The Limits of Statistical Methodology: Why A "Statistically Significant" Number of Published Scientific Research Findings are False

Joseph Wayne Smith

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies [1–3] to the most modern molecular research [4,5]. There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims [6–8]. However, this should not be surprising. It can be proven that most claimed research findings are false. Here I will examine the key factors that influence this problem and some corollaries thereof.

The Essay section contains opinion pieces on topics of broad interest to a general medical audience.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research is not most appropriately represented and summarized by p-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on p-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread. However, here we will target relationships that investigators claim exist, rather than null findings.

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a 2 × 2 table in which research findings are compared against the gold standard of true relationships in a scientific field. In a research field both true and false hypotheses can be made about the presence of relationships. Let R be the ratio of the number of "true relationships" to "no relationships" among those tested in the field. R is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2 × 2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2 × 2 table, one gets PPV = $(1 - \beta)R/(R - \beta R + \alpha)$. A research finding is thus

Citation: Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.

Abbreviation: PPV, positive predictive value

John P. A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, United States of America. E-mail: jioannid@cc.uoi.gr

Competing Interests: The author has declared that no competing interests exist.

DOI: 10.1371/journal.pmed.0020124

(Ioannidis, 2005a)

1. Introduction

In 2005, John Ioannidis published a now widely cited paper, "Why Most Published Research Findings are False," in which he pointed out that in many sciences there is a high rate of non-replication and also a high rate of failure of confirmation, due to a number of factors (Ioannidis, 2005a, 2005b.) In the next section, I'll discuss the basic

problems with the methodology of statistical significance. One principal reason for holding that most published research findings are false, is basing research on a single study assessed by the methodology of statistical significance, with a *p*-value less than 0.05. But Ioannidis also mentioned other factors that collectively would lead one to conclude that most published research findings are false, such as the use of unreasonably small samples, and outright fraud, which has been found to be more common that one would have suspected, or feared (Smith & Smith, 2023a).

Since the publication of Ioannidis's paper, and even before that (see, e.g., de Long & Kang, 1992), there have been other papers published also proposing that "most published research findings are false" (Tabarrok, 2005; Moonesinghe et al., 2007; Diekmann, 2011; Freedman, 2010). As well, running parallel to this issue, there has been deep concern in the literature about a "reproducibility crisis" in psychology and other sciences (Yong, 2015; Simmons, 2011). For example, in an attempt to replicate results in 98 original papers in three psychology journals, one research team found only 39 of 100 replication attempts successful (with two replication attempts duplicated by separate research teams) (Open Science Collaboration, 2015). While 97 percent of the original studies found significance, only 36 percent of the replications found significance (Open Science Collaboration, 2015).

Matters are even worse in cancer biology research, where only six of 53 high-profile peer-reviewed papers could be replicated, the problem arising from the fact that the basic cell line animal models themselves were inadequate (Begley & Ellis, 2012). Further, similar problems have been found in neuroscience and genetics research. Button et al. have concluded:

> [T]he average statistical power of studies in the neurosciences is very low. The consequences of this include overestimates of effect size and low reproducibility of result. (Button et al., 2013: p. 365)

The same is true of genetics research (Ioannidis & Trikalinos, 2005). In general, "the cumulative (total) prevalence of irreproducible preclinical research exceeds 50 %," with the estimated range being from 51 to 89 percent (Ioannidis, 2008; Freedman et al., 2015; Hartshorne et al., 2012; Everett & Earp, 2015).

There is no doubt, as Button (et al. 2013) note, that small sample sizes in research is one factor undermining the reliability of such research. But as Higginson and Munafò have argued, the "institutional incentive structure of academia," and the "publish or perish" mentality, especially the desire for publications in journals with a High Impact Factor (IF), leads researchers to pursue small samples, in order to get publishable results quickly, and maintain the continuity of their careers (Higginson & Munafò, 2016). They show, using an ecological model, that scientists seek to maintain their "fitness" (academic survivability) and thus would conduct research producing novel results with small studies in order to publish quickly and reduce research costs, having only 10-40 percent statistical power. Thus, roughly half of published studies in the sciences will be false, with erroneous conclusions (Higginson & Munafò, 2016).

Richard Horton, writing in *The Lancet*, lamented the precarious state of scientific research:

> The case against science is straightforward; much of the scientific literature, perhaps half, may be simply untrue. Afflicted by studies with small sample sizes, tiny effects, invalid exploratory analyses, and flagrant conflicts of interest, together with an obsession for pursuing fashionable trends of dubious importance, science has taken a turn towards darkness. (Horton, 2015: p. 1380)

Horton observes that scientists no longer have an incentive to be "right" in the disinterested pursuit of truth, since academic incentives reward only those who are innovative and productive, however wrong they might be. Ironically, as shown by the "Matthew effect" (Merton, 1968), the professional academic establishment might be, perhaps, implicitly aware of this problem, because in one experiment, papers that had previously been published were resubmitted to journals under different titles. The majority were rejected, not because prior publication was detected, but because of the poor quality of the papers. Yet, the errors were not originally detected (Peters & Ceci, 1982).

Worse still, reviewers were found in one study to have failed to detect all errors deliberately inserted into a paper for review—and the reviewers were peer-review experts in that field (Godlee et al., 1998; Jefferson, et al., 2002). R. Smith, commenting in the *Journal of the Royal Society of Medicine*, concluded that scientific peer review—a review by experts—is a process merely based on "belief" (faith), not strict rationality:

> So, peer review is a flawed process, full of easily identified defects with little evidence that it works. Nevertheless, it is likely to remain central to science and journals because there is no obvious alternative, and scientists and editors have a continuing belief in peer review. How odd that science should be rooted in belief. (Smith, 2006)

Moreover, scientific experts are biased in many ways, including selectively reporting data (Ioannidis et al., 2014); and even outright fraud and the use of "false" data is more frequent than is often thought by mainstream scientists (Martin, 1992; Vogel, 2011). The limitations of peer review, and the reproducibility crisis, have been discussed by me in other papers (Smith, 2023; Smith & Smith, 2023a, 2023b), and an ingenious response to the reproducibility/replication crisis has been given by Robert Hanna (Hanna, 2023a, 2023b), namely, that properly conducted empirical scientific research does not require reproducibility/replication anyway. It remains to see whether mainstream social and biological scientists take up this idea.

However, beyond these larger issues, the focus of the rest of the paper will be on another issue mentioned by Ioannidis, problems with statistical methodology. It will be argued that critics such as Gigerenzer are right to suppose that problems such as the reproducibility/replication crisis—assuming that reproducibility/replication is even required for properly conducted empirical scientific research, which, as I've mentioned, is open to question (Hanna, 2023a, 2023,b)—can be viewed as a product of

"statistical ritual and associated delusions" (Gigerenzer, 2018). We will now explore these "delusions."

2. Troubles in Statistical Paradise

There is considerable debate about the correct interpretation, and epistemological merits of significance testing, with many methodologists maintaining that the current approach is not scientific. For example, a paper published in *Nature* on 20 March 2019 with 800 signatories (Amrhein et al., 2019), summarized the most recent debates and criticisms, since there seems from the literature to have been something of a cycle of criticisms and doubts expressed about this statistical method, over the decades. Amrhein et al., proposed that the very idea of statistical significance should be retired:

> [W]e are calling for a stop to the use of *P* values in the conventional, dichotomous way … to decide whether a result refutes or supports a scientific hypothesis. (Amrhein et al., 2019).

The reason for this, which has often been given in the critical literature, is

> that all statistics, including *P* values and confidence intervals, naturally vary from study to study, and often do so to a surprising degree. In fact, random variation alone can easily lead to large disparities in *P* values, far beyond falling just to either side of the 0.05 threshold. (Amrhein et al., 2019)

On theoretical grounds, statistical significance can be spurious, arising from pure noise factors (McShane et al., 2019: p. 235). One reason that this is possible is that a result with a *p*-value below 0.05 is not necessarily evidence of a causal relationship (Holmon, et al., 2001), and indeed,

> researchers typically take the rejection of the sharp point null hypothesis of zero effect and zero systematic error as positive or even definitive evidence in favour of some preferred alternative hypothesis—a logical fallacy. (McShane et al., 2019, 237)

Thus, many statisticians and methodologists believe that the very idea of statistical significance should "expire" (Hurlbert et al., 2019). These problems will now be examined in more depth.

