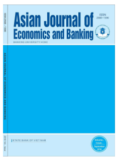
Asian Journal of Economics and Banking

ISSN 2588-1396

http://ajeb.buh.edu.vn/Home



Reality-Based Probability & Statistics: Solving the Evidential Crisis

William M. Briggs [†]

Independent Researcher, New York, NY, United States

Article Info

Received: 29/01/2019
Accepted: 14/02/2019
Available online: In Press

Keywords

Artificial intelligence, Cause, Decision making, Evidence, Falsification, Machine learning, Model selection, Philosophy of probability, Prediction, Proper scores, P-values, Skill scores, Tests of stationarity, Verification.

JEL classifications

Abstract

It is past time to abandon significance testing. In case there is any reluctance to embrace this decision, proofs against the validity of testing to make decisions or to identify cause are given. In their place, models should be cast in their reality-based, predictive form. This form can be used for model selection, observable predictions, or for explaining outcomes. Cause is the ultimate explanation; yet the understanding of cause in modeling is severely lacking. All models should undergo formal verification, where their predictions are tested against competitors and against reality.

1 NATURE OF THE CRISIS

We create probability models either to explain how the uncertainty in some observable changes, or to make probabilistic predictions about observations not yet revealed; see e.g. [7, 93, 86] on explanation versus prediction. The observations need not be in the future, but can be in the past but as yet unknown, at least to the modeler.

These two aspects, explanation and

[†] Corresponding author: William M. Briggs, Independent Researcher, New York, NY, United States. Email address: matt@wmbriggs.com

prediction, are not orthogonal; neither does one imply the other. A model that explains, or seems to explain well, may not produce accurate predictions; for one, the uncertainty in the observable may be too great to allow sharp forecasts. For a fanciful yet illuminating example, suppose God Himself has said that the uncertainty in an observable y is characterized by a truncated normal distribution (at 0) with parameters 10 and 1 million. The observable and units are centuries until the End Of The World. We are as sure as we are of the explanation of y—*if* we call knowledge of parameters and the model an explanation, an important point amplified later. Yet even with this sufficient explanation, our prediction can only be called highly uncertain.

Predictions can be accurate, or at least useful, even in the absence of explanation. I often use the example, discussed below, of spurious correlations: see the website [**?**] for scores of these. The yearly amount of US spending on science, space, and technology correlates 0.998 with the yearly number of Suicides by hanging, strangulation, and suffocation. There is no explanatory tie between these measures, yet because both are increasing (for whatever reason), knowing the value of one would allow reasonable predictions to be made of the other.

We must always be clear what a model's goal is: explanation or prediction. If it is explanation, then while it may seem like an unnecessary statement, it must be said that we do not need a model to tell us what we observed. All we have to do is look. Measurement-error models, incidentally, are not an exception; see e.g. [21]. These models are used when what was observed was not what was wanted; when, for example, we are interested in y but measure $z = y + \tau$, with τ representing the measurement uncertainty. Measurement-error models are in this sense predictive.

For ordinary problems, again we do not need a model if our goal is to state what occurred. If we ran an experiment with two different advertisements and tracked sales income, then a statement like the following becomes certainly true or certainly false, depending on what happened: "Income under ad A had a higher mean than under ad B." That is, it will be the case that the mean was higher under A or B, and to tell all we have to do is look. No model or test is needed, nor any special expertise. We do not have to restrict our attention to the mean: there is no uncertainty in any observable question that can be asked—and answered without ambiguity or uncertainty.

This is not what happens in ordinary statistical investigations. Instead of just looking, models are immediately sought, usually to tell us what happened. This often leds to what I call the Deadly Sin of Reification, where the model becomes more real than reality. In our example, a model would be created on sales income conditioned on or as a function of advertisement (and perhaps other measures, which are not to the point here). In frequentist statistics, a null significance hypothesis test would follow. A Bayesian analysis might focus on a Bayes factor; e.g. [71].

It is here at the start of the modeling process the evidential crisis has as its genesis. The trouble begins because typically the reason for the model has not been stated. Is the model meant to be explanative or predictive? Different goals lead, or should lead, to different decisions, e.g. [78, 79, 78]. The classical modeling process plunges ahead regardless, and the result is massive over-certainty, as will be demonstrated; see also the discussion in Chapter 10 of [11].

The significance test or Bayes factor asks whether the advertisement had any "effect". This is causal language. A cause is an explanation, and a complete one if the full aspects of a cause are known. Did advertisement A *cause* the larger mean income? Those who do testing imply this is so, if the test is passed. For if the test is not passed, it is said the differences in mean income were "due to" or "caused by" chance. Leaving aside for now the question whether chance or randomness can cause anything, if chance was not the cause, because the test was passed, then it is implied the advertisements were the cause. Yet if the ads were a cause, they are of a very strange nature. For it will surely be the case that not every observation of income under one advertisement was higher than every observation under the other, or higher in the same exact amount. The implies inconstancy in the cause. Or, even more likely, it implies an improper understanding of cause and the nature of testing, as we shall see.

If the test is passed, cause is implied, but then it must follow the model would evince good predictive ability, because if a cause truly is known, good predictions (to whatever limits are set by nature) follow. That many models make lousy predictions implies testing is not revealing cause with any consistency. Recall cause was absurd in the spurious correlation example above, even though any statistical test would be passed. Yet useful predictions were still a possibility in the absence of a known cause.

It follows that testing conflates explanation and prediction. Testing also misunderstands the nature of cause, and confuses exactly what explanation is. Is the cause making changes in the observable? Or in the *parameter* of an *ad hoc* model chosen to represent uncertainty in the observable? How can a material cause change the size or magnitude of an unobservable, mathematical object like a parameter? The obvious answer is that it cannot, so that our ordinary understanding of cause in probability models is, at best, lacking. It follows that cause has become too easy to ascribe cause between measures ("x") and observables ("y"), which is a major philosophical failing of testing.

