gerd-Rainer Weber Germany
Prof Donald Weetman UK
Dr Charlotte Wiin-Christensen Denmark
Dr Aksel Wiin-Nielsen Denmark
Dr James Wilson USA

Business Members
Dr Alfred Bader UK
Mr John Boler UK
Mr Charles Bottoms UK
Dr Francisco Capella Gómez-Acebo Spain
Mr Richard Courtney UK
Mr Michael Gough USA
Dr Claes Hall UK
Mr Richard Hallett UK
Mr Peter Henry UK
Mr Holger Heuseler Germany
Mr Graham Horne
Dr Warwick Hughes Australia
Dr Kelvin Kemm South Africa
Mr Peter Plumley
Dr John Rae UK
Dr Michael Rogers UK
Mr Peter Toynbee Australia
Dr Wynne Davies UK

Mission Statement
The European Science and Environment Forum is an independent, non-profit-making alliance of scientists whose aim is to ensure that scientific debates are properly aired, and that decisions which are taken, and action that is proposed, are founded on sound scientific principles.

The ESEF will be particularly concerned to address issues where it appears that the public and their representatives, and those in the media, are being given misleading or one-sided advice. In such instances the ESEF will seek to provide a platform for scientists whose views are not being heard, but who have a contribution to make.

Members are accepted from all walks of life and all branches of science. There is no membership fee. Members will be expected to offer their services in contributing to ESEF publications on issues where their expertise is germane.

Purpose of ESEF
The European Science and Environment Forum is a Charitable Company Limited by Guarantee (No.1060751). It was established in 1994 to inform the public about scientific debates. Our chosen method for achieving this objective is to provide a forum for scientific opinions that are usually not heard in public policy debates.

Our primary role is to provide an independent voice to the media, the general public and the educators, and by doing so, we aim to provide balance on scientific issues. A secondary role is to contribute to the scientific debate itself. Many of our authors will simplify papers that they originally wrote for the peer reviewed scientific literature. ESEF's tertiary role is to advise scientists how to present their findings to the media, and how the media will perceive, and may use, the information. We hope that this will provide dialogue and understanding between these two important institutions.

How was ESEF formed?
ESEF was formed in 1994 by Roger Bate (Director of the Environment Unit at the Institute of Economic Affairs, London), Dr John Emsley (Science Writer in Residence at Imperial College London University) and Professor Frits Böttcher (Director of the Global Institute for the Study of Natural Resources in The Hague). The issue of climate change was the initiation for the meeting. All three thought that the debate had been unduly one-sided and they wanted to provide a forum for scientists to publish their arguments for public consumption. The media, and via them the public, tended to only hear the so-called consensus view presented by government and intergovernmental science panels.

Of course climate change is not the only issue where member scientists consider that the media debate is not balanced and that there are many environmental and public health issues which are not fully discussed in the public arena either.

ESEF decided on a mission to provide the media and the public with accessible first-hand research of leading scientists in their fields, as an alternative to reports received from specialist journals, government departments or single-issue pressure groups. As Einstein is reputed to have said: "Make science as simple as possible, but no simpler". Our aim is to provide science simplified as far as possible. Our members are from fields as diverse as nuclear physics, biochemistry, glaciology, toxicology and philosophy of science. We intend on liaising between the media and our expert members to provide an independent voice on subjects germane to various public policy debates.

To maintain its independence and impartiality, ESEF accepts funding only from charities, and the income it receives is from the sale of its publications. Such publications will automatically be sent to members. Copies will be sent to selected opinion formers within the media and within government.