There are deep, troublesome problems with statistics, related to the philosophical foundations of statistical inference, which cast doubt on the objectivity of statistical evidence (Kaye, 1986; Fienberg et al., 1995). A battle of epochal scale has, and is continuing between the Bayesian and traditional Neyman-Pearson methods of hypothesis testing. The Neyman-Pearson model is a hybrid of Ronald Fisher's method and that of Jerzy Neyman and Egon Pearson. The hybrid method is usually known as the Null Hypothesis Significance Test (NHST). Stated simply, two hypotheses are formulated. The first is a statistical hypothesis called the null hypothesis (or restricted hypothesis) H_o and the second is called the alternative or research hypothesis H_r. The

null hypothesis states that there are no statistically significant differences between the populations from which the two samples are taken, so that observed differences arise by chance alone. The alternative or research hypothesis is a proposition in probabilistic form about aspects of the data, which is operationalized through a parameter θ. The null hypothesis might posit that $\theta = 0$ and the research hypothesis that $\theta \neq 0$. Under the assumption that the null hypothesis is true, a test statistic, such as chi-square or *t* statistic in linear regression analysis, these being a function of θ and the collected data, is then computed. A *p* value is then determined; as Nickerson summarizes:

> Application of NHST to the difference between the two means yields a value of *p*, the theoretical probability that if two samples of the size of those used had been drawn at random from the same population, the statistical test would have yielded a statistic (e.g., *t*) as large or larger than the one obtained. (Nickerson 2000: p. 242).

A significance level $\propto$ is specified and the null hypothesis is rejected only if the *p* value is not greater than $\propto$, often set at 0.05, and the experiment is statistically significant at the 0.05 level. Thus, either the null hypothesis is rejected or there is a failure to reject the null hypothesis.

Rejecting the null hypothesis is conventionally taken to be indirect evidence for the research hypothesis, since chance has supposedly been eliminated as an explanation for sample differences. Nevertheless, it is a fallacy to treat failure to disconfirm as confirmation, and to suppose that if H_o is rejected that the theory is established as true: it still may be false, but not by chance (Oakes, 1986: p. 83). NHST does not tell us the answer to the question, "Given these data, what is the probability that H_o is true?" (Cohen, 1994: p. 997). Rather, it tells us that "Given that H_o is true, what is the probability of these (or more extreme data)?" (Cohen, 1990).

NHST has been subjected to searching criticism (Selvin, 1957; Nunnally, 1960; Rozeboom, 1960; Lykken, 1968; Baken, 1966; Morrison & Henkel eds, 1970; Carver, 1978; Glass et al., 1981; Guttman, 1985; Rosenthal & Rubin 1985; McCloskey, 1986; Pratt, 1987; Chow, 1988, 1996, 1998; Loftus, 1991; Schmidt, 1991; Goodman, 1993; Frick, 1996; Kirk, 1996; Albelson, 1997; Berger et al., 1997; Harlow et al. eds, 1997; Hagan, 1997; Harris, 1997; Hunter, 1997; Shrout, 1997; Johnson, 1997; Gelman & Stern, 2006; Albert, 2002; Gliner et al., 2002; Hubbard & Bayarri, 2003; Morgan, 2003; Fidler et al. 2004; Banasiewicz, 2005). Many critics argue that the method lacks a sound scientific basis (Bakan, 1966; Carver, 1978; Gigerenzer, 1998; Sterne & Hunter, 2001; Schmidt & Hunter, 2002; Anderson et al., 2000; Armstrong, 2007; Wasserstein & Lazar, 2016; Hurlbert, et al., 2019). More generally, the criticisms are many and fundamental (Hurlbert & Lombardi, 2009).

For example, if the sample size is large enough statistical significance can occur for trivial effects (Ziliak & McCloskey, 2008). *P*-values depend upon sample size and with a large enough sample, the null hypothesis may be rejected (Berkson, 1938; Rozeboom, 1960; Grant, 1962; Bakan, 1966; Johnson, 1999, 2005; Shrader-Frechette,

2008). As I noted above, statistical significance can be generated by "pure noise" (Carney et al., 2010; Bem, 2011). Most null hypotheses are known to be false before any data are collected: indeed, nearly all null hypotheses are false *a priori*. (Ziliak & McCloskey, 2008). Further, it is not necessarily the case that a small *p* value shows strong evidence against the null; according to statisticians Berger and Sellke:

> [A]ctual evidence against a null (as measured, say, by posterior probability or comparative likelihood) can differ by *an order of magnitude* from the *P* value. For instance, data that yield a *P* value of .05, when testing a normal mean, result in a posterior probability of the null of *at least* .30 for *any* objective prior distribution. (Berger & Sellke, 1987: p. 112)

Correspondingly, they conclude: "P values can be highly misleading measures of the evidence provided by the data against the null hypothesis" (Berger & Sellke, 1987: p. 112; see also Berger & Berry, 1988; Simberloff, 1990). The difference between "significant" and "not significant" has been shown to be not itself statistically significant, as no sharp demarcation is possible, conceptually (Rosnow & Rosenthal, 1989; Gelman & Stern, 2006).

Bayesians Colin Howson and Peter Urbach, in *Scientific Reasoning: The Bayesian Approach*, are highly critical of significance testing (Howson & Urbach, 2006). For example, they point out that the chi-square test is "used to test theories asserting that some population has a particular, continuous probability distribution, such as the normal distribution" and to test such a theory,

> the range of possible results of some sampling trial would be divided into several intervals and the number of subjects falling into each would be compared with the 'expected' number. (Howson & Urbach, 2006: p. 139)

However, "the test … is … vitiated by the absence of any principled rule for partitioning the outcomes into separate intervals or cells, for not all partitions "lead to the same inferences when the significance test is applied" (Howson & Urbach, 2006: p. 139). They conclude that there is no epistemic basis for the chi-square test. Furthermore, Lindley's paradox, which shows that a well-supported hypothesis can be rejected in significance tests (Lindley, 1957; Loftus, 1996), indicates that

> the classical thesis that a null hypothesis may be rejected with greater confidence, the greater the power of the test is not borne out; indeed, the reverse trend is signalled. (Howson & Urbach, 2006: p. 154)

In their opinion, Lindley's paradox "shows unanswerably and decisively that inferences drawn from significance tests have no inductive significance whatsoever" (Howson & Urbach, 2006: p. 154). Likewise, they are skeptical about the epistemic cogency of classical estimates:

> classical 'estimates' are not estimates in any normal or scientific sense, and, like judgments of 'significance' and 'non-significance', they carry no inductive meaning at all. Therefore, they cannot be used to arbitrate between rival theories or to determine practical policy. (Howson & Urbach, 2006: p. 182)

In conclusion, they reject frequentism in favor of Bayesianism, as "classical methods are set altogether on the wrong lines, and are based on ideas inimical to scientific method" (Howson & Urbach, 2006: p. 182).

McShane et al. also believe that the problems with null hypothesis significance testing (NHST) remain unresolved, even by measures such as modified *p*-value thresholds, Bayes factors, and confidence intervals, holding that

> it seldom makes sense to calibrate evidence as a function of *p*-values or other purely statistical measures. (McShane et al., 2019: p. 236)

These criticisms are independent of Bayesian considerations and will hold even if Bayesianism is rejected on independent grounds, as I will argue below.