This is the true crisis. Tests based on p-values, or Bayes factors, or on any criteria revolving around parameters of models not only misunderstand cause, and mix up explanation and prediction, they also produce massive over-certainty. This is because it is believed that when a test has been passed, the model has been validated, or proved true in some sense, or if not proved true, then at least proved useful, even when the model has faced no external validation. If a test is passed, the theories that led to the model form in the minds of researchers are then embraced with vigor, and the uncertainty due these theories dissolves. These attitudes have led directly to the reproducibility crisis, which is by now well documented; e.g. [19, 22, 60, 80, 84, 3].

Model usefulness or truth is in no way conditional on or proved by hypothesis tests. Even stronger, usefulness and truth are not coequal. A model may be useful even if it is not known to be true, as is well known. Now *usefulness* is not a probability concept; it is a matter of decision, and decision criteria vary. A model that is useful for one may be of no value to another; e.g. [42]. On top of its other problems, testing conflates decision and usefulness, assuming, because it makes universal statements, that decisions must have the same consequences for all model users.

Testing, then, must be examined in its role in the evidential crisis and whether it is a favorable or unfavorable means of providing evidence. It will be argued that it is entirely unfavorable, and that testing should be abandoned in all its current forms. Its replacements must provide an understanding of what explanation is and restore prediction and, most importantly, verification to their rightful places in modeling. True verification is almost non-existent outside the hard sciences and engineering, fields where it is routinely demanded models *at least* make reasonable, verified predictions. Verification is shockingly lacking in all fields where probability models are the main results. We need to create or restore to probability and statistics the kind of reality-based modeling that is found in those sciences where the reality-principle reigns.

The purposes of this overview article are therefore to briefly outline the arguments against hypothesis testing and parameter-based methods of analysis, present a revived view of causation (explanation) that will in its fullness greatly assist statistical modeling, demonstrate predictive methods as substitutes for testing, and introduce the vital subject of model verification, perhaps the most crucial step. Except for demonstrating the flaws of classical hypothesis testing, which arguments are by now conclusive, the other areas are positively ripe with research opportunities, as will be pointed out.

2 NEVER USE HYPOTHESIS TESTS

The American Statistical Association has announced that, at the least, there are difficulties with p-values, [102]. Yet there is no official consensus on what to do about these difficulties, an unsurprising finding given that the official Statement on p-values was necessarily a bureaucratic exercise. This seeming lack of consensus is why readers may be surprised to learn that every use of a p-value to make a statement for or against the so-called null hypothesis is fallacious or logically invalid. Decisions made using p-values always reflect not probabilistic evidence, but are pure acts of will, as [76] originally criticized. Consequently, p-values should never be used for testing. Since it is p-values which are used to reject or accept ("fail to reject") hypotheses in frequentism, because every use of p-values is logically flawed, it means that there is no logical justification for null hypothesis significance testing, which ought to be abandoned.

2.1 Retire P-values Permanently

It is not just that p-values are used incorrectly, or that their standard level is too high, or that there are good uses of them if one is careful. It is that there exists *no* theoretical basis for their use in making statements about null hypotheses. Many proofs of this are provided in [13] using several arguments that will be unfamiliar or entirely new to readers. Some of these are amplified below.

Yet it is also true that sometimes p-values seem to "work", in the sense that they make, or seem to make, decisions which comport with common sense. When this occurs, it is not because the p-value itself has provided a useful measure but because the modeler himself has. This curious situation occurs because the modeler has likely, relying on outside knowledge, identified at least some causes, or partial causes, of the observable, and because in some cases the p-value is akin to a (loose) proxy for the predictive probabilities to be explained below.

Now some say (e.g. [4]) that the solution to the p-value crisis is to divide the magic number, a number which everybody knows and need not be repeated, by 10. "This simple step would immediately improve the reproducibility of scientific research in many fields," say these authors. Others say (e.g. [45]) that taking the negative log (base 2) of p-values would fix them. But these are only glossy cosmetic tweaks which do not answer the fundamental objections.

There is a large and growing body of critiques of p-values, e.g. [5, 39, 25, 98, 77, 100, 80, 1, ?, 81, 26, 46, 47, 57, ?]. None of these authorities recommend using p-values in any but the most circumscribed way. And several others say not to use them at all, at any time, which is also our recommendation; see [69, 99, 106, 59, 11].

There isn't space here to survey every argument against p-values, or even all the most important ones against them. Readers are urged to consult the references, and especially [13]. That article gives new proofs against the most common justifications for p-values.

2.2 Proofs of P-value Invalidity

Many of the proofs against p-values' validity are structured in the following way: calculation of the p-value does not begin until it is *accepted* or assumed the null is true: p-values only exist when the null is true. This is demanded by frequentist theory. Now if we start by accepting the null is true, logically there is only one way to move from this position and show the null is false. That is if we can show that some contradiction follows from assuming the null is true. In other words, we need a proof by contradiction by using a classic *modus tollens* argument:

- If "null true" is true then a certain proposition Q is true;
- $\neg$ Q (this proposition Q is false in fact);
- Then "null true" is false; i.e. the null is false.

Yet there is no proposition Q in frequentist theory consistent with this kind of proof. Indeed, under frequentist theory, which must be adhered to if p-values have any hope of justification, the only proposition we know is true about the p-value is that assuming the null is true the p-value is uniformly distributed. This proposition (the uniformity of p) is the only Q available. There is no theory in frequentism that makes any other claim on the value of p except that it can equally be any value in $(0, 1)$. And, of course, every calculated p (except in circumstances to be mentioned presently) will be in this interval. Thus what we actually have is:

- If "null true" then Q="$p \sim$ U$(0, 1)$";
- $p \in [0, 1]$ (note the now-sharp bounds).
- Therefore...what?

