Studies in Computational Intelligence 809

Vladik Kreinovich
Nguyen Ngoc Thach
Nguyen Duc Trung
Dang Van Thanh *Editors*

Beyond Traditional Probabilistic Methods in Economics

Springer

Editors
Vladik Kreinovich
Department of Computer Science
University of Texas at El Paso
El Paso, TX, USA

Nguyen Ngoc Thach
Banking University HCMC
Ho Chi Minh City, Vietnam

Nguyen Duc Trung
Banking University HCMC
Ho Chi Minh City, Vietnam

Dang Van Thanh
TTC Group
Ho Chi Minh City, Vietnam

ISSN 1860-949X ISSN 1860-9503 (electronic)
Studies in Computational Intelligence
ISBN 978-3-030-04199-1 ISBN 978-3-030-04200-4 (eBook)
https://doi.org/10.1007/978-3-030-04200-4

Library of Congress Control Number: 2018960912

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Everything Wrong with P-Values Under One Roof

William M. Briggs(✉)

[redacted]

matt@wmbriggs.com

Abstract. P-values should not be used. They have no justification under frequentist theory; they are pure acts of will. Arguments justifying p-values are fallacious. P-values are not used to make all decisions about a model, where in some cases judgment overrules p-values. There is no justification for this in frequentist theory. Hypothesis testing cannot identify cause. Models based on p-values are almost never verified against reality. P-values are never unique. They cause models to appear more real than reality. They lead to magical or ritualized thinking. They do not allow the proper use of decision making. And when p-values seem to work, they do so because they serve a loose proxies for predictive probabilities, which are proposed as the replacement for p-values.

Keywords: Causation · P-values · Hypothesis testing
Model selection · Model validation · Predictive probability

1 The Beginning of the End

It is past time for p-values to be retired. They do not do what is claimed, there are better alternatives, and their use has led to a pandemic of over-certainty. All these claims will be proved here.

Criticisms of p-values are as old as the measures themselves. None was better than Jerzy Neyman's original, however, who called decisions made conditional on p-values "acts of will"; see [1,2]. This criticism is fundamental: once the force of it is understood, as I hope readers agree, it is seen there is no justification for p-values.

Many are calling for an end to p-value-drive hypothesis testing. An important recent paper is [3] which concludes that given the many flaws with p-values "it is sensible to dispense with significance testing altogether." The book *The Cult of Statistical Significance* [4] has had some influence. The shift away from formal testing, and parameter-based inference, is also called for in [5].

There are scores of critical articles. Here is an incomplete, small, but representative list: [6–18]. The mood that was once uncritical is changing, best demonstrated by the critique by [19], which leads with the modified harsh words of Sir Thomas Beecham, "One should try everything in life except incest, folk

V. Kreinovich et al. (Eds.): ECONVN 2019, SCI 809, pp. 22–44, 2019.
https://doi.org/10.1007/978-3-030-04200-4_2

dancing and calculating a P-value." A particularly good resource of p-value criticisms is the web page "A Litany of Problems With p-values" compiled and routinely updated by Harrell [20].

Replacements, tweaks, manipulations have all been proposed to save p-values, such as lowering the magic number. Prominent among these is Benjamin et al. [21], who would divide the magic number by 10. There are many others suggestions which seek to put p-values in their "proper" but still respected place. Yet none of the proposed fixes solve the underlying problems with p-values, which I hope to demonstrate below.

Why are p-values used? To say something about a theory's or hypothesis's truth or goodness. But the relationship between a theory's truth and p-values is non-existent by design. Frequentist theory forbids speaking of the probability of a theory's truth. The connection between a theory's truth and Bayes factors is more natural, e.g. [22], but because Bayes factors focus on unobservable parameters, and rely just as often on "point nulls" as do p-values, they too exaggerate evidence for or against a theory. It is also unclear in both frequentist and Bayesian theory what *precisely* a hypothesis or theory is. The definition is usually taken to mean non-zero value of a parameter, but that parameter, attached to a certain measurable in a model (the "X"), does not say how the observable (the "Y") itself changes in any causal sense. It only says how our *uncertainty* in the observable changes. Probability theories and hypotheses, then, are epistemic and not ontic statements; i.e., they speak of our knowledge of the observable, given certain conditions, and not on what causes the observable.

This means probability models are only needed when causes are unknown (at least in some degree; there are rare exceptions). Though there is some disagreement on the topic, e.g. [23–25], there is no ability for a wholly statistical model to identify cause. Everybody agrees models can, and do, find correlations. And because correlations are not causes, hypothesis testing cannot find causes, nor does it claim to in theory. At best, hypothesis testing highlights possibly interesting relationships. So that finding a correlation is all a p-values or Bayes factor, of indeed any measure, can do. But correlations exist whether or not they are identified as "significant" by these measures. And that identification, as I show below, is rife with contradictions and fallacies. Accepting that, it appears the only solution is to move from purely a hypothesis testing (frequentist or Bayes) scheme to a predictive one in which the model claimed to be good or true or useful can be verified and tested against reality. See the latter chapters of [26] for a complete discussion of this.

Now every statistician knows about at least these limitations of p-values (and Bayes factors), and all agree with them to varying extent (most disputes are about the nature of cause, e.g. contrast [25,26]). But the "civilians" who use our tools do not share our caution. P-values, as we all know, work like magic for most civilians. This explains the overarching desire for p-value hacking and the like. The result is massive over-certainty and a much-lamented reproducibility crisis; e.g. see among many others [27,28]; see too [13].

The majority—which includes all users of statistical models, not just careful academics—treat p-values like ritual, e.g. [8]. If the p-value is less than the magic number, a theory has been proved, or taken to be proved, or almost proved. It does not matter that frequentist statistical theory insists that this is not so. It is what everybody believes. And the belief is impossible to eradicate. For that reason alone, it's time to retire p-values.

Some definitions are in order. I take probability to be everywhere conditional, and nowhere causal, in the same manner as [26,29–31]. Accepting this is not strictly necessary for understanding the predictive position, which is compared with hypothesis testing below, but understanding the conditional nature of all probability required is for a complete philosophical explanation. Predictive philosophy's emphasis on observables and measurable values which only inform uncertainty in observables is the biggest point of departure between hypothesis testing, which assumes probability is real and, at times, even causal.

Predictive probabilities make an apt, easy, and verifiable replacement for p-values; see [26,32] for fuller explanations. Predictive probability is demonstrated in the schematic equation:

$$\Pr(\mathrm{Y}|\text{new X}, \mathrm{DMA}), \tag{1}$$

where Y is the proposition of interest. For example, Y = "$y > 0$", Y = "yellow", Y = "$y < -1$ or $y > 1$ but not $y = 0$ if x_3 = 'Detroit'"; basically, Y is any proposition that can be asked (and answered!). D is the old data, i.e. prior measures X and the observable Y (where the dimension of all is clear from the context), both of which may have been measured or merely assumed. The model characterizing uncertainty in Y is M, usually parameterized, and A is a list of assumptions probative to M and Y. Everything thought about Y goes into A, even if it is not quantifiable. For instance, in A is information on the priors of the parameters, or *whatever* other information that is relevant to Y. The new X are those values of the measures that must be assumed or measured each time the probability of Y is computed. They are necessary because they are in D, and modeled in M.