There is a further challenging critique of this field of statistics, alleging that scientific inference makes only a limited use of formal statistical inference, applying the statistical toolkit to random samples of data (Guttman, 1985; Gigerenzer, 2004; Hubbert et al., 2019). Many of the harder sciences than psychology and the social sciences, such as physics, astrophysics, cosmology, and chemistry, engage in mathematical model construction of physical phenomena usually by ordinary and partial differential equations, aiming to produce testable hypotheses, are subject to carefully designed experiments and/or observational studies, which are most often not random at all, with the aim being to obtain empirically replicable and generalizable data (Harman, 1965). As Hubbard et al. state:

> Scientific inference is better viewed as being grounded in abductive (explanatory) reasoning. Abduction—sometimes termed inference to the best explanation … takes as its locus the studying of facts and proposing a theory about the causal mechanisms generating them. Thus, abduction is a method of scientific inference geared toward the development of feasible and best explanations for the stubborn facts we possess. Like detective work, this approach mirrors the behavior of practicing scientists. And it is not beholden to methods of formal statistical inference. (Hubbard et al., 2019: p. 96)

Others agree: "Much of causal inference is beyond statistics" (Shadish & Cook, 1999: p. 298). "Statistical inference … is fundamentally incompatible with 'most' science" (Gunter & Tong, 2016-2017: p. 1). More generally, statistical methods are limited in most physical sciences: "the estimation of fixed population parameters from random samples is limited" (Guttman, 1985; Gigerenzer, 2004).

3. A Critique of Bayesianism

Some methodologists advocate the view that significance tests should be replaced by alternative methods, "the new statistics" of estimation confidence intervals, and meta-analysis (Cumming, 2012, 2014; Gelman, 2014; Morey et al., 2014, 2016). There are many sound points presented by Cumming, including 25 guidelines for improving psychological research, for example, not trusting any *p*-value, and to accept that any results are "one possibility from an infinite sequence" (Cumming, 2014: p. 8). However, as far as presenting an alternative statistical framework to NHST goes, there are many published criticisms. In general, critics of this approach argue from a Bayesian perspective, that the frequentist approach using confidence intervals, leads to inconsistent inferences, and that confidence intervals do not solve the existing problems with null hypothesis significance testing (Dienes, 2011). However, as we will now see, Bayesianism itself does not fare any better, and has its own conceptual difficulties.

This major competing school of thought, Bayesianism, holds that the inductive support for hypotheses is assessed on the basis of subjective and objective factors. The subjective factor is the prior probability of a hypothesis before the evidence is assessed. It is subjective, because epistemic subjects will frequently differ in their prior probabilities Pr(*h*), for a hypothesis *h*. The objective factor consists of direct inference probabilities that a hypothesis *h* is supported by evidence *e*. More explicitly,

> Bayes's theorem relates these direct inference probabilities with a subject's prior probabilities to produce the subject's *posterior probability*, the subject's probability judgment after the evidence has been considered. Bayes' theorem relates the *posterior* (or later coming) probability of a hypothesis Pr(*h*/*e*) to Pr(*h*), Pr(*e*/*h*) and Pr(*e*) so that knowing the values of the last three terms will enable the calculation of Pr*(h/e)* as:
>
> Bayes' Theorem: $$Pr(h/e) = \frac{Pr(e/h).Pr(h)}{Pr(e)}$$
>
> for Pr(*h*), Pr(*e*) > 0 (Smith et al., 1999 : p. 33).

Therefore:

> scientific inference as involves moving from the prior probability Pr(*h*) of a hypothesis to its posterior probability Pr(*h*/*e*) on the basis of the evidence collected, such that if Pr(*h*/*e*) > Pr(*h*) then *e* confirms or supports *h*. If Pr(*h*/*e*) < Pr(*h*) then *e* disconfirms or refutes *h*. (Smith et al., 1999: P. 33)

Just as the conventional significance testing approach has been subject to extensive criticism, so too has the Bayesian approach. The critics of the Bayesian approach believe that it has severe limitations and cannot provide a complete statistical methodology for the sciences, with critics raising problems about the limits of rationality and the cognitive capacities of Bayesian subjects, questioning the claim

that degrees of justification are Bayesian probabilities, and demonstrating the mathematical and computational intractability of Bayesian methods for even simple problems (Kyburg, 1978, 1993; Hyliand & Zeckhauser, 1979; Garber, 1983; Sowden, 1984; Humberg, 1987; Van Fraassen, 1988; Earman, 1989; Eells, 1990; Howson, 1991; Zynda, 1995; Wagner, 1997; Barnes, 1999; Gunn et al., 2016). What is interesting about this debate, if one adopts a neutral standpoint, is that the experts seem to make telling criticisms of opposing statistical methodologies without begging the question and assuming that their own position is correct, such as (i) the argument from the computational intractability of the Bayesian approach, that holds even if significance tests face independent criticisms, which says that *p* values are misunderstood as posterior probabilities of the null hypothesis, and (ii) the common fallacious deduction of "no difference" from "no significant difference," and "non-significant" with "no effect" (Hill, 1965: pp. 299-300; Greenland, 2011). This raises the threat of epistemological skepticism, only this time for the sciences. It certainly raises a very severe challenge to expert knowledge, and there are many astonishing claims made in the technical literature.

Let us consider one of the core foundational challenges to Bayesianism, which is the argument that there are no good reasons for believing that epistemic subjects, have any degree of confidence assignments that in general obey the axioms of the Pascalian probability calculus (Kaplan, 1989). The critical allegation to be considered is that there is little reason for supposing that betting provides a method for demonstrating the existence of degrees of belief (Milne, 1991).

The Bayesian claims that degrees of belief exist because he/she can measure them. The standard Bayesian argument for this, to paraphrase the argument by Glymour, is as follows. No rational agent will accept a bet where a loss is expected, but a rational agent will accept a bet where a gain is expected. The degree of belief in proposition P is the highest amount U that a person will pay to receive U+V for a fixed V, if P is true, but if P is not true, nothing will be received. The expected gain on paying U is zero, if U is the greatest amount willing to be paid for the bet. If P is the case, then the agent's gain is V, but if P is not the case, the gain is –U. Therefore:

$$V.Pr(P) + (–U).Pr(\sim P) = 0.$$

Since

$$Pr(\sim P) = 1 – Pr(P),$$

then:

$$Pr(P) = U / (U+V)$$

Thus, the rational agent striving to maximize expected-gain will make a bet if the expected gain is greater than zero. The degree of belief will be determined by the betting odds accepted (Glymour, 1980: pp. 69-70).

However, the problem with this argument is that it is circular. For the rational agent to contemplate betting at all in this situation, so that the betting odds are accepted, requires positing a wealth of prior beliefs about the betting set-up itself: namely, that the bet will pay if he/she wins, that the set-up is fair and so on. Thus, the argument *presupposes* degrees of belief rather than proving their existence. As well, there are many beliefs about which we may have a feeling of plausibility, but where we are not prepared to gamble because maximizing expected gain is socially inappropriate. The juror's belief about an accused person's guilt or innocence is an example. Betting language seems inappropriate in the context.

Beyond this though, even if there are degrees of belief, as we have seen, most people cannot reasonably attach a specific number to a required level of confidence and this task is even more difficult, perhaps impossible for them to do, when a large set of evidence is presented. The numbers produced to be plugged into Bayes' theorem will be essentially arbitrary (Humphreys, 1988).