First notice that we cannot move from observing $p \in (0, 1)$, which is almost always true in practice, to concluding that the null is true (or has been "failed to be rejected"). This would be the fallacy of affirming the consequent. On the other hand, in the cases where $p \in \{0,1\}$, which happens in practical computation when the sample size is small or when the number of parameters is large, then we have found that p is *not* in $(0,1)$, and therefore it follows that the null *is* false by *modus tollens*. But this is an absurd conclusion when $p = 1$. For any $p \in (0,1)$ (not-sharp bounds), it never follows that "null true" is false. There is thus no justification for declaring, believing, or deciding the null is true or false, except in ridiculous scenarios (p identical to 0 or 1).

Importantly, there is no statement in frequentist theory that says if the null is true, the p-value will be small, which would contradict the proof that it is uniformly distributed. And there is no theory which shows what values the p-value will take if the null is false. There is thus no Q which allows a proof by contradiction. Think of it this way: we begin by declaring "The null is true"; therefore, it becomes almost impossible to move from that declaration to concluding it is false.

Other attempts at showing usefulness of the p-value, despite this uncorrectable flaw, follow along lines developed by [58], quoting John Tukey: "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." As Holmes says, "This translates into an inference which suggests that if we observe data X, which is very unlikely if A is true (written $P(\mathrm{X}|\mathrm{A}) < \epsilon$), then A is not plausible."

Now "not plausible" is another way to say "not likely" or "unlikely", which are words used to represent probability, quantified or not. Yet in frequentist theory it is *forbidden* to put probabilities to fixed propositions, like that found in judging model statement A. Models are either true or false (a tautology), and no probability may be affixed to them. P-values in practice are, indeed, used in violation of frequentist theory all the time. Everybody takes wee p-values as indicating evidence that A is likely true, or is true *tout court*. There simply is no other use of p-values. Every use therefore is wrong; or, on the charitable view, we might say frequentists are really closet Bayesians. They certainly *act* like Bayesians in practice.

For mathematical proof, we have that Holmes's statement translates to this:

$$\Pr\left(\mathrm{A}|\mathrm{X}\ \&\ \Pr(\mathrm{X}|\mathrm{A}) = \text{small}\right) = \text{small}. \tag{1}$$

I owe part of this example to Hung Nguyen (personal communication). Let A be the theory "There is a six-sided object that on each activation must show only one of the six states, just one of which is labeled 6." Let X = "2 6s in a row." We can easily calculate $\Pr(\mathrm{X}|\mathrm{A}) = 1/36 < 0.05$. Nobody would reject the "hypothesis" A based on this thin evidence, yet the p-value is smaller than the traditional threshold. And with X = "3 6s in a row", $\Pr(\mathrm{X}|\mathrm{A}) = 1/216 < 0.005$, which is lower than the newer threshold advocated by some. Most importantly, there is no way to calculate (1): we cannot compute the probability of A, first because theory forbids it, and second because there is no way to tie the evidence of the conditions to A. Arguments like this to justify p-values fail.

A is the only theory under consideration, so A is all we have. If we use it, we assume it is true. It does not help to say we have an alternate in the proposition "A or not-A", for that proposition is always true because is a tautology, and it is always true regardless of whether

A is true or false. What people seem to have in mind, then, are more extreme cases. Suppose X = "100 6s in a row", so that $\Pr(X|A) \approx 1.5 \times 10^{-78}$, a very small probability. But here the confusion of specifying the purpose of the model enters. What was the model A's purpose? If it was to explain or allow calculations, it has done that. Other models, and there are an infinite number of them, could better explain the observations, in the sense that these models could better match the old observations. Yet what justification is there for their use? How do we pick among them?

If our interest was to predict the future based on these past observations, that implies A could still be true. Everybody who has ever explained the gambler's fallacy knows this is true. When does the gambler's fallacy become false and an alternate, predictive model based on the suspicion the device might be "rigged" become true? There is *no* way to answer these questions using just the data! Our suspicion of device rigging relates to cause: we think a different cause is in effect than if A were true. Cause, or rather knowledge of cause, must thus come from outside the data (the X). This is proved formally below.

The last proofs against p-value use are not as intuitive, and also relate to knowledge of cause. We saw in Section 1 that spending on science was highly correlated to suicides. Many other spurious correlations will come to mind. We always and rightly reject these, even though formal hypothesis testing (using p-values or other criteria) say we should accept them. What is our justification for going against frequentist theory in these cases? That theory never tells us when testing should be adhered to and when it shouldn't, except to imply it should *always* be used. Many have developed various heuristics to deal with these cases, but none of them are valid within frequentism. The theory says "reject" or "accept (fail to reject)", and that's it. The only hope is that, in the so-called long run (when, as Keynes said, "we shall all be dead"), the decisions we make will be correct at theoretically specified rates. The theory does not the justify arbitrary and frequent departures from testing that most take. That these departures are anyway taken signals the theory is not believed seriously. And if it is not taken seriously, it can be rejected. More about the delicate topic is found in [50, 52, 11].

Now regardless whether the previous argument is accepted, it is clear we are rejecting the spurious correlations because we rightly judge there is no causal connection between the measures, even though the "link" between the measures is verified by wee p-values. Let us expand that argument. In, for example, generalized linear models we *begin* modeling efforts with

$$\mu = g^{-1}(\beta_1 x_1 + \cdots + \beta_p x_p),$$

where μ is a parameter in the distribution said to represent uncertainty in observable y, g is some link function, and the x_i are explanatory measures of some sort, connected through g to μ via the coefficients β_i.

What happened to $x_{p+1}, x_{p+2}, \cdots$? An *infinity* of x have been tacitly excluded without benefit of hypothesis tests. This may seem an absurd point, but it is anything but. We exclude in models for observable y such measures as "The inches of peanut butter in the jar belonging to our third-door-down neighbor" (assuming y is about some unrelated subject) because we recognize, as with the spurious correlations, that there can be no possible causal connection between a relative stranger's peanut butter and an our observable of interest.