A book could be written summarizing all of the literature for and against p-values. Here I tackle only the major arguments against p-values. The first arguments are those showing they have no or sketchy justification, that their use reflects, as Neyman originally said, acts of will; that their use is even fallacious. These will be less familiar to most readers. The second set of arguments assume the use of p-values, but show the severe limitations arising from that use. These are more common. Why p-values seem to work is also addressed. When they do seem to work it is because they are related to or proxies for the more natural predictive probabilities.

The emphasis in this paper is philosophical not mathematical. Technical mathematical arguments and formula, though valid and of interest, must always assume, tacitly or explicitly, a philosophy. If the philosophy on which a mathematical argument is based is shown to be in error, the "downstream" mathematical arguments supposing this philosophy are thus not independent evidence for or against p-values, and, whatever mathematical interest they may have, become irrelevant.

2 Arguments Against P-Values

2.1 Fisher's Argument

A version of an argument given first by Fisher appears in every introductory statistics book. The original argument is this, [33]:

> Belief in a null hypothesis as an accurate representation of the population sampled is confronted by a logical disjunction: Either the null hypothesis is false, or the p-value has attained by chance an exceptionally low value.

A logical disjunction would be a proposition of the type "Either it is raining or it is not raining." Both parts of the proposition relate to the state of rain. The proposition "Either it is raining or the soup is cold" is a disjunction, but not a logical one because the first part relates to rain and the second to soup. Fisher's "logical disjunction" is evidently not a logical disjunction because the first part relates to the state of the null hypothesis and the second to the p-value.

Fisher's argument can be made into a logical disjunction, however, by a simple fix. Restated: Either the null hypothesis is false and we see a small p-value, or the null hypothesis is true and we see a small p-value. Stated another way, "Either the null hypothesis is true or it is false, and we see a small p-value." The first clause of this proposition, "Either the null hypothesis is true or it is false", is a tautology, a necessary truth, which transforms the proposition to (loosely) "TRUE and we see a small p-value." Adding a logical tautology to a proposition does not change its truth value; it is like multiplying a simple algebraic equation by 1. So, in the end, Fisher's dictum boils down to: "We see a small p-value."

In other words, in Fisher's argument a small p-value has no bearing on any hypothesis (any hypothesis unrelated to the p-value itself, of course). Making a decision about a parameter or data because the p-value takes any particular value is thus always fallacious: it is not justified by Fisher's argument, which is a non sequitur. The decision made using p-values may be serendipitously correct, of course, as indeed any decision based on any criterion might be. Decisions made by researchers are often likely correct because experimenters are good at controlling their experiments, and because (as we will see) the p-value is a proxy for the predictive probability, but if the final decision is dependent on a p-value it is reached by a fallacy. It becomes a pure act of will.

2.2 All P-Values Support the Null?

Frequentist theory claims that, assuming the truth of the null, we can equally likely see any p-value whatsoever, i.e. the p-value under the null is uniformly distributed. That is, assuming the truth of the null, we deduce we can see any p-value between 0 and 1. It is thus asserted the following proposition is true:

$$\text{If the null is true, then } p \in (0, 1). \tag{2}$$

where the bounds may or may not be not sharp, depending on one's definition of probability.

We always do see any value between 0 and 1, and so it might seem that *any* p-value confirms the null. But it is not a formal argument to then say that the null is true, which would be the fallacy of affirming the consequent.

Assume the bounds on the p-value's possibilities are sharp, i.e. $p \in [0, 1]$. Now it is not possible to observe a p-value *except* in the interval $[0, 1]$. So that if the null hypothesis is judged true a fallacy of affirming the consequent is committed, and if the null is rejected, i.e. judged false, a non sequitur fallacy is committed. It does not follow from the premise (2) that any particular p-value confirms the falsity (or unlikelihood) of the null.

If the bounds were not sharp, and a p-value *not* in $(0, 1)$ was observed, then it would logically follow that the null would be false, from the classic modus tollens argument. That is, if either $p = 0$ or $p = 1$, which can occur in practice (given obvious trivial data sets), then it is not true that the null is true, which is to say, the null would be false. But that means an observed $p = 1$ would declare the null false! The only way to validly declare the null false, to repeat, would be if $p = 0$ or $p = 1$, but as mentioned, this doesn't happen except in trivial cases. Using any other value to reject the null does not follow, and thus any decision is again fallacious.

Other than those two extreme cases, then, any observed $p \in (0, 1)$ says nothing logically about the null hypothesis. At no point in frequentist theory is it proved that

$$\text{If the null is false, then } p \text{ is wee.} \tag{3}$$

Indeed, as just mentioned, all frequentist theory states is (2). Yet practice, and not theory, insists small p-value are evidence the null is false. Yet not quite "not false", but "not true". It is said the null "has not been falsified." This is because of Fisher's reliance on the then popular theory of Karl Popper that propositions could never be affirmed but only falsified; see [34] for a discussion of Popper's philosophy, which is now largely discredited among philosophers of science, e.g. [35].

2.3 Probability Goes Missing

Holmes [36] wrote "Data currently generated in the fields of ecology, medicine, climatology, and neuroscience often contain tens of thousands of measured variables. If special care is not taken, the complexity associated with statistical analysis of such data can lead to publication of results that prove to be irreproducible." These words every statistician will recognize as true. They are true because of the use of p-values and hypothesis testing.

Holmes defines the use of p-values in the following very useful and illuminating way:

> Statisticians are willing to pay “some chance of error to extract knowledge” (J.W. Tukey) using induction as follows.
> “If, given A $\implies$ B, then the existence of a small ϵ such that $P(B) < \epsilon$ tells us that A is probably not true.”
> This translates into an inference which suggests that if we observe data X, which is very unlikely if A is true (written $P(\text{X}|\text{A}) < \epsilon$), then A is not plausible.

The last sentence had the following footnote: “We do not say here that the probability of A is low; as we will see in a standard frequentist setting, either A is true or not and fixed events do not have probabilities. In the Bayesian setting we would be able to state a probability for A.”

We have just seen in (2) (A $\implies$ B in Holmes's notation) that because the probability of B (conditional on what?) is low, it most certainly does not tell us A is probably not true. Nevertheless, let us continue with this example.

In my notation, Holmes's statement translates to this:

$$\Pr\left(\text{A}|\text{X}\,\&\,\Pr(\text{X}|\text{A}) = \text{small}\right) = \text{small}. \qquad (4)$$

This equation is equally fallacious. First, under the theory of frequentism the statement “fixed events do not have probabilities” is true. Under objective Bayes and logical probability anything can have a probability: under these systems, the probability of *any* proposition is always conditional on assumed premises. Yet every frequentist acts as if fixed events *do* have probabilities when they say things like “A is not plausible.” *Not plausible* is a synonym for *not likely*, which is a synonym for *of low probability*. In other words, every time a frequentist uses a p-value, he makes a probability judgment, which is forbidden by the theory he claims to hold. In frequentist theory A has to believed or rejected with certainty. Any uncertainty in A, quantified or not, is, as Holmes said, forbidden.

Frequentists may believe, if they like, that singular events like A cannot have probabilities, but then they cannot, via a back door trick using imprecise language, give A a (non-quantified) probability after all. This is an inconsistency.