No Bayesian has shown how the Bayesian methodology could be *practically* applied in a real evidential situation—for example, criminal trials involving thousands of items of evidence to consider. The use of Bayesian methodology in law can serve as a test case. For example, the updating of probabilities by Bayesian conditionalization, where a mere 30 pieces of evidence is introduced, would need the consideration of billions of probabilities (Bergman & Moore, 1991). Justice David Hodgson, of the Supreme Court of New South Wales, had this to say about the practical application of Bayes' Theorem to a legal problem in evidence:

> As an exercise, I have written a judgment for the hypothetical case, which applies Bayes' theorem, and set it out in an Appendix. It required two assumptions of prior probabilities of hypotheses, and twelve Bayesian steps, each involving two assumptions of numerical probabilities of evidence, given the truth or falsity of hypotheses: twenty-six guesses in all. In all twenty-six, I found I had virtually no confidence in the numbers I initially selected (in some cases partly because of unsureness of exactly what question I was asking, as well as because I just had to guess the answer); and I felt I had to check the numbers against the plausibility of the results, and then adjust (and re-adjust) the numbers, in order to arrive at numbers in which I had very slightly more confidence. (That is, I had to cheat.) Such little confidence as I ended up with depended very heavily on my common-sense assessment of the plausibility of the intermediate results and the conclusion.
>
> I think my hypothetical case shows that, for ordinary contested cases, it is fanciful to envisage a process by which a court manipulates probabilities fixed upon for certain basic statements (premisses) to arrive at a decision of the case (conclusion). In all steps from the premisses to the conclusion, a judge will generally have in the forefront of her mind the actual particular circumstance of the case, and will be making common

> sense judgments of (non-quantitative) probability in making these steps (as well as in determining upon the premisses). Indeed, the ultimate decision on the facts will generally itself be a common-sense judgment of non-quantitative probability concerning the overall situation, of very much the same kind as gave rise to the premisses—and very often the judge will (rightly) be more confident of reaching a correct overall conclusion 'on the balance of probabilities' than of assigning even approximate numerical probabilities to the premisses. (Hodgson, 1995: p. 56)

Philosophers Kevin Kelly and Clark Glymour are skeptical that Bayesianism captures the logic of scientific justification and have said:

> the sweeping consistency conditions implied by Bayesian ideals are computationally and mathematically intractable even for simple logical and statistical examples. (Kelly & Glymour, 2004: p. 95)

Indeed, Kelly and Glymour claim that Bayesian confirmation "is not even the right *sort* of thing to serve as an explication of scientific justification" (Kelly & Glymour, 2004: p. 95), because:

> Bayesian confirmation is just a change in the current output of a particular strategy or method for updating degrees of belief, whereas scientific justification depends on the truth-finding *performance* of the methods we use, whatever they might be. (Kelly & Glymour, 2004, pp. 95-96)

In particular,

> conditional probabilities can fluctuate between high and low values any number of times as evidence accumulates, so an arbitrary high degree of confirmation tells us nothing about how many fluctuations might be forthcoming in the future or about whether an alternative method might have required fewer. (Kelly & Glymour, 2004: p. 95-96)

For more critical argumentation along these lines, see also (Kelly & Schulter, 1995; Allen, 1996-1997; Ligertwood, 1996-1997; Norton, 2011).

Bayesianism also faces the difficulty of explaining where the initial priors come from in order to start the inferential process (Simpson & Orlov, 1979-1980). A logician sympathetic to Bayesianism, Patrick Suppes, has pointed out that "there is an almost total absence of a detailed discussion of the highly differentiating nature of past experience in forming a prior" (Suppes, 2007: p. 441). About this problem R.A. Fisher has said that Bayesians

> seem forced to regard mathematical probability, not as an objective quantity measured by observable frequencies, but as measuring merely psychological tendencies, theorems … which are useless for scientific purposes. (Fisher, 1960: pp. 6-7)

Similarly, Redmayne concluded this about subjective Bayesianism:

> When the only constraint on rational belief is coherence among a belief set, it can seem that anything goes. (Redmayne, 2003: p. 276)

For example, in a criminal law context, if the prior probability of guilt is taken to be zero, then as Eggleston puts it, "no amount of evidence could justify a conviction, since to assume an initial probability of zero is to postulate that guilt is impossible" (Eggleston, 1991: p. 276). However, on the other hand, "no lawyer would accept the proposition that the case should start with any particular presumption as to the probability of guilt" (Eggleston, 1991: p. 276). Rawling has argued that a Bayesian juror starting from an initial presumption of innocence will virtually never reach a judgment of "guilty beyond a reasonable doubt" (Eggleston, 1991, 276; Rawling, 1999).

Finally, to continue my legal example, even if we did grant that jurors had degrees of belief that ideally obeyed the Pascalian probability calculus, there is another reason for regarding Bayesianism as unsatisfactory. As Shafer has observed, there is *a constructive* character to personalistic probability judgments: these probability opinions are not ready made in a subject's mind. Rather, such probabilities arise from matching the problem at hand to background canonical examples where there are known probabilities (Shafer, 1986). As Shafer puts it, this process of construction involves us

> constructing an argument, an argument that draws an analogy between our actual evidence and the knowledge of objective probabilities in a complex physical experiment or game of chance. (Shafer, 1986: p. 802)

In doing this, that is, in constructing an explanatory structure that accounts for the evidence at hand, the juror (or epistemic subject in general) does not attempt to form a conjunction C_1 & C_2 & … & C_n of statements and then obtain a probability for them via the multiplication rule. Rather, what is conducted is the attempt to assess whether the plaintiff or the defendant's explanatory structures more adequately account for the evidence as a whole. In this sense, personalistic probabilities at the final stage will be relevant only to entire *systems* of evidence, not to isolated evidential propositions as Bayesians suppose (Pardo, 2000).

I conclude that the Bayesian position is flawed as a general decision theory for many good reasons, but in particular, that rationality, as defined by Bayesians, is simply not a general feature of human interaction (Colman, 2003).

4. The Limits of Probability Theory

There are many unsolved logical problems facing probability theory, especially involving infinite events (Hild, 2000; Shackel, 2007; Hájek, 1997, 2003, 2007). For example, what is the probability of an infinite sequence of heads tossed with an unbiased coin (Williamson, 2007)? Assume that the coin is "fair" by hypothesis. Multiplying the conjunctive probabilities leads to a sequence converging to 0

probability. Yet, an infinite sequence of heads is one logical possibility. Williamson argues that the use of infinitesimal probabilities does not resolve the contradiction:

> Cantor showed that some natural, apparently compelling forms of reasoning fail for infinite sets. This moral applies to forms of probabilistic and decision-theoretic reasoning in a more radical way than may have been realised. Infinitesimals do not solve the problem. (Williamson, 2007: p. 179)

Another relevant problem is that of the definition of conditional probability as a ratio of unconditional probabilities (Hájek, 2003):

$$\Pr(A/B) = \frac{\Pr(A\&B)}{\Pr(B)}, \quad \Pr(B) > 0$$

Hájek notes that that zero probability events are not necessarily impossible and can be of real scientific interest. He points out that Kolmogorov deals with this problem by analyzing conditional probability as a random variable. But even here there are problems because conditional probabilities can be defined in situations where the ratio is undefined because Pr(A&B) and Pr(B) are undefined. For example, if there is an urn with 90 red balls and 10 white balls, well mixed, the probability of drawing a red ball given that a ball is drawn at random is 0.9. However, the ratio analysis gives:

$$\frac{\Pr(\text{X draws a red ball \& X draws a ball at random from the urn})}{\Pr(\text{X draws a ball at random from the urn})}$$

which does not have a defined numerator nor denominator (Hájek,2007).

Apart from these logical problems facing probability, one of the most important unsolved philosophical/methodological problems involving probabilities is the reference class problem: any sentence, event, or proposition can be classified in various ways; hence the probability of the sentence, event, or proposition, is dependent upon the classification (Colyvan et al., 2001; Kaye, 2004; Pardo, 2007; Colyvan & Regan, 2007; Rhee, 2007; Allen & Pardo, 2007a). The reference problem is not merely a problem for probabilistic evidence but as Roberts explains, is more general:

> Every factual generalisation implies a reference class, and this in turn entails that the reference class problem is an inescapable concomitant of inferential reasoning and fact-finding in legal proceedings. (Roberts, 2007: p. 245)

Nevertheless, the problem has frequently been discussed in the narrower context of probability problems by leading theorists such as John Venn (Venn, 1876) and Hans Reichenbach (Reichenbach, 1949: p. 374). Although the problem has been regarded by many inductive logicians as providing a decisive refutation of the frequentist interpretation of probability, the reference problem also arises for classical, logical, propensity and subjectivist Bayesian interpretation as well (Hájek, 2007). The

reference class problem has also been discussed in a legal context, and if the problem turns out to be insuperable for one area of human cognitive activity, then this establishes a general problem.