Now these rejections mean we are willing to forgo testing at some times. There is nothing in frequentism to say which times hypothesis testing should be rejected and which times it must be used, except, as mentioned, to suggest it always must be used. Two people looking at similar models may therefore come to different conclusions: one claiming a test is necessary to verify his hypothesis, the other rejecting the hypothesis out of hand. Then it is also true many would keep an x_j in a model even if the p-value associated with it is large if there is outside knowledge this x_j is causally related in some way to the observable. Another inconsistency.

So not only do we have proof that all use of p-values are nothing except expressions of will, we have that the testing process itself is rejected or accepted at will. There is thus no theoretical justification for hypothesis testing—in its classical form.

There are many other arguments against p-values that will be more familiar to readers, such as how increasing sample size lowers p-values, and that p-value "significance" is no way related to real-world significance, and so on for a very long time, but these are so well known we do not repeat them, and they are anyway available in the references.

There is however one special, or rather frequent, case in economics and econometrics, where it seems testing is not only demanded, but necessary, and that is in so-called tests of stationarity. A discussion of this problem is held in abeyance until after cause has been reviewed, because it impossible to think about stationarity without understanding cause. The answer can be given here, though: testing is not needed.

We now move to the replacement for hypothesis tests, where we turn the subjectivity found in p-values to our benefit.

3 MODEL SELECTION USING PREDICTIVE STATISTICS

The shift away from formal testing, and parameter-based inference, is called for in for example [44]. We echo those arguments and present an outline of what is called the reality-based or predictive approach. We present here only the barest bones of predictive, reality-based statistics. See the following references for details about predictive probabilities: [24, 37, 38, 61, 62, 66, 14, 12]. The main benefits of this approach are that it is theoretically justified wholly within probability theory, and therefore has no arbitrariness to it, that it unlike hypothesis testing puts questions and answers in terms of observables, and that it better accords with the true uncertainty inherent in modeling. Hypothesis testing exaggerates certainty through p-values, as discussed above.

Since the predictive approach won't be as familiar as hypothesis testing, we spend a bit more time up front before moving to how to apply it to complex models.

3.1 Predictive Probability Models

All probability models fit into the following schema:

$$\Pr(y \in s|\mathrm{M}), \tag{2}$$

where y in the observable of interest (the dimension will be assumed by the context), s a subset of interest, so that "$y \in s$" forms a verifiable proposition.

We can, at least theoretically, measures its truth or falsity. That is, with s specified, once y is observed this proposition will either be true or false; the probability it is true is predicated on M, which can be thought of as a complex proposition. M will contain *every* piece of evidence considered probative to $y \in s$. This evidence includes all those premises which are only tacit or implicit or which are logically implied by accepted premises in M. Say M insists uncertainty in y follows a normal distribution. The normal distribution with parameters μ and σ in this schema is written

$$\begin{aligned}&\Pr(y \in s|\text{normal}(\mu, \sigma))\\ =&\frac{1}{2}\left[\text{erf}\left(\frac{s_2-\mu}{\sigma\sqrt{2}}\right) - \text{erf}\left(\frac{s_1-\mu}{\sigma\sqrt{2}}\right)\right],\end{aligned} \tag{3}$$

where s_2 is the supremum and s_1 the infimum of s when $s \in \mathbb{R}$, and assuming s is continuous. In real decisions, s can of course be any set, continuous or not, relevant to the decision maker. M is the implicit proposition "Uncertainty in y is characterized by a normal distribution with the following parameters." Also implicit in M are the assumptions leading to the numerical approximation to (3) because of course the error function is not analytic ($\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$). Since these approximations vary, the probability of $y \in s$ will also vary, essentially creating *new* or *different* M for every different approximation. This is not a bug, but a feature. It is also a warning that it would be better to explicitly list all premises and assumptions that go into M so that ambiguity can be removed.

It must be understood that each calculation of $\Pr(y \in s|\text{M}_i)$ for every different M_i is correct and true (barring human error). The index is arbitrary and ranges over all M under consideration. It might be that a proposition in a particular M_j is itself known to be false, where it is known to be false conditioned on premises not in M_j: if this knowledge were in M_j it would contradict itself. But this outside knowledge does not make $\Pr(y \in s|\text{M}_j)$ itself false or wrong. That probability is still correct *assuming* M_j is correct. For instance, consider M_1 = "There are 2 black and 1 red balls in this bag and nothing else and one must be drawn out", $\Pr(\text{black drawn}|\text{M}_1) = 2/3$, and this probability is true and correct even if it is discovered later that there are 2 and not 1 red balls (= M_2). All our discovery means is that $\Pr(\text{black drawn}|\text{M}_1) \neq \Pr(\text{black drawn}|\text{M}_2)$. This simple example should be enough to clear up most controversies over prior and model selection, as explained below.

It is worth mentioning that (3) holds no matter what value of y is observed. This is because, unless as the case may be $s \equiv \mathbb{R}$ or $s \equiv \emptyset$,

$$\begin{aligned}&\Pr(y \in s|\text{normal}(\mu, \sigma))\\ &\neq \Pr(y \in s|y, \text{normal}(\mu, \sigma)).\end{aligned}$$

The probability of $y \in s$ conditioned on observing the value of y will be extreme (either 0 or 1), whereas the probability of $y \in s$ not conditioning on knowing the value will not be extreme (i.e. in $(0, 1)$). We must always keep careful track of what is on the right side of the conditioning bar |.

It is usually the case that values for parameters, such as in (3), are not known. They may be given or estimated by some outside method, and these methods of estimation are usually driven by conditioning on observations. In some cases, parameter values are deduced; e.g. such as knowledge that μ in (3) must be 0 in some engineering example. Whatever is the case, each change in estimate, observation, or deduction results in a *new* M. Comparisons between probabilities is thus always a comparison between models. Which model is best can be answered by appeal to the model only in those cases where the model itself has been deduced by premises which are either true themselves, or are accepted as true by those interested in the problem at hand. These models are rare enough, but they do exist; see [11] (Chapter 8) for examples. In other cases, because most models are *ad hoc*, appeal to which model is best must come via exterior methods, such as by the validation process, as demonstrated below.