Let that pass and consider more closely (4). It helps to have an example. Let A be the theory “There is a six-sided object that when activated must show one of the six sides, just one of which is labeled 6.” And, for fun, let X = “6 6s in a row.” We are all tired of dice examples, but there is still some use in them (and here we do not have to envisage a real die, merely a device which takes one of six states). Given these facts, Pr(X|A) = small, where the value of “small” is much weer than the magic number (it's about 2×10^{-5}). We want

$$\Pr\left(\text{A}|6\ 6\text{s on six-sided device}\,\&\,\Pr(6\ 6\text{s}|\text{A}) = 2 \times 10^{-5}\right) = ? \qquad (5)$$

It should be obvious there is no (direct) answer to (5). That is, unless we magnify some implicit premise, or add new ones entirely.

The right-hand-side (the givens) tell us that if we accept A as true, then 6 6s are a possibility; and so when we see 6 6s, if anything, it is evidence in favor of A's truth. After all, something that A said could happen did happen. An implicit premise might be that in noticing we just rolled 6 6s in a row, there were other possibilities beside A we should consider. Another implicitly premise is that we notice we can't identify the precise *causes* of the 6s showing (this is just some mysterious device), but we understand the causes must be there and are, say, related to standard physics. These implicit premises can be used to infer A. But they cannot reject it.

We now come to the classic objection, which is that no alternative to A is given. A is the only thing going. Unless we add new implicit premises to (5) that give us a hint about something beside A. Whatever this premise is, it cannot be "Either A is true or something else is", because that is a tautology, and in logic adding a tautology to the premises changes nothing about the truth status of the conclusion.

Now if you told a frequentist that you were rejecting A because you just saw 6 6s in the row, because "another number is due", he'd probably (rightly) accuse you of falling prey to the gambler's fallacy. The gambler's fallacy can only be judged were we to add more information to the right hand side of (5). This is the key. *Everything* we are using as evidence for or against A goes on the right hand side of (5). Even if it is not written, it is there. This is often forgotten in the rush to make everything mathematical and quantitative.

In our case, to have any evidence of the gambler's fallacy would entail adding evidence to the RHS of (5) that is similar to "We're in a casino, where I'm sure they're careful about the dice, replacing worn and even 'lucky' ones; plus, the way they make you throw the dice make it next to impossible to physically control the outcome." That, of course, is only a small summary of a large thought. All evidence that points to A or away from it that we consider is there on the right hand side, even if it is, I stress again, not formalized.

For instance, suppose we're on 34th street in New York City at the famous Tannen's Magic Store and we've just seen the 6 6s, or even 20 6s, or however many you like, by some dice labeled "magic". What of the probability then? The RHS of (5) in *that* situation changes dramatically, adding possibilities other than A, by implicit premise.

In short, it is not the observations alone in (5) that get you anywhere. It is the extra information we add that does the trick, as it were. Most important of all—and this cannot be overstated—*whatever* is added to (5), then (5) *is no longer* (5), *but something else*! That is because (5) specifies all the information it needs. If we add to the right hand side, we change (5) into a new equation.

Once again it is shown there is no justification for p-values, except the appeal to authority which states wee p-values cause rejection.

2.4 An Infinity of Null Hypotheses

An ordinary regression model is written $\mu = \beta_0 x_1 + \cdots + \beta_0 x_p$, where μ is the central parameter of the normal distribution used to quantify uncertainty in the observable. Hypothesis tests help hone the eventual list of measures appearing on the right hand side. The point here is not about regression *per se*, but about all probability models; regression is a convenient, common, and easy example.

For every measure included in a model, an infinity of measures have been tacitly excluded, exclusions made without benefit of hypothesis tests. Suppose in a regression the observable is patient weight loss, and the measures the usual list of medical and demographic states. One potential measure is the preferred sock color of the third nearest neighbor from the patient's main residence. It is a silly measure because, we judge using outside common-sense knowledge, that this neighbor's sock color cannot have any causal bearing on our patient's weight loss. The point is not that nobody would add such a measure—nobody would—but that it could have been but was excluded without the use of hypothesis testing.

Sock color *could* have been measured and incorporated into the model. That it wasn't proves two things: (1) that inclusion and exclusion of measures in models can and are made without guidance of p-values and hypothesis tests, and (2) since there are an infinity of possible measures for every model, we always must make many judgments without p-values. There is no guidance in frequentist (or Bayesian) theory that says use p-values here, but use your judgment there. One man will insist on p-values for a certain X, and another will use judgment. Who is right? Why not use p-values everywhere? Or judgment everywhere? (The predictive method uses judgment aided by probability and decision.)

The only measures put into models are those which are at least suspected to be in the "causal path" of the observable. Measures which may, in part, be directly involved with the efficient and material cause of the observable are obvious, such as adding sex to medical observable models, because it is known differences in biological sex cause different things to happen to many observables. But those measures which might cause a change in the direct partial cause, or a change in the change and so on, like income in the weight loss model, also naturally find homes (income does not directly cause weight loss, but might cause changes which in turn cause others etc. which cause weight loss). Sock color belongs to this chain only if we can tell ourselves a just-so story of how this sock color can cause changes in other causes etc. of eventual causes of the observable. This can *always* be done: it only takes imagination.

The (initial) knowledge or surmise of material or efficient causes comes from *outside* the model, or the evidence of the model. Models begin with the assumption of measures included in the causal chain. A wee p-value does not, however, confirm a cause (or cause of a cause etc.) because non-causal correlations happen. Think of seeing a rabbit in a cloud. P-values, at best (see the Sect. 3 below) highlight large correlations.

It is also common that measures with small correlations, i.e. with large p-values, where there are known, or highly suspected, causal chains between the X and Y are not expunged from models; i.e. they are kept regardless what they p-value said. These are yet more cases where p-values are ignored.

The predictive approach is agnostic about cause: it accepts conditional hypotheses and surmises and outside knowledge of cause. The predictive approach simply says the best model is that which makes the best verified predictions.

2.5 Non-unique Adjustments

This criticism is similar to the infinity of hypotheses. P-values are often adjusted for multiple tests using methods like Bonferroni corrections. There are no corrections for those hypotheses rejected out of hand without the benefit of hypothesis tests.

Corrections are not used consistently. For instance, in model selection and in interim analyses, which is often informal. How many working statisticians have heard the request, “How much more data do I need to get significance?” It is, of course, except under the most controlled situations, impossible to police abuse. This is contrasted with the predictive method, which reports the model in a form which can be verified by (theoretically) anybody. So that even if abuse, such as confirmation bias, was used in building the model, it can still be checked. Confirmation bias using p-values is easier to hide. The predictive method does not assume a true model in the frequentist senses: instead, all models are conditional on the premises, evidence, and data assumed.

Harrell [20] says, “There remains controversy over the choice of 1-tailed vs. 2-tailed tests. The 2-tailed test can be thought of as a multiplicity penalty for being potentially excited about either a positive effect or a negative effect of a treatment. But few researchers want to bring evidence that a treatment harms patients... So when one computes the probability of obtaining an effect larger than that observed if there is no true effect, why do we too often ignore the sign of the effect and compute the (2-tailed) p-value?”

The answer is habit married to the fecundity of two-tailed tests at producing wee p-values.