The reference class problem has been discussed in the. jurisprudential literature, in the case of *United States v Shonubi* (1992, 1995, 1997). A Nigerian citizen, Charles Shonubi, was convicted of smuggling heroin into New York by the Kennedy airport. Shonubi had made seven previous drug-smuggling trips. Since sentencing is based on the total quantity of drugs smuggled, the prosecution estimated the quantity of heroin smuggled on those prior trips. In the trial, the US Second Circuit Court of Appeals did not allow the statistical evidence. Consequently, Shonubi was prosecuted on the basis of the actual quantity of drugs in his possession at the time he was arrested. The statistical data were based upon estimates using the reference class of other Nigerians smuggling heroin into Kennedy airport using Shonubi's method of ingesting balloons containing heroin paste. But if use were made of a different reference class to which Shonubi also belonged, a conflicting probability would have been be obtained.

Ronald J. Allen and Michael S. Pardo, in their paper "The Problematic Value of Mathematical Models of Evidence" (Allen & Pardo, 2007a), have concluded that the reference class problem shows the epistemological limits of mathematical models of evidence for, at least, law:

> The reference-class problem demonstrates that objective probabilities based on a particular class of which an item of evidence is a member cannot typically (and maybe never) capture the probative value of that evidence for establishing facts relating to a specific event. The only class that would accurately capture the 'objective' value would be the event itself, which would have a probability of one or zero, respectively. (Allen & Pardo, 2007a: p. 114).

There may be "practical" solutions to the reference class problem, because people make statistical inferences regularly in daily life (Cheng, 2009, 2089). Nevertheless, the theoretical issue, like that of making inductive inferences, is to show that such inferences are *justified*. Thus, Mike Redmayne concludes that the reference class problem is not intractable, but merely shows that probability judgments are relative to our evidence pool (Redmayne, 2008: p. 288). Agreed: but the issue in the debate is whether or not a rationally justified choice can be made between *prima facie* plausible, but conflicting probabilities, generated from different reference classes. Saying that our probability judgments are relative to our evidence pool, is true, but in fact only restates the problem: what is the "correct" evidence pool?

5. Conclusion

In this essay, I have examined the question raised by John Ioannidis, of why most published research findings, primarily in the social and biomedical sciences, are false. There are many reasons for this, such as small sample sizes, and even fraud, which

when exposed lead to substantial numbers of papers being retracted. There is also the quality control issue as well, whereby journals are reluctant to publish refutations of papers, so that there is a build-up of intellectual "rubbish," just as a creek might get clogged up with weeds. However, as I discussed above, the crisis of statistical methodology is also genuinely important, for if the foundational methodologies are flawed, then we cannot have reasoned faith in the conclusions reached. And that is precisely the situation in disciplines like psychology, for example, as far as much or even most empirical scientific research in those disciplines goes. Therefore, a constructive or healthy skepticism about empirical science is strongly recommended. At the same time, however, since constructive or healthy skepticism is itself a product of human rationality, then a cautious optimism about human rationality is also strongly recommended.

REFERENCES

(Abelson, 1997). Abelson, R.P. "On the Surprising Longevity of Flogged Horses: Why There is a Case for the Significance Test." *Psychological Science* 8: 12-15.

(Albert, 2002). Albert, M. "Resolving Neyman's Paradox." *British Journal for the Philosophy of Science* 53: 69-76.

(Allen & Pardo, 2007a). Allen, R.J & Pardo, M.S. "The Problematic Value of Mathematical Models of Evidence." *Journal of Legal Studies* 36: 107-140.

(Allen & Pardo, 2007b). Allen, R.J. & Pardo, M.S. "Probability, Explanation and Inference: A Reply." *International Journal of Evidence and Proof* 11: 307-317.

(Allen, 1996-1997). Allen, R.J. "Rationality, Algorithms and Juridical Proof: A Preliminary Inquiry." *International Journal of Evidence and Proof* 1: 254-275.

(Amrhein et al., 2019). Amrhein, V. et al. "Scientists Rise Up Against Statistical Significance," *Nature*. 20 March. Available online at URL = <https://www.nature.com/articles/d41586-019-00857-9.>

(Anderson et al., 2000). Anderson, D.R. et.al. "Null Hypothesis Testing: Problems, Prevalence and an Alternative." *Journal of Wildlife Management* 64: 912-923.

(Armstrong, 2007). Armstrong, J.S. "Significance Tests Harm Progress in Forecasting." *International Journal of Forecasting* 23: 321-327.

(Bakan, 1966). Bakan, D. "The Effect of Significance in Psychological Research." *Psychological Bulletin* 66: 423-437.

(Banasiewicz, 2005). Banasiewicz, A.D. "Marketing Pitfalls of Statistical Significance Testing." *Marketing Intelligence and Planning* 23: 515-528.

(Barnes, 1999). Barnes, E.C. "The Quantitative Problem of Old Evidence." *British Journal for the Philosophy of Science* 50: 249-264.

(Begley & Ellis, 2012). Begley, C.G. & Ellis, L.M. "Drug Development: Raise Standards for Preclinical Cancer Research." *Nature* 483: 531-533.

(Bem, 2011). Bem, D. J. "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect." *Journal of Personality and Social Psychology* 100: 407-425.

(Berger & Sellke, 1987). Berger, J.O. & Sellke, T. "Testing a Point Null Hypothesis: The Irreconcilability of *P* Values and Evidence." *Journal of the American Statistical Association* 82: 112-122.

(Berger & Berry, 1988). Berger, J.O. & Berry, D.A. "Statistical Analysis and the Illusion of Objectivity." *American Scientist* 76: 159-165.

(Berger et al., 1997). Berger, J. et.al. "Unified Frequentist and Bayesian Testing of a Precise Hypothesis." *Statistical Science* 12: 133-160.

(Bergman & Moore, 1991). Bergman, P. and Moore, A. "Mistrial by Likelihood Ratio: Bayesian Analysis Meets the F-Word." *Cardozo Law Review* 13: 589-619.

(Berkson, 1938). Berkson, J. "Some Difficulties of Interpretation Encountered in the Application of the Chi-Squared Test." *Journal of the American Statistical Association* 33: 526-536.

(Button, et al., 2013). Button, K.S. et al. "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience." *Nature Reviews/Neuroscience* 14: 365-376.

(Carney et al., 2010). Carney, D.R. et al., "Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance." *Psychological Science* 21: 1363-1368.

(Carver, 1978). Carver, R. "The Case Against Statistical Significance Testing." *Harvard Educational Review* 48: 378-399.

(Cheng, 2009). Cheng, E.K. "A Practical Solution to the Reference Class Problem." *Columbia Law Review* 109: 2081-2105.

(Chow, 1988). Chow, S.L. "Significance Test or Effect Size?" *Psychological Bulletin* 103: 105-110.

(Chow, 1996). Chow, S.L. *Statistical Significance: Rationale, Validity and Utility*. London: Sage.

(Chow, 1998). Chow, S.L. "Precis of *Statistical Significance: Rationale, Validity and Utility*." *Behavioral and Brain Sciences* 21: 169-239.

(Cohen, 1990). Cohen, J. "Things I Have Learned (So Far)." *American Psychologist* 45: 1304-1312.

(Cohen, 1994). Cohen, J. "The Earth is Round ($p < 0.05$)." *American Psychologist* 49: 997-1003.

(Colman, 2003). Colman, A.M. "Cooperation, Psychological Game Theory, and Limitations of Rationality in Social interaction." *Behavioral and Brain Sciences* 26: 139-198.

(Colyvan et al., 2001). Colyvan, M. et al. "Is It a Crime to Belong to a Reference Class?" *Journal of Political Philosophy* 9: 168-181.

(Colyvan & Regan, 2007). Colyvan, M. & Regan, H.M. "Legal Decisions and the Reference Class Problem." *International Journal of Evidence and Proof* 11: 274-285.

(Cumming, 2012). Cumming, G. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. London: Routledge.