In general, models are *ad hoc*, being used for the sake of convenience or because of custom. Consider the normal model used above. If the parameters are unknown, guesses may be made, labeled $\hat{\mu}$ and $\hat{\sigma}$. It is in general true then that $\Pr(y \in s|\text{normal}(\mu,\sigma)) \neq \Pr(y \in s|\text{normal}(\hat{\mu},\hat{\sigma}))$, with the equality occurring (for any s) only when $\mu = \hat{\mu}$ and $\sigma = \hat{\sigma}$. Thus we might say $\text{M}_1 = \text{normal } (\mu,\sigma)$ (where the values may be known or supplied) and $\text{M}_2 = \text{normal } (\hat{\mu},\hat{\sigma})$ (where guesses are used). Again, the guesses are usually driven by methods applied to observations. Maximum likelihood estimation is common enough. So that it would be better to write

$$\text{M}_2 = \text{normal } (\text{MLE}(\hat{\mu},\hat{\sigma},\text{D}_n)),$$

where D_n indicates the n previous observations of y (the data). Method of moments is another technique, so that we might write

$$\text{M}_3 = \text{normal } (\text{MOM}(\hat{\mu},\hat{\sigma},\text{D}_n)),$$

And in general $\Pr(y \in s|\text{M}_2) \neq \Pr(y \in s|\text{M}_3)$. Both of these probabilities (for any s) are correct. Whether or not one is better or more useful than another we have yet to answer. Importantly, we then have, for example,

$$\text{M}_4 = \text{normal } (\text{MLE}(\hat{\mu},\hat{\sigma},\text{D}_{n+m})),$$

where M_2 is identical to M_4 except for the addition (or even subtraction) of m observations. Again, in general, $\Pr(y \in s|\text{M}_2) \neq \Pr(y \in s|\text{M}_4)$. However, it is usually accepted that adding more observations provides better estimates, so that it follows M_4 is superior to M_2. However, it might not be that M_4 is more useful than M_2. Usefulness is not a probability or statistical concept: truth is, and we have already proven the probabilities supplied by all these instances are correct. About usefulness, more presently.

The predictive model selection process begins in this example like this:

$$\Pr(y \in s|\text{M}_4(\text{D}_{n+m})) = p + \delta, \qquad (4)$$
$$\Pr(y \in s|\text{M}_2(\text{D}_n)) = p \qquad (5)$$

Following [64], we say the additional observations m are relevant if $\delta \neq 0$, and irrelevant if $\delta = 0$. There are obvious restrictions on the value of δ, i.e. $\delta \in [-1,1]$ (the bounds may or may not be sharp depending on the problem). If adding new observations does not change the probability of $y \in s$, then adding these points has provided no additional usefulness. Relevance, as is clear, depends on s as well as M. Adding new observations may be relevant for some s (say, in the tails) but not for others (say, near the median); i.e. $\delta = \delta(s)$. As the s of interest themselves depend on the decisions made by the user of the probability model, relevance cannot be totally a probability concept, but must contains aspects of decision. The form (4) would be useful in developing sample size calculations, which are anticipated to be similar to Bayesian sample size methods, e.g. [70]. This is an open area of research.

The size of the critical value δ is also decision dependent. No universal value exists, or should exist, as with p-values. The critical value may not be constant but can depend on s; e.g. large relative changes in rarely encountered y may not be judged important. Individual problems must "reset" the value of $\delta(s)$ each time.

As useful as the form (4) will be to planning experiments, it is of more interest to use it with traditional model selection. Here then is a more general form:

$$\Pr(y \in s|\mathrm{M}_1) = p + \delta, \qquad (6)$$

$$\Pr(y \in s|\mathrm{M}_2) = p. \qquad (7)$$

This is, of course, identical in form to (4), which shows the generality of the predictive method. M_1 may be any model that is not logically deducible from M_2; if M_2 is deducible from M_1, then $\delta \equiv 0$. What's perhaps not obvious is that (6) can be used both before and after data is taken: before data, it is akin to (4); after, we have genuine predictive probabilities.

A Bayesian approach is assumed here, though a predictive approach under frequentism can also be attempted (but not recommended). We have the schematic equation

$$\Pr(y \in s|\mathrm{M}_\theta) = \int_\Theta \Pr(y \in s|\theta\mathrm{M}_\theta)\Pr(\theta|\mathrm{M}_\theta)d\theta. \qquad (8)$$

Here M is a parameterized probability model, possibly containing observations; and of course it also contains *all* (tacit, explicit and implicit) premises used to justify the model. If (8) is calculated before taking data, then $\Pr(\theta|\mathrm{M}_\theta)$ is the prior distribution and $\Pr(y \in s|\theta\mathrm{M}_\theta)$ the likelihood. In this case, (8) is the prior predictive distribution (allowing s to vary, of course). If the calculations are performed after data has been taken, M_θ is taken to represent the model plus data, and thus $\Pr(\theta|\mathrm{M}_\theta)$ becomes the posterior, and (8) becomes the posterior predictive distribution.

All models have a posterior predictive form, though most models are not "pushed through" to this final form. See [6] for derivation of predictive posteriors for a number of common models, computed with so-called reference priors. Now here it worth saying that (8) with an M_θ which assumes as part of its list of premises a certain prior is *not* the same as another posterior predictive distribution with the same model form but with a different prior. Some are troubled by this. They should not be. Since all probability is conditional on the assumptions made, changing the assumptions changes the probability. Again, this is not a bug, but a feature. After all, if we change the (almost always) *ad hoc* model form we also change the probability, and this is never bothersome. We have already seen that adding new data points also changes the probability, and nobody ever balks at that, either. It follows below that everything said about comparing models with respect to measures x_i applies to comparing models with different priors.