2.6 P-Values Cannot Identify Cause

Often when a wee p-value is seen in accord with some hypothesis, it will be taken as implying that the cause, or one of the causes, of the observable has been verified. But p-values cannot identify cause; see [37] for a full discussion. This is because parameters inside probability models are not (or almost never) representations of cause, thus any decision based upon parameters cannot confirm nor deny any cause. Regression model parameters in particular are not representations of cause.

It helps to have a semi-fictional example. Third-hand smoking, which is not fictional [38], is when items touched by second-hand smokers, who have touched things by first-hand smokers, are in turn touched by others, who become “third-hand smokers”. There is no reason this chain cannot be continued indefinitely. One gathers data from x-hand smokers (which are down the touched-smoke chain somewhere) and non-x-hand smokers and the presence or absence of a list of maladies. If in some parameterized model relating these a wee p-value is found for one of the maladies, x-hand smoking will be said to have been “linked to” the malady. This “linked to” only means a “statistically significant result” was found, which in turn only means wee p-value was seen.

Those keen on promoting x-hand smoking as causing the disease will take the "linked to" as statistical validation of cause. Careful statisticians won't, but stopping the causal interpretation from being used is by now an impossible task. This is especially so when even statisticians use "linked to" without carefully defining it.

Now if x-hand smoking caused the particular disease, then it would always do so, and statistical testing would scarcely be needed to ascertain this because each individual exposed to the cause would be always contract the disease—unless the cause were blocked. What blocks this cause could be various, such as a person's particular genetic makeup, or state of hand calluses (to block absorption of x-hand smoke), or whether a certain vegetable was eaten (that somehow cancels out the effect of x-hand smoke), and so on. If these blocking causes were known (the blocks are also causes), again statistical models would not be needed, because all we would need know is whether any x-hand-smoke-exposed individual had the relevant blocking mechanism. Each individual would get the disease for certain unless he had (for certain) a block.

Notice that (and also see below the criticism that p-values are not always believed) models are only tested when the causes or blocks are not known. If causes were known, then models would not be needed. In many physical cases, cause or block can be demonstrated by "bench" science, and then the cause or block becomes known with certainty. It may not be known how this cause or block interacts or behaves in the face of multiple other potential causes or blocks, of course. Statistical models can be used to help quantify this kind of uncertainty, given appropriate experiments. But then this cause or block would not be added or expunged from a model regardless of the size of its p-value.

It can be claimed hypothesis tests are only used where causes or blocks are unknown, but testing cannot confirm unknown causes or blocks.

2.7 P-Values Aren't Verified

One reason for the reproducibility crisis is the presumed finality of p-values. Once a "link" has been "validated" with a wee p-value, it is taken by most to mean the "link" definitely exists. This thinking is enforced since frequentist theory forbids assigning a probability measure to any "link's" veracity. The wee-p-confirmed "link" enters the vocabulary of the field. This thinking is especially rife in purely statistically driven fields, like sociology, education, and so forth, where direct experimentation to identify cause is difficult or impossible.

Given the ease of finding wee p-values, it is no surprise that popular theories are not re-validated when in rare instances they are attempted to be replicated. And then not every finding can be replicated at least because of the immense cost and time involved. So, many spurious "links" are taken as true or causal.

Using Bayes factors, or adjusting the magic number lower, would not solve the inherent problem. Only verifying models can, i.e. testing them against reality. When a civil engineer proposes a new theory for bridge construction, testing via simulation and incorporating outside causal knowledge provides guidance whether the new bridge built using the theory will stand or fall. But even given a positive judgment from this process does not mean the new bridge will stand. The only way to know with any certainty is to build the bridge and see. And, as readers will know, not every new bridge does stand. Even the best considered models fail.

What is true for bridges is true for probability models. P-value-based models are never verified against reality using new, never before seen or used in any way data. The predictive approach makes predictions that can, and must, be verified. Whatever measures are assumed results in probabilistic predictions about the observable. These predictions can be checked in theory by anybody, even without having the data which built the model, in the same way even a novice driver can understand whether the bridge under him is collapsing or not. How verification is done is explained elsewhere. e.g. [26,32,39–41].

A change in practice is needed. Models should only be taken as preliminary and unproved until they can be verified using outside, never-before-seen or used data. Every paper which uses statistical results should announce "This model has not yet been verified using outside data and is therefore unproven." The practice of printing wee p-values, announcing "links", and then moving on to the next model must end. This would move statistics into the realm of the harder sciences, like physics and chemistry, which take pains to verify all proposed models.

2.8 P-Values Are Not Unique

We now begin the more familiar arguments against p-values, with some added insight. As all know, the p-value is never unique, and is dependent on *ad hoc* statistics. Statistics themselves are not unique. The models on which the statistics are computed are, with very rare exceptions in practice, also *ad hoc*; thus, they are not unique. The rare exceptions are when the model is deduced from first principles, and are therefore parameter-free, obviating the need for hypothesis testing. The simplest examples of fully deduced models are found in introductory probability books. Think of dice or urn examples. But then nobody suggests using p-values on these models.

If in any parameterized model the resulting p-value is not wee, or otherwise has not met the criteria for publishing, then different statistics can be sought to remedy the "problem." An amusing case found its way into the *Wall Street Journal*, [42].

The paper reported that Boston Scientific (BS) introduced a new stent called the Taxus Liberte. The company did the proper experiments and analyzed their data using a Wald test. This give them a p-value that was just under the magic number, a result which is looked upon with favor by the Food and Drug Administration. But a competitor charged that the Wald statistic is not one they would have used. So they hired their own statistician to reevaluate their rival's data. This statisticians computed p-values for several other statistics and discovered each of these were a fraction larger than the magic number. This is when the lawyers entered the story, and where we exit it.

Now the critique that the model and statistic is not unique must be qualified. Under frequentism, probability is said to exist unconditionally; which is to say, the moment a parameterized model is written—somehow, somewhere—at "the limit" the "actual" or "true" probability is created. This theory is believed even though alternate parameterized models for the same observable may be created, which in turn create their own "true" values of parameters. All rival models and parameters are thus "true" (at the limit), which is a contradiction. This is further confused if probability is believed to be ontic, i.e. actually existing as apples or pencils exist. It would seem that rival models battle over probability somehow, picking one which is the truly true or really true model (at the limit).

Contrast this with the predictive approach, which accepts all probability is conditional. Probability at the limit may never need be referenced. All is allowed to remain finite (asymptotics can of course be used as convenient approximations). Changing any assumptions changes the model by definition, and all probability is epistemic. Different people using different models, or even using the same models, would come to different conclusions quite naturally.

2.9 The Deadly Sin of Reification

If in some collection of data a difference in means between two groups is seen, this difference is certain (assuming no calculation mistakes). We do not need to do any tests to verify whether the difference is real. It was seen: it is real. Indeed, any question that can be asked of the observed data can be answered with a simple yes or no. Probability models are not needed.

Hypothesis testing acknowledges the observed difference, but then asks whether this difference is "really real". If the p-value is wee, it is; if not, the observed real difference is declared not really real. It will even be announced (by most) "No difference was found", a very odd thing to say. If it does not sound odd to your ears, it shows how successful frequentist theory is. The attitude that actual difference is not really real comes from assuming probability is ontic, that we have only sampled from an infinite reality where the model itself is larger and realer than the observed data. The model is said to have "generated" the value in some vague way, where the notion of the causal means by which the model does this forever recedes into the distance the more it is pursued. The model is reified. It becomes better than reality.