(Cumming, 2014). Cumming, G. "The New Statistics: Why and How." *Psychological Science* 25: 7-29.

(de Long & Lang, 1992). de Long, J.B. and Lang, K. "Are All Economic Hypotheses False?" *Journal of Political Economy* 100: 1257-1272.

(Diekmann, 2011). Diekmann, A. "Are Most Published Research Findings False?" *Journal of Economics and Statistics* 231: 628-635.

(Dienes, 2011). Dienes, Z. "Bayesian Versus Orthodox Statistics; Which Side Are You On?" *Perspectives on Psychological Science* 6: 274-290.

(Earman, 1989). Earman, J. "Old Evidence, New Theories: Two Unresolved Problems in Bayesian Confirmation Theory." *Pacific Philosophical Quarterly* 70: 323-340.

(Eells, 1990). Eells, E. *Bayesian Problems of Old Evidence in Scientific Theories* Minneapolis MN: Univ. of Minnesota Press.

(Eggleston, 1991). Eggleston, R. "Similar Facts and Bayes' Theorem." *Jurimetrics Journal* 31: 275-287.

(Everett & Earp, 2015). Everett, J.A.C. & Earp, B.D. "A Tragedy of the (Academic) Commons: Interpreting the Replication Crisis in Psychology as a Social Dilemma for Early-Career Researchers." *Frontiers in Psychology* 6. 5 August. Available online at URL = <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2015.01152/full>.

(Fidler et al., 2004). Fidler, F. et al. "Editors Can Lead Researchers to Confidence Intervals, But They Can't Make Them Think: Statistical Reform Lessons from Medicine." *Psychological Science* 15: 119-126.

(Fienberg et al., 1995). Fienberg, S.E. et.al. "Understanding and Evaluating Statistical Evidence in Litigation." *Jurimetrics Journal* 36: 1-32.

(Fisher, 1960). Fisher, R.A. *The Design of Experiments* 7th edn., New York: Hafner.

(Freedman et al., 2015). Freedman, L.P. et al. "The Economics of Reproducibility in Preclinical Research." *PLOS Biology* 13. 9 June. Available online at URL = <https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002165>.

(Freeman, 2010). Freedman, D.H. *Wrong.* New York: Little Brown and Company.

(Frick, 1996). Frick, R.W. "The Appropriate Use of Null Hypothesis Testing." *Psychological Methods* 1: 379-390.

(Garber, 1983). Garber, D. "Old Evidence and Logical Omniscience in Bayesian Confirmation Theory." In J. Earman (ed.), *Testing Scientific Theories.* Minneapolis MN: Univ. of Minnesota Press. Pp. 99-132.

(Gelman & Stern, 2006). Gelman, A. & Stern, H. "The Difference Between 'Significant' and 'Not Significant' is Not Itself Statistically Significant." *American Statistician* 60: 328-331.

(Gelman, 2014). Gelman, A. "The Fallacy of Placing Confidence in Confidence Intervals." Available online at URL = <http://andrewgelman.com/2014/12/11/fallacy-placing-confidence-confidence-intervals/.>

(Gigerenzer, 1998). Gigerenzer, G. "We Need Statistical Thinking, Not Statistical Rituals." *Behavioral and Brain Sciences* 21: 199-200.

(Gigerenzer, 2004). Gigerenzer, G. "Mindless Statistics." *Journal of Socio-Economics* 33: 587-606.

(Gigerenzer, 2018). Gigerenzer, G. "Statistical Rituals: The Replication Delusion and How We Got There." *Advances in Methods and Practices in Psychological Science* 1, 2: 198-218.

(Glass et al., 1981). Glass, G.V. et al. *Meta-Analysis in Social Research*. Beverly Hills CA: Sage Publications.

(Gliner et al., 2002). Gliner, J. et al. "Problems with Null Hypothesis Significance Testing (NHST): What Do the Textbooks Say?" *Journal of Experimental Education* 71: 83-92.

(Glymour, 1980). Glymour, C. *Theory and Evidence.* Princeton NJ: Princeton Univ. Press.

(Godlee et al., 1998). Godlee, F. et al. "Effect on the Quality of Peer Review of Blinding Reviewers and Asking Them to Sign Their Reports: A Randomized Control Trial." *JAMA* 280: 237-240.

(Goodman, 1993). Goodman, S.N. "*P* Values Hypothesis Tests, and Likelihood: Implications for Epidemiology of a Neglected Historical Debate." *American Journal of Epidemiology* 137: 485-496.

(Grant, 1962). Grant, D.A. "Testing the Null Hypothesis and the Strategy and Tactics of Investigating Theoretical Models." *Psychological Review* 69: 54-61.

(Greenland, 2011). Greenland, S. "Null Misinterpretation in Statistical Testing and Its Impact on Health Risk Assessment." *Preventive Medicine* 53: 225-228.

(Gunn et al., 2016). Gunn, L.J. et al. "Too Good to be True: When Overwhelming Evidence Fails to Convince." *Proceedings of the Royal Society A* 472. 23 March. Available online at URL = <https://royalsocietypublishing.org/doi/10.1098/rspa.2015.0748>.

(Gunter& Tong, 2016-2017). Gunter, B. & Tong, C. "A Response to the ASA Statement on Statistical Significance and *P*-Values." *NV-ASA Newsletter* 14: 1-3.

(Guttman, 1985). Guttman, L. "The Illogic of Statistical Inference for Cumulative Science." *Applied Stochastic Models and Data Analysis* 1: 3-10.

(Hagan, 1997). Hagan, R.L. "In Praise of the Null Hypothesis Statistical Test." *American Psychologist* 52: 15-24.

(Hájek, 1997). Hájek, A. "Mises Redux-Redux: Fifteen Arguments against Finite Frequentism." *Erkenntnis* 45: 209-227.

(Hájek, 2003). Hájek, A. "What Conditional Probability Could Not Be." *Synthese* 137: 273-323.

(Hájek, 2007). Hájek, A. "The Reference Class Problem is Your Problem Too." *Synthese* 156: 563-585.

(Hanna, 2023a). Hanna, R. "Empirical Science with Uncertainty but Without Reproducibility." *Against Professional Philosophy*. December 10. Available online at URL = <https://againstprofphil.org/2023/12/10/empirical-science-with-uncertainty-but-without-reproducibility/>.

(Hanna, 2023b). Hanna, R. "The End of Peer Review and the Matrix of Ideas." *Against Professional Philosophy*. Available online at URL = <https://againstprofphil.org/2023/12/17/the-end-of-peer-review-and-the-matrix-of-ideas/>.

(Harlow, et al., eds, 1997). Harlow, L.L. et al. eds. *What If There Were No Significance Tests?* Mahwah NJ: Lawrence Erlbaum.

(Harman, 1965). Harman, G. "Inference to the Best Explanation." *Philosophical Review* 74: 88-95.

(Harris, 1997). Harris, R. J. "Significance Tests Have Their Place." *Psychological Science* 8: 8-11.

(Hartshorne et al., 2012). Hartshorne, J. et al. "Tracking Replicability as a Method of Post-Publication Open Evaluation." *Frontiers in Computational Neuroscience* 6: 1-13.

(Higginson & Munafò, 2016). Higginson, A.D. and Munafò, M.R. "Current Incentives for Scientists Lead to Underpowered Studies with Erroneous Conclusions." *PLOS Biology* 14. 10 November. Available online at URL = <https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2000995>.

(Hild, 2000). Hild, M. "Trends in the Philosophy of Probability." *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 31: 419-422.

(Hill, 1965). Hill, A.B. "The Environment and Disease: Association or Causation?" *Proceedings of the Royal Society of Medicine* 58: 295-300.

(Hodgson, 1995). Hodgson, D. "Probability: The Logic of the Law—A Response." *Oxford Journal of Legal Studies* 15: 51-68.

(Holman et al., 2001). Holman, C.L. et al. "A Psychometric Experiment in Causal Inference to Estimate Evidential Weights used by Epidemiologists." *Epidemiology* 12: 246-255.