The immediate and obvious benefit of (8) is that direct, verifiable, reality-based probabilistic predictions are made. Since y is observable, we have a complete mechanism in place for model specification and model verification, as we shall soon see.

It is more usual to write (8) in a form which indicates both past data and potentially helpful measures x. Thus

$$\Pr(y \in s|\mathrm{XD}_n\mathrm{M}_\theta) = \int_\Theta \Pr(y \in s|\theta\mathrm{XD}_n\mathrm{M}_\theta)\Pr(\theta|\mathrm{XD}_n\mathrm{M}_\theta)d\theta, \tag{9}$$

where $\mathrm{X} = (x_1, x_2, \cdots, x_p)$ represents new or assumed values of measures x, and $\mathrm{D}_n = (y_n, \mathrm{X}_n)$ are the previous n observations (again, the dimension of y etc. will be obvious in context). Usually or by assumption $\Pr(\theta|\mathrm{XD}_n\mathrm{M}_\theta) = \Pr(\theta|\mathrm{D}_n\mathrm{M}_\theta)$.

For model selection, which assumes, but need not, the same prior and observations, but which must choose between x, we finally have

$$\Pr(y \in s|\mathrm{XD}_n\mathrm{M}_1) = p + \delta, \tag{10}$$
$$\Pr(y \in s|\mathrm{XD}_n\mathrm{M}_2) = p. \tag{11}$$

Everything said above about δ applies here, too. Briefly, addition of subtraction of any number of x from one model to the other will move δ away from 0. The distance it moves is a measure of importance of the change, but only with respect to a decision made by the user of the model. At some s, a $|\delta| > 0.01$ may be crucial to one decision maker but useless to a second, who may require in his application a $|\delta| > 0.20$ before acting. This assumes the same models and same observations for both decision makers—and same s. The predictive approach has thus removed a fundamental flaw of hypothesis testing, which set one number for significance for all problems and decisions everywhere. There was no simple way in hypothesis testing to use a p-value in decisions with different costs and losses; yet predictive probability is suited for just this purpose. And since these are predictive probabilities, the full uncertainty of the model and data are accounted for, which fixes a second fundamental problem of hypothesis tests, which revolved around unobservable parameters. Recall that we can be as certain as possible of parameter values but still wholly uncertain in the value of the observable. This point is almost nowhere appreciated, but it becomes glaringly obvious when data is analyzed.

3.2 Comparing other Model Selection Measures

Now (10) shares certain similarities with Bayes factors. These may be written in this context as

$$\begin{aligned}\mathrm{BF} &= \frac{\Pr(\mathrm{D}_n|\mathrm{M}_1, \mathrm{E})}{\Pr(\mathrm{D}_n|\mathrm{M}_2, \mathrm{E})} \\ &= \frac{\Pr(\mathrm{M}_1|\mathrm{D}_n, \mathrm{E})}{\Pr(\mathrm{M}_2|\mathrm{D}_n, \mathrm{E})}\frac{\Pr(\mathrm{M}_2|\mathrm{E})}{\Pr(\mathrm{M}_1|\mathrm{E})}.\end{aligned} \tag{12}$$

This form implies there must be outside or extra-model evidence E from which we could assess $\Pr(\mathrm{M}_1|\mathrm{E})$ and $\Pr(\mathrm{M}_2|\mathrm{E})$. If E suggests these are the two and the only two models under consideration, then it follows $\Pr(\mathrm{M}_1|\mathrm{E}) = 1 - \Pr(\mathrm{M}_2|\mathrm{E})$. It is difficult to see what this E might be, except for the simple case where the obvious deduction $\Pr(\mathrm{M}_1|\mathrm{E}) = \Pr(\mathrm{M}_1|\mathrm{E}) = 1/2$ is made. But then that would seem to hold for any two models, regardless the number of changes made between models; i.e. we make the same deduction whether one parameter changes between models or whether there are $p > 1$ changes. Of course, such problem-dependent E might very well exist, and in practice they always do, as the infinity of hypotheses example above proved. In those cases where it is explicit it should be used if the decision is to pick one model over another.

The BF also violates the predictive nature of the problem. It asks us to choose a model as the once and final model, which is certainly a sometime goal, but the measure here is not predictive in terms of the observable. If our interest is predictive, and with any finite n, there will still be positive probability for either model being true. We need not choose in these cases, but can form a probability weighted average of $y \in s$ assuming both models might be true. This is another area insufficiently investigated: why throw away a model that may be true when what one really wants are good predictions?

We can in any case see that the BF exaggerates differences as hypothesis tests did, though not in the same manner. For one, the explicit setting of s is removed in the BF, whereas in the predictive $\delta = \delta(s)$ it is assumed any model may be judged useful for some s and not for others. The BF sort of performs an average over all s. A large (or small) BF value may be seen, but because we're comparing ratios of probabilities the measure is susceptible to swings in the denominator toward 0. Of course, the predictive (10) can be written in ratio form, i.e. $\frac{p+\delta}{p}$, and this may in some cases be helpful in showing how predictive measures are related to Bayes factors and other information criteria, such as the Bayesian Information Criterion, AIC, minimum message length, and so on. All these are open research questions. However, only the predictive method puts the results in a form that are directly usable and requires no additional interpretation by model users. This latter benefit is enormous. How much easier is to to say to a decision maker, "Given our model and data, the probability for $y \in s$ is p" than to say "Given our data and that we accept our model is false, then we expect to see values of this *ad hoc* statistic $p \times 100\%$ of the time, if we can repeat the experiment that generated our data an infinite number of times"? The question answers itself.

How to pick the s? They should *always* be related to the decisions to be made, and so s will vary for different decision makers. However, it should be the case that for some common problems natural s arise. This too is an open area of research.