The predictive method is, as said, agnostic about cause. It takes the observed difference as real and given and then calculates the chance that such differences will be seen in *new* observations. Predictive models can certainly err and can be fooled by spurious correlations just as frequentist ones can (though far less frequently). But the predictive model asks to be verified: if it says differences will persist, this can be checked. Hypothesis tests *declare* they will be seen (or not), end of story.

If the difference is observed but the p-value not wee, it is declared that chance or randomness *caused* the observed difference; other verbiage is to say the observed difference is "due to" chance, etc. This is causal language, but it is false. Chance and randomness do not exist. They are purely epistemic. They therefore cannot cause anything. Some thing or things caused the observed difference. But it cannot have been chance. The reification of chance comes, I believe, from the reluctance of researchers to say, “I have no idea what happened.”

If *all*—and I mean this word in its strictest sense—we allow is X as the potential cause (or in the causal path) of an observed difference, then we must accept that X is the cause regardless of what a p-value says to do with X (usually, of course, the parameter associated with X). We can say “Either X is the cause or something else is”, but this will always be true, even in the face of knowledge X is not a cause. This argument is only to reinforce the idea that knowledge of cause must come from outside the probability model. Also that chance is never a cause. And that any probability model that gives non-extreme predictive probabilities is always an admission that we do not know all the causes of the observable. This is true (and for chance and randomness, too) even for quantum mechanical observations, the discussion of which would take us too far afield here. But see [26], Chap. 5 for a discussion.

2.10 P-Values Are Magic

Every working statistician will have a client who has been reduced to grief after receiving the awful news that the p-value for their hypothesis was larger than the magic number, and therefore unpublishable. “What can we do to make it smaller?” ask many clients (I have had this happen many times). All statisticians know the tricks to oblige this request. Some do oblige.

Gigerenzer [8] calls p-value hunting a ritualized approach to doing science. As long as the proper (dare we say magic) formulas are used and the p-values are wee, science is said to have been done. Yet is there any practical, scientific difference between a p-value of 0.49 and 0.051? Are the resulting post-model decisions made always so finely tuned and hair-breadth crucial that the tiny step between 0.49 and 0.51 throws everything off balance? Most scientists, and all statisticians, will say no. But most will act as if the answer is yes. A wee p-value is mesmerizing.

The counter-argument to abandoning p-values in the fact of this criticism is better education. But that education would have to overcome decades of beliefs and actions that the magic number is in fact magic. The word preferred is not *magic*, of course, but *significant.* Anyway, this educational initiative would have to cleanse all books and material that bolsters this belief, which is not possible.

2.11 P-Values Are Not Believed When Convenient

In any given set of data, with some parameterized model, its p-value are assumed true, and thus the decisions based upon them sound. Theory insists on this. The decisions “work”, whether the p-value is wee or not wee.

Suppose a wee p-value. The null is rejected, and the “link” between the measure and the observable is taken as proved, or supported, or believable, or whatever it is “significance” means. We are then directed to act as if the hypothesis is true. Thus if it is shown that per capita cheese consumption and the number of people who died tangled in their bed sheets are “linked” via a wee p, we are to believe this. And we are to believe all of the links found at the humorous web site Spurious Correlations, [43].

I should note that we can either accept that grief of loved ones strangulated in their beds drives increased cheese eating, *or* that cheese eating causes sheet strangulation. This is joke, but also a valid criticism. The direction of causal link is not mandated by the p-value, which is odd. That means the direction comes from outside the hypothesis test itself. Direction is thus (always) a form of prior information. But prior information like this is forbidden in frequentist theory. Everybody dismisses, as they should, these spurious correlations, but they do so using prior information. They are thus violating frequentist theory.

Suppose next a non-wee p-value. The null has been "accepted" in any practical sense. There is the idea, started by Fisher, that if the p-value was not wee that one should collect more data, and that the null is not accepted but that we have failed to reject it. Collecting more data will lead to a wee p-value eventually, even when the correlations are spurious (this is a formal criticism, given below). Fisher did not have in mind spurious correlations, but genuine effects, where he took it the parameter represented something real in the causal chain of the observable. But this is a form of prior information, which is forbidden because it is independent (I use this word in its philosophical not mathematical sense) of the p-value. The p-value then becomes a self-fulfilling prophecy. It must be, because we started by declaring the effect was real. This practice does not make any finding false, as Cohen pointed out [9]. But if we knew the effect was real before the p-value was calculated, we know it even after. And we reject the p-values that do not conform to our prior knowledge. This, again, goes against frequentist theory.

2.12 P-Values Base Decisions on What Did Not Occur

P-values calculate the probability of what did not happen on the assumption that what did not happen should be rare. As Jefferys [44] famously said: "What the use of P[-value] implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred."

Decisions should instead be conditioned of what did happen and on uncertainty in the observable itself, and not on parameters (or functions of them) inside models.

2.13 P-Values Are Not Decisions

If the p-value is wee, a decision is made to reject the null hypothesis, and vice versa (ignoring the verbiage "fail to reject"). Yet the consequences of this decision are not quantified using the p-value. The decision to reject is just the same, and therefore just as consequential, for a p-value of 0.05 as one of 0.0005. Some have the habit of calling especially wee p-values as "highly significant", and so forth, but this does not accord with frequentist theory, and is in fact forbidden by that theory because it seeks a way around the proscription of applying probability to hypotheses. The p-value, as frequentist theory admits, is not related in any way to the probability the null is true or false. Therefore the size of the p-value does not matter. Any level chosen as "significant" is, as proved above, an act of will.

A consequence of the frequentist idea that probability is ontic and that true models exist (at the limit) is the idea that the decision to reject or accept some hypothesis should be the same for all. Steve Goodman calls this idea "naive inductivism", which is "a belief that all scientists seeing the same data should come to the same conclusions," [45]. That this is false should be obvious enough. Two men do not always make the same bets even when the probabilities are deduced from first principles, and are therefore true. We should not expect all to come to agreement on believing a hypothesis based on tests concocted from *ad hoc* models. This is true, and even stronger, in a predictive sense, where conditionality is insisted upon.

Two (or more) people can come to completely different predictions, and therefore difference decisions, even when using the same data. Incorporating decision in the face of uncertainty implied by models is only partly understood. New efforts along these lines using quantum probability calculus, especially in economic decisions, are bound to pay off, see e.g. [46].

A striking and in-depth example of how using the same model and same data can lead people to *opposite* beliefs and decisions is given by Jaynes in his chapter "Queer uses for probability theory", [30].

2.14 No One Remembers the Definition of P-Values

The p-value is (usually) the conditional probability an *ad hoc* test statistic being larger (in absolute value) than the observed statistic, assuming the null hypothesis is true, given the values of the observed data, and assuming the truth of the model. The probability of exceeding the test statistic assuming the alternate hypothesis is true, or given the null hypothesis is false, given the other conditions, is not known. Nor is the second-most important probability known: whether or not the null hypothesis is true.