(Homburg, 1987). Humburg, J. "The Bayes Rule is Not Sufficient to Justify or Describe Inductive Reasoning." *Erkenntnis* 26: 379-390.

(Horton, 2015). Horton, R. "Offline: What is Medicine's 5 Sigma?" *The Lancet* 385: 1380.

(Howson, 1991). Howson, C. "The Old Evidence Problem." *British Journal for the Philosophy of Science* 42: 547-555.

(Howson & Urbach, 2006). Howson, C. & Urbach, P. *Scientific Reasoning: The Bayesian Approach.* La Salle IL: Open Court.

(Hubbard & Bayarri, 2003). Hubbard, R. & Bayarri, M. J. "Confusion Over Measures of Evidence (P's) Versus Errors (α's) in Classical Statistical Testing (With Comments)." *American Statistician* 57: 171-182.

(Hubbard et al., 2019). Hubbard, R. et al. "The Limited Role of Formal Statistical Inference in Scientific Inference." *American Statistician* 73 (S1): 91-98.

(Humphreys, 1988). Humphreys, P. "Non-Nietzschean Decision Making." In J.H. Fetzer (ed.), *Probability and Causality*. Dordrecht: Kluwer. Pp. 253-268.

(Hunter, 1997). Hunter, J.E. "Needed: A Ban on the Significance Test." *Psychological Science* (Special Section) 8: 3-7.

(Hurlbert & Lombardi, 2009). Hurlbert, S. H. & Lombardi, C. M. "Final Collapse of the Neyman-Pearson Decision Theoretic Framework and Rise of the NeoFisherian." *Annales Zoologica Fennici* 46: 311-349.

(Hurlbert, et al., 2019). Hurlbert, S. H. et al. "Coup de Grâce for a Tough Old Bull: 'Statistically Significant' Expires." *American Statistician* 73 (S1): 352-357.

(Hyliand & Zeckhauser, 1979). Hyliand, A. and Zeckhauser, R. "The Impossibility of Bayesian Group Decision Making with Separate Aggregation of Beliefs and Values." *Econometrica* 47: 1321-1336.

(Ioannidis & Trikalinos, 2005). Ioannidis, J.P.A. & Trikalinos, T. A. "Early Extreme Contradictory Estimates May Appear in Published Research: The Proteus Phenomenon in Molecular Genetics Research and Randomized Trials." *Journal of Clinical Epidemiology* 58: 543-549.

(Ioannidis, 2005a). Ioannidis, J.P.A. "Why Most Published Research Findings are False." *PLOS* Medicine 2, 8: 0696-0701.

(Ioannidis, 2005b). Ioannidis, J.P.A. "Contradictions in Highly Cited Research—Reply." *JAMA* 294: 2695-2696.

(Ioannidis, 2008). Ioannidis, J.P.A. "Why Most Discovered True Associations are Inflated." *Epidemiology* 19: 640-648.

(Ioannidis, 2014). Ioannidis, J.P.A. et al. "Publication and Other Reporting Biases in Cognitive Sciences: Detection, Prevalence, and Prevention." *Trends in Cognitive Sciences* 18, 5: 235-241.

(Jefferson et al., 2002). Jefferson, T. et al. "Effects of Editorial Peer Review: A Systematic Review." *JAMA* 287: 2784-2786.

(Johnson, 1999). Johnson, D.H. "The Insignificance of Statistical Significance Testing." *Journal of Wildlife Management* 63: 763-772.

(Johnson, 2005). Johnson, D.H. "What Hypothesis Tests are Not: A Response to Colgrave and Ruxton." *Behavioral Ecology* 16: 323-324.

(Kaplan, 1989). Kaplan, M. "Bayesianism without the Black Box." *Philosophy of Science* 56: 48-69.

(Kaye, 1986). Kaye, D.H. "Is Proof of Statistical Significance Relevant?" *Washington Law Review* 61: 1333-1365.

(Kaye, 2004). Kaye, D. H. "Logical Relevance: Problems with the Reference Population and DNA Mixtures in *People v Pizarro*." *Law, Probability and Risk* 3: 211-220.

(Kelly & Glymour, 2004). Kelly, K.T. & Glymour, C. "Why Probability Does Not Capture the Logic of Scientific Justification." In C. Hitchcock (ed.), *Contemporary Debates in Philosophy of Science*. Malden MA: Blackwell. Pp. 94-114.

(Kelly & Schulte, 1995). Kelly, K. and Schulte, O. "The Computable Testability of Theories with Uncomputable Predictions." *Erkenntnis* 42: 29-66.

(Kirk, 1996). Kirk, R.F. "Practical Significance: A Concept Whose Time Has Come." *Educational and Psychological Measurement* 56: 746-759.

(Kyburg, 1978). Kyburg, H. "Subjective Probability: Criticisms, Reflections and Problems." *Journal of Philosophical Logic* 7: 157-180.

(Kyburg, 1993). Kyburg, H. "The Scope of Bayesian Reasoning," in D. Hull et al. (eds.), *Philosophy of Science Association 1992*. East Lansing MI: Philosophy of Science Association. Pp. 139-152.

(Ligertwood, 1996-1997). Ligertwood, A. "Bayesians and the World Out There." *International Journal of Evidence and Proof* 1: 321-325.

(Lindley, 1957). Lindley, D.V. "A Statistical Paradox." *Biometrika* 44: 187-192.

(Loftus, 1991). Loftus, G.R. "On the Tyranny of Hypothesis Testing in the Social Sciences." *Contemporary Psychology* 36: 102-105.

(Lykken, 1968). Lykken, D.T. "Statistical Significance in Psychological Research." *Psychological Bulletin* 70: 151-159.

(Martin, 1992). Martin, B. "Scientific Fraud and the Power Structure of Science." *Prometheus* 10: 83-98.

(McCloskey, 1986). McCloskey, D. "Why Economic Historians Should Stop Relying on Statistical Tests of Significance and Lead Economists and Historians into the Promised Land." *Newsletter of the Cliometrics Society* 2: 5-7.

(McShane, 2019). McShane, B.B. "Abandon Statistical Significance." *American Statistician* 73 (S1): 235-245.

(Merton, 1968). Merton, R.K. "The Matthew Effect in Science." *Science* 159: 56-63.

(Milne, 1991). Milne, P. "Annabel and the Bookmaker: An Everyday Tale of Bayesian Folk." *Australasian Journal of Philosophy* 69: 98-102.

(Moonesinghe et al., 2007). Moonesinghe, R. et al. "Most Published Research Findings are False—But a Little Replication Goes a Long Way." *PLoS Medicine* 4, 2: 0218-0221.

(Morey et al., 2014). Morey, R.D. et al. "Why Hypothesis Tests are Essential for Psychological Science: A Comment on Cumming." *Psychological Science* 25: 1289-1290.

(Morey et al., 2016). Morey, R.D. et al. "The Fallacy of Placing Confidence in Confidence Intervals." *Psychological Bulletin and Review* 23: 103-123.

(Morgan, 2003). Morgan, P.L. "Null Hypothesis Significance Testing: Philosophical and Practical Considerations of a Statistical Controversy." *Exceptionality* 11: 209-221.

(Morrison & Henkel eds., 1970). Morrison, D.E. and Henkel, R.E. (eds.), *The Significance Test Controversy.* Chicago: Aldine.

(Nickerson, 2000). Nickerson, R.S. "Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy." *Psychological Methods* 5: 241-305.

(Norton, 2011). Norton, J. D. "Challenges to Bayesian Confirmation Theory." In P.S. Bandyopadhyay and M. Forster (eds.), *Philosophy of Statistics: Vol. 7, Handbook of the Philosophy of Science*. Amsterdam: North Holland. Pp. 391-437.

(Nunnally, 1960). Nunnally, J. "The Place of Statistics in Psychology." *Educational and Psychological Measurement* 20: 641-650.

(Oakes, 1986). Oakes, M. *Statistical Inference: A Commentary for the Social and Behavioral Sciences*. New York: Wiley.