3.3 Example

Here is a small example, chosen for its ease in understanding. The Boston Housing dataset is comprised of 506 observations of Census tract median house prices (in \$1,000s), along with 13 potential explanatory measures, the most interesting of which is nox, the atmospheric nitric oxides concentration (parts per 10 million), [54]. The idea was that high nox concentrations would be associated with lower prices, where "associated" was used as a causal word. To keep the example simple yet informative, we only use some of the measures: crim, per capita crime rate by town; chas, Charles River border indicator; rm, average number of rooms per dwelling; age, proportion of owner-occupied units built prior to 1940; dis, weighted distances to five Boston employment centres; tax, full-value property-tax rate; and b, a function of the proportion of blacks by town. The dataset is available in the R package `mlbench`. All examples in this paper use R version 3.4.4, and the Bayesian computation package `rstanarm` version 2.17.4 with default priors.[a].

The original authors used regression of price on the given measures. The ordinary ANOVA table is given in Table 1.

[a] Code for all examples is available at http://wmbriggs.com/post/26313/

Table 1 The ANOVA table for the linear regression of median house prices (in $1,000s) on a group of explanatory measures. All variables would pass ordinary hypothesis tests.

```
Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.035298   4.193677  -2.631  0.00877
crim         -0.110783   0.036748  -3.015  0.00270
chas1         4.043991   1.010563   4.002 7.25e-05
nox         -11.068707   4.145901  -2.670  0.00784
rm            7.484931   0.381730  19.608  < 2e-16
age          -0.068859   0.014473  -4.758 2.57e-06
dis          -1.146715   0.207274  -5.532 5.12e-08
tax          -0.006627   0.002289  -2.895  0.00395
b             0.012806   0.003142   4.076 5.33e-05
```

The posterior distribution of the parameters of the regression mimic the evidence in the ANOVA table. These aren't shown because the interest is not on parameters, but observables. Most researchers would be thrilled by the array of wee p-values, figuring the model must be on to something. We shall see this hope is not realized.

What does the predictive analysis show? That's a complicated question because there is no single, universal answer like there is in hypothesis-testing, parameter-centric modeling. This makes the method more burdensome to implement, but since the predictive method can answer any question about observables put to it, it's generality and potential are enormous.

We cannot begin without asking questions about observables. This is implicit in the classical regression, too, only the questions there have nothing directly to do with observables, and so nobody really cares about them. Here is a question which I thought interesting. It may be of no interest to any other decision maker, all of whom may ask different questions and come to different judgments of the model.

The third quartile observed housing price was $35,000. What is the predicted probability prices would be higher than that given different levels of nox *for data not yet observed*? The answer for old data can be had by just looking. In order to answer that, we also have to specify values for crim, chas, and *all* the other measures we chose to put into the model. I picked median observed values for all. The `stan_glm` method was used to form a regression of the same mathematical form as the classic procedure, and the `posterior_predict` method from that package was used to form posterior predictive distributions, i.e. eq. (9). These are solved using resampling methods; for ease of use and explanation the default values on all methods were used.

Fig. 1 shows the relevance plot for the models with and without nox. This is the predictive probability of housing prices greater than $35,000 with all measures are set at their median value, and with nox varying from its minimum to maximum observed values. The lines are not smooth because they are the result of a resampling process; larger resamples would produce smoother fig-

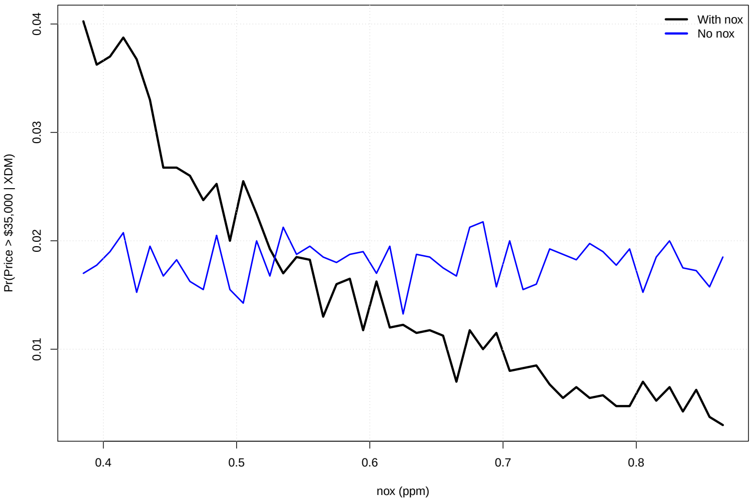


Fig. 1 Relevance plot, or the predictive probability of housing prices greater than $35,000, using models with nox (black line) and without (blue). All other measures are set at their median value.

ures; however, these are adequate for our purposes.

The predictive probability of high housing prices goes from about 4% with the lowest levels of nox, to something near 0% at the maximum nox values. The predictive probability in the model without nox is about 1.8% on average. The original p-value for nox was 0.008, which all would take as evidence of strong effect. Yet for this question the probability changes are quite small. Are these differences (a ± 2% swing) in probability enough to make a difference to a decision maker? There is no single answer to that question. It depends on the decision maker. And there would still not be an answer until it was certain the other measures were making a difference. Now experience with the predictive method shows that often a measure will be predictively useful, but which also gives a large p-value; but we also see cases where the measure shows a wee p-value but does not provide any real use in predictive situations. Every measure has to be checked (and this is easily automated). We don't do this here because it would take us too far afield.

What might not be clear but needs to be is that we can make predictions for any combination of X, for any function of Y. Usefulness of X (any of its constituent parts) is decided with respect to these functions of Y, which in turn are demanded by the decisions to be made. Usefulness is in-sample usefulness, with the real test of any model being verification, which is discussed below. Anticipating that, we have a most useful plot, which is the probability prediction of Y for every old observed X, supposing that old X were new. This is shown in Fig 2.