It is the second-most important probability because most null hypotheses are "point nulls", because continuous parameters take fixed single values, which because parameters live on the continuum, "points" have a probability of 0. The most important probability, or rather probabilities, is that of Y given X, and Y given X's absence, where it is assumed (as with p-values) X is part of the model. This is a direct measure of relevance of X. If the conditional probability of Y given X (in the model) is a, and the probability of Y given X's absence is also a, then X is irrelevant, conditional on the model and other information listed in (1). If X is relevant, the difference in probabilities because a matter of individual decision, not a mandated universal judgment, as with p-values.

Now frequentists do not accept the criticism of the point null having zero probability, because according to frequentist theory parameters (the uncertainty in them) do not have probabilities. Again, once any model is written, parameters come into existence (somehow) as some sort of Platonic form at the limit. They take "true" values there; it is inappropriate in the theory to use probability to express uncertainty in their unknown values. Why? It is not, after all, thought wrong to express uncertainty in unknown observables using probability. The restriction to probability only on observables has no satisfactory explanation: the difference just exists by declaration. See [47–49] for these and other unanswerable criticisms of frequentist theories (including those in the following paragraphs) well known to philosophers, but somehow more-or-less unknown to statisticians.

Rival models, i.e. those with different parameterizations (Normal versus Weibull model, say) somehow create parameters, too, which are also "true". Which set of parameters are the truest? Are all equally true? Or are all models merely crude approximations to *the* true model which nobody knows or can know? Frequentists might point to central limit theorems to answer these questions, but it is not the case all rival models converge to the same limit, so the problem is not solved.

Here is one of a myriad of examples showing failing memories, from a paper whose intent is to teach proper p-value use: [50] says, "The p value is the probability to obtain an effect equal to or more extreme than the one observed presuming the null hypothesis of no effect is true; it gives researchers a measure of the strength of evidence against the null hypothesis."

The p-value is mute on the size of an effect (and also on what an effect is; see above). And though it is widely believed, this conclusion is false, accepting the frequentist theory in which p-values are embedded. "Strength" is not a measure of probability, so just what is it? It is never defined formally inside frequentist theory. The discussion below on why p-values sometimes seem to work is relevant here.

2.15 Increasing the Sample Size Lowers P-Values

Large and increasing sample sizes show low and lowering p-values. Even small differences become "significant" eventually. This is so well known there are routine discussions warning people to, for instance, not conflate clinical versus statistical "significance", e.g. [51]. What is statistical significance? A wee p-value. And what is a wee p-value? Statistical significance.

Suppose the uncertainty in some observable y_0 in a group 0 is characterized by a normal distribution with parameters $\theta_0 = a$ and with a σ also known; and suppose the same for the observable y_1 in a group 1, but with $\theta_1 = a + 0.00001$. The groups represent, say, the systolic blood pressure measures of people who live on the same block but with even (group 0) and odd (group 1) street addresses. We are in this case *certain* of the values of the parameters. Obviously, $\theta_1 - \theta_0 = 0.00001$ with certainty. P-values are only calculated with observed measures, and here there are none, but since there is a certain difference, we would expect the "theoretical" p-value to be precisely 0. As it would be for *any* sized difference in the θs.

This by itself is not especially interesting, except that it confirms low p-values can be found for small differences, which here flows from the knowledge of the true difference in the parameters. The p-value would (or should) in these cases always be "significant".

Now a tradition has developed to call the difference in parameters the "effect size", borrowing language used by physicists. In physics (and similar fields) parameters are often written as direct or proxy causes and can then be taken as effects. This isn't the case for the vast, vast majority of statistical models. Parameters are not ontic or causal effects. They represent only changes in our epistemic knowledge.

This is a small critique, but the use of p-values, since they are parameter-centric, encourages this false view of effect. Parameter-focused analyses of any kind always exaggerates the certainty we have in any measure and its epistemic influence on the observable. We can have absolute certainty of parameter values, as in the example just given, but that does not translate into large differences in the probability of new differences in the observable. If that example, $\Pr(\theta_1 > \theta_0|\mathrm{DMA}) = 1$, but for most scenarios $\Pr(\mathrm{Y}_1 > \mathrm{Y}_0|\mathrm{DMA}) \approx 0.5$. That means frequentist point estimates bolstered by wee p-values, or Bayesians parameter posteriors, all exaggerate evidence. Given that nearly all analyses are parameter-centric, we do not only have a reproducibility crisis, we have an over-certainty crisis.

2.16 It Ain't Easy

Tests for complicated decisions do not always exist; the further we venture from simple models and hypotheses, the more this is true. For instance, how to test whether groups 3 or 4 exceed some values but not group 1 when there is indifference about group 2, and where the values depend in some way on the state of other measures (say, these other measures being in some range)?

This is no problem at all for predictive statistics. Any question that can be conceived, and can theoretically be measured, can be formulated in probability in a predictive model.

P-values also make life too easy for modelers. Data is "submitted" to software (a not uncommon phrase), and if wee p-values are found, after suitable tweaking, everybody believes their job is done. I don't mean that researchers don't call for "future work", which they will always do, but the belief that the model has been sufficiently proved. That the model just proposed for, say, this small set of people existing in one location for a small time out of history, and having certain attributes, somehow then applies to all people everywhere. This is not *per se* a p-value criticism, but p-values do make this kind of thinking easy.

2.17 The P-Value for What?

Neyman fixed "test level", which is practically identical with p-values fixed at the magic number, are for tests on the whole, and not for the test at hand, which is itself in no way guaranteed to have a Type I or even Type II error level. These numbers (whatever they might mean) apply to infinite sets of tests. And we haven't got there yet.

2.18 Frequentists Become Secret Bayesians

That is because people argue: For most small p-values I have seen in the past, I believe the null has been false (and vice versa); I now see a new small p-value, therefore the null hypothesis in this new problem is likely false. That argument works, but it has no place in frequentist theory (which anyway has innumerable other difficulties). It is the Bayesian-like interpretation. Newman's method is to accept with finality the decisions of the tests as certainty. But people, even ardent frequentists, cannot help but put probability, even if unquantified, on the truth value of hypotheses. They may believe that by omitting the quantification and only speaking of the truth of the hypothesis as "likely", "probable" or other like words, that they have not violated frequentist theory. If you don't write it down as math, it doesn't count! This is, of course, false.

3 If P-Values Are Bad, Why Do They Sometimes Work?

3.1 P-Values Can Be Approximations to Predictive Probability

Perhaps the most-used statistic is the t (and I make this statement without benefit of a formal hypothesis test, you notice, and you understood it without one, too), which is in its numerator the mean of one measure minus the mean of a second. The more the means of measures under different groups differ, the smaller the p-value will in general be, with the caveats about standard deviations and sample sizes understood.

Now consider the objective Bayesian or logical probability interpretation of the same observations, taken in a predictive sense. The probability the measure with the larger observed mean exhibits in new data larger values than the measure with the smaller mean increases the larger t is (with similar caveats). That is, loosely,

$$\text{As } \ t \to \infty, \quad \Pr(Y_2 > Y_1 | \text{DMA}, t) \to 1, \tag{6}$$

where D is the old data, M is a parameterized model with its host of assumptions (such as about the priors) A, and t the t-statistic for the two groups Y_2 and Y_1, assuming the group 2 has the larger observed mean. As t increases, so does in general the probability Y_2 will be larger than Y_1, again with the caveats understood (most models will converge not to 1, but to some number larger than 0.5 less than 1). Since this is a predictive interpretation, the parameters have been "integrated out." (In the observed data, it will be *certain* if the mean of one group was larger than the other.) This is an abuse of notation, since t is derived from D. It is also a cartoon equation meant only to convey a general idea; it is, as is obvious enough, true in the normal case (assuming finite variance and conjugate or flat priors).