(Open Science Collaboration, 2015). Open Science Collaboration. "Estimating the Reproducibility of Psychological Science." *Science* 349: aac4716-1—aac4716-8.

(Pardo, 2000). Pardo, M.S. "Juridical Proof, Evidence and Pragmatic Meaning: Toward Evidentiary Holism." *Northwestern University Law Review* 95: 399-442.

(Pardo, 2007). Pardo, M.S. "Reference Classes and Legal Evidence." *International Journal of Evidence and Proof* 11: 255-258.

(Peters & Ceci, 1982). Peters, D. & Ceci, S. "Peer-Review Practices of Psychological Journals: The Fate of Submitted Articles, Submitted Again." *Behavioral and Brain Sciences* 5: 187-255.

(Pratt, 1987). Pratt, J.W. "Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence: Comment." *Journal of the American Statistical Association* 82: 123-125.

(Rawling, 1999). Rawling, P. "Reasonable Doubt and the Presumption of Innocence: The Case of the Bayesian Juror." *Topoi* 18: 117-126.

(Redmayne, 2003). Redmayne, M. "Objective Probability and the Assessment of Evidence." *Law, Probability and Risk* 2: 275-294.

(Redmayne, 2008). Redmayne, M. "Exploring the Proof Paradoxes." *Legal Theory* 14: 281-309.

(Reichenbach, 1949). Reichenbach, H. *The Theory of Probability*. Berkeley CA: Univ. of California Press.

(Rhee, 2007). Rhee, R.J. "Probability, Policy and the Problem of the Reference Class." *International Journal of Evidence and Proof* 11: 286-291.

(Roberts, 2007). Roberts, P. "From Theory into Practice: Introducing the Reference Class Problem." *International Journal of Evidence and Proof* 11: 243-254.

(Rosenthal & Rubin, 1985). Rosenthal, R. and Rubin, D.B. "Statistical Analysis: Summarizing Evidence Versus Establishing Facts." *Psychological Bulletin* 97: 527-529.

(Rosnow & Rosenthal, 1989). Rosnow, R.L. & Rosenthal, R. "Statistical Procedures and the Justification of Knowledge in Psychological Science." *American Psychologist* 44: 1267-1284.

(Rozeboom, 1960). Rozeboom, W.W. "The Fallacy of the Null-Hypothesis Significance Test." *Psychological Bulletin* 57: 416-428.

(Schmidt & Hunter, 2002), Schmidt, F.L. and J.E. Hunter, J.E. "Are There Benefits from NHST?" *American Psychologist* 57: 65-66.

(Schmidt, 1991). Schmidt, F.L. "Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers." *Psychological Methods* 1: 115-129.

(Selvin, 1957). Selvin, H.C. "A Critique of Tests of Significance in Survey Research." *American Sociological Review* 22: 519-527.

(Shackel, 2007). Shackel, N. "Bertrand's Paradox and the Principle of Indifference." *Philosophy of Science* 74: 150-175.

(Shadish & Cook, 1999). Shadish, W. R. & Cook, T. D. "Comment—Design Rules: More Steps Towards a Complete Theory of Quasi-Experimentation." *Statistical Science* 14: 294-300.

(Shafer, 1986). Shafer, G. "The Construction of Probability Arguments." *Boston University Law Review* 66: 799-816.

(Shrader-Frechette, 2008). Shrader-Frechette, K. "Statistical Significance in Biology: Neither Necessary Nor Sufficient for Hypothesis Acceptance." *Biological Theory* 3: 12-16.

(Shrout, 1997). Shrout, P. "Should Significance Tests Be Banned?" *Psychological Science* 8: 1-2.

(Simberloff, 1990). Simberloff, D. "Hypotheses, Errors and Statistical Assumptions." *Herpetologica* 46: 351-357.

(Simmons, 2011). Simmons, J.P. "False Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22: 1359-1366.

(Simpson & Orlov, 1979-1980). Simpson, S. and Orlov, M. "Comment: An Application of Logic to the Law." *University of New South Wales Law Journal* 3: 415-425.

(Smith, 2023). Smith, J. W. "Against the Academics: Peering at the Problem of Peer Review." *Against Professional Philosophy*. June 18. Available online at URL = <https://againstprofphil.org/2023/06/18/against-the-academics-peering-at-the-problem-of-peer-review/>.

(Smith, 1999). Smith, J.W. et al. *The Bankruptcy of Economics*. London: Macmillan.

(Smith & Smith, 2023a). Smith, J.W. & Smith, S. "Corruption, Falsity, and Fraud: The Epistemological Crisis of Professional Academic Research." *Against Professional Philosophy*. 17 September. Available online at URL = < <https://againstprofphil.org/2023/09/17/corruption-falsity-and-fraud-the-epistemological-crisis-of-professional-academic-research/>.

(Smith & Smith, 2023b). Smith, J. W. & Smith, S. "From Scientific Reproducibility to Epistemic Humility." *Against Professional Philosophy*. 3 December. Available online at URL = <https://againstprofphil.org/2023/12/03/from-scientific-reproducibility-to-epistemic-humility/>.

(Smith, 2006). Smith, R. "Peer Review: A Flawed Process at the Heart of Science and Journals." *Journal of the Royal Society of Medicine* 99: 178-182.

(Sowden, 1984). Sowden, L. "The Inadequacy of Bayesian Decision Theory," *Philosophical Studies* 45: 293-313.

(Stein, 1996-1997). Stein, A. "Judicial Fact-Finding and the Bayesian Method: The Case for Deeper Scepticism About Their Combination." *International Journal of Evidence and Proof* 1: 25-47.

(Sterne & Smith, 2001). Sterne, J.A.C. and Smith, G.D. "Sifting the Evidence—What's Wrong with Significance Tests? *British Medical Journal* 322: 226-231.

(Suppes, 2007). Suppes, P. "Where do Bayesian Priors Come From?" *Synthese* 156: 441-471.

(Tabarrok, 2005). Tabarrok, A. "Why Most Published Research Findings are False." *Marginal Revolution*. 2 September. Available online at URL = <http://marginalrevolution.com/marginalrevolution/2005/09/why_most_publis.html. >.

United States v Shonubi, 802 F Supp 859 (EDNY, 1992).

United States v Shonubi, 998 F2d 84 (2d Cir., 1993).

United States v Shonubi, 895 F Supp 480 (EDNY, 1995).

United States v Shonubi, 962 F. Supp 370 (EDNY, 1997).

United States v Shonubi, 103 F3d 1085 (2d Cir, 1997).

(Van Fraassen, 1988). Van Fraassen, B.C. "The Problem of Old Evidence." In D.F. Austin (ed.), *Philosophical Analysis*. Norwell: Kluwer. Pp. 153-165.

(Venn, 1876). Venn, J. *The Logic of Chance*. 2nd edn., London: Macmillan.

(Vogel, 2011). Vogel, G. "Scientific Misconduct: Psychologist Accused of Fraud on 'Astonishing Scale'." *Science* 334: 579.

(Wagner, 1997). Wagner, C.G. "Old Evidence and New Explanation." *Philosophy of Science* 64: 677- 691.

(Wasserstein & Lazar, 2016). Wasserstein, R. L. & Lazar, N.A. "The ASA Statement on *P*-Values: Context, Process, and Purpose." *American Statistician* 70: 129-133.

(Williamson, 2007). Williamson, T. "How Probable is an Infinite Sequence of Heads?" *Analysis* 67: 173-180.

(Yong, 2015). Yong, E. "How Reliable Are Psychology Studies?" *The Atlantic*. 27 August. Available online at URL = <http://www.theatlantic.com/science/archive/2015/08/psychology-studies-reliability-reproducibility-nosek/4024661>.

(Ziliak & McCloskey, 2008). Ziliak, S.T. & McCloskey, D.N. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives.* Ann Arbor MI: Univ. of Michigan Press.

(Zynda, 1995). Zynda, L. "Old Evidence and New Theories." *Philosophical Studies* 77: 67-95.