What (6) says is that the p-value in this sense is a proxy for the predictive probability. And it's the predictive probability all want, since again there is no uncertainty in the past data. When p-values work, they do so because they are representing reasonable predictions about future values of the observables.

This is only rough because those caveats become important. Small p-values, as mentioned above, are had just by increasing sample size. With a fixed standard deviation, and miniscule difference between observed means, a small p-value can be got by increasing the sample size, but the probability the observables differ won't budge much beyond 0.5.

Taking these caveats into consideration, why not use p-values, since they, at least in the case of t- and other similar statistics, can do a reasonable job approximating the magnitude of the predictive probability? The answer is obvious: since it's easy to get, and it is what is desired, calculate the predictive probability instead of the p-value. Even better, with predictive probabilities none of the caveats must be worried about: they take care of themselves in the modeling. There will be no need of any discussions about clinical versus statistical significance. Wee p-values can lead to small or large predictive probability differences. And all we need are the predictive probability differences.

The interpretation of predictive probabilities is also natural and easy to grasp, a condition which is certainly false with p-values. If you tell a civilian, "Given the experiment, the probability your blood pressure will be lower if you take this new drug rather than the old is 70%", he'll understand you. But if you tell him that if the experiment were repeated an infinite number of times, and if we assume the new drug is no different than the old, then a certain test statistic in each of these infinite experiments will be larger than the one observed in the experiment 5% of the time, he won't understand you.

Decisions are easier and more natural—and verifiable—using predictive probability.

3.2 Natural Appeal of Some P-Values

There is a natural and understandable appeal to some p-values. An example is in tests of psychic abilities, [52]. An experiment will be designed, say guessing numbers from 1 to 100. On the hypothesis that no psychic ability is present, and the only information the would-be psychic has is that the numbers will be in a certain set, and where knowledge of successive numbers is irrelevant (each time it's 1–100, and it's not numbered balls in urns), then the probability of guessing correctly can be deduced as 0.01. The would-be psychic will be asked to guess more than once, and his total correct out of n is his score.

Suppose conditional on this information the probability of the would-be psychic's score assuming he is only guessing is some small number, say, much lower than the magic number. The lower this probability is, the more likely, it is thought, of the fellow having genuine psychic powers. Interestingly, a probability at or near the magic number in psychic would be taken by no one as conclusive evidence. The reason is that cheating and sloppy and misleading experiments are far from unknown. But those suspicions, while true, do not accord with p-value theory, which has no way to incorporate anything but quantifiable hypotheses (see the discussion above about incorporating prior information).

But never mind that. Let's assume no cheating. This probability of the score assuming guessing, or the probability of scores at least as large as the one observed, functions as a p-value. Wee ones are taken as indicating psychic ability, or at least as indicating psychic ability is likely. Saying ability is “likely” is forbidden under frequentist theory, as discussed above, so when people do this they are acting as predictivists. Nor can we say the small p-value confirms psychic powers are the cause of the results. Nor chance.

So what do the scores mean? Same thing batting averages do in baseball. Nobody bats a thousand, nor do we expect psychics to guess correctly 100% of the time. Abilities differ. Now a high batting average, say from Spring Training, is taken as a predictive of a high batting average in the regular season. This often does not happen—the prediction does not verify—and when it doesn’t Spring Training is taken as a fluke. The excellent performance during Spring Training will be put down to a variety of causes. One of these won’t be good hitting ability.

A would-be psychic’s high score is the same thing. Looks good. Something caused the hits. What? Could have been genuine ability. Let’s get to the big leagues and really put him to the test. Let magicians watch him. If the would-be psychic doesn’t make it there, and so far none have, then the prior performance just like in baseball will be ascribed to any number of causes, one of which may be cheating.

In other words, even when a p-value seems natural, it is again a proxy for a predictive probability or an estimate of ability assuming cause (but not proving it).

4 What Are the Odds of That?

As should be clear, many of the arguments used against p-values could for the most part also be used against Bayes factors. This is especially so if probability is taken as subjective (where a bad burrito can shift probabilities in any direction), where the notion of cause becomes murky. Many of the arguments against p-values can also be marshaled against using point (parameter) estimation. As said, parameter-based analyses exaggerates evidence, often to extent that is surprising, especially if one is unfamiliar with predictive output. Parameters are too often reified as “the” effects, when all they are, in nearly all probability models, are expressions of uncertainty in how the measure X affects the uncertainty in the observable Y. Why not then speak directly of the how changes in X, and not in some *ad hoc* uninteresting parameter, relate to changes in the uncertainty of Y? About the mechanics of how to decide which X are relevant and important in a model, I leave to other sources, as mentioned above.

People often quip, when seeing something curious, “What are the odds of that?” The probability of any observed thing is 1, conditional on its occurence. It happened. There is therefore no need to discuss its probability—*unless* one wanted to make predictions of future possibilities. Then the conditions on which the curious thing are stated dictate the probability. Different people can come to different conditions, and therefore come to different probabilities. As often happens. This isn’t so with frequentist theory, which must embed every event in some unique not-debatable infinite sequence in which, at the limit, probability becomes real and unchangeable. But nothing is actually infinite, only potentially infinite. It is these fundamental differences in philosophy that drive many of the criticisms of p-values, and therefore of frequentism itself. Most statisticians will not have read these arguments, given by authors like Hájek [47,49], Franklin [29,53], and Stove [54] (the second half of this reference). They are therefore urged to review them. The reader does not now have to believe frequentism is false, as these authors argue, to grasp the arguments against p-values above. But if frequentism is false, then p-values are ruled moot *tout court*.

A common refrain in the face of criticisms like these is to urge caution. "Use p-values wisely," it will be said, or use them "in the proper way." But there is no wise or proper use of p-values. They are not justified in any instance.

Some think p-values are justified by simulations which purport to show p-values behave as expected when probabilities are known. But those who make those arguments forget that there is nothing in a simulation that was not first put there. All simulations are self-fulfilling. The simulation said, in some lengthy path, that the p-value should look like this, and, lo, it did. There is also, in most cases, reification of probability in these simulations. Probability is taken as real, ontic. When all simulations do is manipulate known formulas given known and fully expected input. That it, simulations begin by stating that given an input u produce via this long path p. Except that semi-blind eyes are turned to u, which makes it "random", and therefore makes p ontic. This is magical thinking. I do not expect readers to be convinced by this telegraphic and wholly unfamiliar argument, given how common simulations are, so see Chap. 5 in [26] for a full explication. This argument will seem more shocking the more one is convinced probability is real.

Predictive probability takes the model not as true or real as in hypothesis testing, but as the best summary of knowledge available to the modeler (some models can be deduced from first principles, and thus have no parameters, and are thus true). Statements made about the model are therefore more naturally cautious. Predictive probability is no panacea. People can cheat and fool themselves just as easily as before, but the exposure of the model in a form that can be checked by anybody will propel and enhance caution. P-value-based models say 'Here is the result, which you must accept.' Rather, that is what theory directs. Actual interpretation often departs from theory dogma, which is yet another reason to abandon p-values.

Future work is not needed. The totality of all arguments insists that p-values should be retired immediately.

References

1. Neyman, J.: Philos. Trans. R. Soc. Lond. A **236**, 333 (1937)
2. Lehman, E.: Jerzy Neyman, 1894–1981. Technical report, Department of Statistics, Berkeley (1988)

3. Trafimow, D., Amrhein, V., Areshenkoff, C.N., Barrera-Causil, C.J., Beh, E.J., Bilgiç, Y.K., Bono, R., Bradley, M.T., Briggs, W.M., Cepeda-Freyre, H.A., Chaigneau, S.E., Ciocca, D.R., Correa, J.C., Cousineau, D., de Boer, M.R., Dhar, S.S., Dolgov, I., Gómez-Benito, J., Grendar, M., Grice, J.W., Guerrero-Gimenez, M.E., Gutiérrez, A., Huedo-Medina, T.B., Jaffe, K., Janyan, A., Karimnezhad, A., Korner-Nievergelt, F., Kosugi, K., Lachmair, M., Ledesma, R.D., Limongi, R., Liuzza, M.T., Lombardo, R., Marks, M.J., Meinlschmidt, G., Nalborczyk, L., Nguyen, H.T., Ospina, R., Perezgonzalez, J.D., Pfister, R., Rahona, J.J., Rodríguez-Medina, D.A., Romão, X., Ruiz-Fernández, S., Suarez, I., Tegethoff, M., Tejo, M., van de Schoot, R., Vankov, I.I., Velasco-Forero, S., Wang, T., Yamada, Y., Zoppino, F.C.M., Marmolejo-Ramos, F.: Front. Psychol. **9**, 699 (2018). https://doi.org/10.3389/fpsyg.2018.00699
4. Ziliak, S.T., McCloskey, D.N.: The Cult of Statistical Significance. University of Michigan Press, Ann Arbor (2008)
5. Greenland, S.: Am. J. Epidemiol. **186**, 639 (2017)
6. McShane, B.B., Gal, D., Gelman, A., Robert, C., Tackett, J.L.: The American Statistician (2018, forthcoming)
7. Berger, J.O., Selke, T.: JASA **33**, 112 (1987)
8. Gigerenzer, G.: J. Socio-Econ. **33**, 587 (2004)
9. Cohen, J.: Am. Psychol. **49**, 997 (1994)
10. Trafimow, D.: Philos. Psychol. **30**(4), 411 (2017)
11. Nguyen, H.T.: Integrated Uncertainty in Knowledge Modelling and Decision Making, pp. 3–15. Springer (2016)
12. Trafimow, D., Marks, M.: Basic Appl. Soc. Psychol. **37**(1), 1 (2015)
13. Nosek, B.A., Alter, G., Banks, G.C., et al.: Science **349**, 1422 (2015)
14. Ioannidis, J.P.: PLoS Med. **2**(8), e124 (2005)
15. Nuzzo, R.: Nature **526**, 182 (2015)
16. Colquhoun, D.: R. Soc. Open Sci. **1**, 1 (2014)
17. Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N., Altman, D.G.: Eur. J. Epidemiol. **31**(4), 337 (2016). https://doi.org/10.1007/s10654-016-0149-3
18. Greenwald, A.G.: Psychol. Bull. **82**(1), 1 (1975)
19. Hochhaus, R.G.A., Zhang, M.: Leukemia **30**, 1965 (2016)
20. Harrell, F.: A litany of problems with p-values (2018). http://www.fharrell.com/post/pval-litany/
21. Benjamin, D., Berger, J., Johannesson, M., Nosek, B., Wagenmakers, E., Berk, R., et al.: Nat. Hum. Behav. **2**, 6 (2018)
22. Mulder, J., Wagenmakers, E.J.: J. Math. Psychol. **72**, 1 (2016)
23. Hitchcock, C.: The Stanford Encyclopedia of Philosophy (Winter 2016 Edition) (2016). https://plato.stanford.edu/archives/win2016/entries/causation--probabilistic
24. Breiman, L.: Stat. Sci. **16**(3), 199 (2001)
25. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge (2000)
26. Briggs, W.M.: Uncertainty: The Soul of Probability, Modeling & Statistics. Springer, New York (2016)
27. Nuzzo, R.: Nature **506**, 50 (2014)
28. Begley, C.G., Ioannidis, J.P.: Circ. Res. **116**, 116 (2015)
29. Franklin, J.: Erkenntnis **55**, 277 (2001)
30. Jaynes, E.T.: Probability Theory: The Logic of Science. Cambridge University Press, Cambridge (2003)

31. Keynes, J.M.: A Treatise on Probability. Dover Phoenix Editions, Mineola (2004)
32. Briggs, W.M., Nguyen, H.T., Trafimow, D.: Structural Changes and Their Econometric Modeling. Springer (2019, forthcoming)
33. Fisher, R.: Statistical Methods for Research Workers, 14th edn. Oliver and Boyd, Edinburgh (1970)
34. Briggs, W.M.: arxiv.org/pdf/math.GM/0610859 (2006)
35. Stove, D.: Popper and After: Four Modern Irrationalists. Pergamon Press, Oxford (1982)
36. Holmes, S.: Bull. Am. Math. Soc. **55**, 31 (2018)
37. Briggs, W.M.: arxiv.org/abs/1507.07244 (2015)
38. Protano, C., Vitali, M.: Environ. Health Perspect. **119**, a422 (2011)
39. Briggs, W.M.: JASA **112**, 897 (2017)
40. Gneiting, T., Raftery, A.E., Balabdaoui, F.: J. R. Stat. Soc. Ser. B Stat. Methodol. **69**, 243 (2007)
41. Gneiting, T., Raftery, A.E.: JASA **102**, 359 (2007)
42. Winstein, K.J.: Wall Str. J. (2008). https://www.wsj.com/articles/SB121867148093738861
43. Vigen, T.: Spurious correlations (2018). http://www.tylervigen.com/spurious-correlations
44. Jeffreys, H.: Theory of Probability. Oxford University Press, Oxford (1998)
45. Goodman, S.N.: Epidemiology **12**, 295 (2001)
46. Nguyen, H.T., Sriboonchitta, S., Thac, N.N.: Structural Changes and Their Econometric Modeling. Springer (2019, forthcoming)
47. Hájek, A.: Erkenntnis **45**, 209 (1997)
48. Hájek, A.: Uncertainty: Multi-disciplinary Perspectives on Risk. Earthscan (2007)
49. Hájek, A.: Erkenntnis **70**, 211 (2009)
50. Biau, D.J., Jolles, B.M., Porcher, R.: Clin. Orthop. Relat. Res. **468**(3), 885 (2010)
51. Sainani, K.L.: Phys. Med. Rehabil. **4**, 442 (2012)
52. Briggs, W.M.: So, You Think You're Psychic? Lulu, New York (2006)
53. Campbell, S., Franklin, J.: Synthese **138**, 79 (2004)
54. Stove, D.: The Rationality of Induction. Clarendon, Oxford (1986)