Price (s) in on the x-axis, and the probability of future prices less than s, given the old data and M, are on the y-axis. A dotted red line at $0 is shown. Now we know based on external knowledge to M that it is impossible prices can be less than $0. Yet the model far too often gives positive probabilities for impossible prices. The

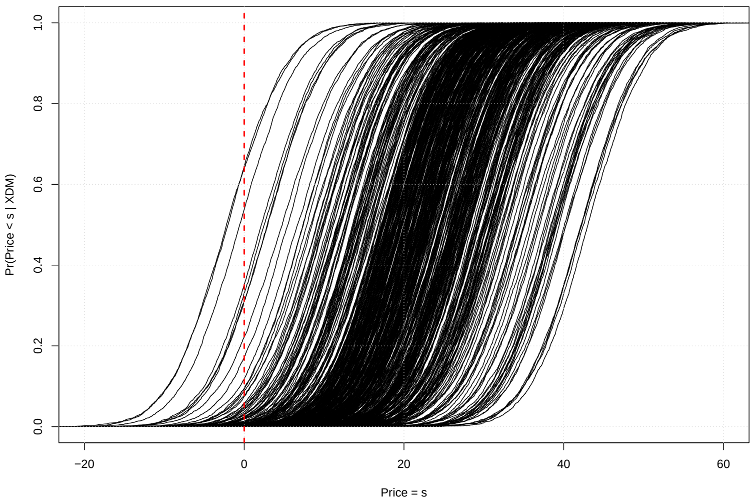


Fig. 2 The probability prediction of housing prices for every old observed X, supposing that old X were new. The vertical red dashed line indicates prices we know to be impossible based on knowledge exterior to M.

worst prediction is about a 65% chance for prices less than $0. I call this phenomenon *probability leakage*, [9]. Naturally, once this is recognized, M should be amended. Yet it never would be recognized using hypothesis testing or parameter estimation: the flaw is only revealed in the predictive form. An ordinary regression is inadequate here. I do not here pursue other models, which are here not the point. What should be fascinating is the conjecture that many, many models in economics, if they were looked at in their predictive sense, would show leakage, and when they do it is another proof the ordinary ways of examining model performance generate over-certainty. For as exciting as the wee p-values were in the ANOVA table above, the excitement was lessened when we looked for practical differences in knowledge of nox. And that excitement turned to disappointment when it was learned the model had too much leakage to be useful in a wide variety of situations.

We have only sketched the many opportunities in predictive methods research. How the predictive choices fit in with more traditional informational criteria, such as given in [65], is largely unknown. It seems clear the predictive approach avoids classic "paradoxes", however, such as for instance given in [67], since the focus in prediction is always on observables and their probabilities.

It should also be clear that nox, useful of not, could not cause a change in housing prices. In order to be a cause, nox would somehow have to seep into realtors' offices and push list prices up or down. This is not a joke, but a necessary condition for nox being an efficient cause. Another possibility is that buyers' or sellers' *perception* of nox caused them to raise or lower prices. This might happen, but it's scarcely likely: how many home owners can even identify what nox is? So it might be that nox causes other things that, in turn or eventually, cause prices to change. Or it could be that nox has nothing to do with the cause of price changes, and that its association with price is a coincidence or the result or "confounders." There is no way to tell by just examining the data. This judgment is strict, and is proven in the next Section.

4 Y CAUSE?

This section is inherently and necessarily more philosophical than the others. It addresses a topic scientists, the closer their work gets to statistics, are more accustomed to treating cavalierly and, it will be seen, sloppily. As such, some of the material will be entirely new to readers. The subject matter is vast, crucial, and difficult. Even though this is the longest section, it contains only the barest introduction, with highlights of the biggest abuses in causal ascription common in modeling. One of the purposes of models we admitted was explanation, and cause is an explanation. There are only two possibilities of cause in any model: cause of change in the observable y or cause of change in the unobservable, non-material parameters in the model used to characterize uncertainty in y. So that when we speak of explanation, we always speak of cause. Cause is *the* explanation of any observable, either directly or through a parameter. Since models speak of cause, or purport to, we need to understand exactly where knowledge of cause arises: in the data itself through the model, or in our minds. We must understand just what cause means, and we must know how, when, and if we really can identify cause. It will be seen that, once again, classic data analysis procedures lead to over-certainty.

4.1 Be Cause

There is great confusion about the role cause and knowledge of cause plays in statistical and econometric models, e.g. [43, 97], which are all species of probability models. I include in this designation all artificial intelligence and so-called machine learning algorithms whose outputs while they may be point predictions are not meant to be taken literally or with absolute certainty, implying uncertainty is warranted in their predictions; hence they are non-standard forms of probability models. We may say all such algorithms, from the statistical to the purely computational, are uncertainty models. Any model which produces anything but certain and known-to-be-certain predictions is an uncertainty model in the sense that probability, quantified or not, must be used to grasp the output.

Can probability or uncertainty models discover cause? Always or never, or only under certain circumstances? What does cause mean? These are all large topics, impossible to cover completely in a small review article. So here we have the limited goal of exploring the confusing and varying nature of cause in probability models in common use, and in contrasting modern, Humean, Popperian, and Descartean notions of cause with the older but resurgent Aristotelian ideas of cause that, it will be argued, should be embraced by researchers for the many benefits it contains.

For good or bad, and mostly a whole lot of bad, causal language is often used to convey results of probability models; see [49]. Results which are merely probable are far too often taken as certain. [36] relates a history of attempts at assigning cause in linear models, beginning with Yule in 1899. These attempts have largely been similar: they begin by specifying a parameterized linear model. One or more of the parameters are then taken to be effects of causes. Parameter estimates are often called "effect size", though the causes thought to generate these effects are not well specified. Models are often written in causal-like form (to be described below), or cause is conceived by drawing figurative lines or "paths" between certain parameters. The conclusion cause exists or does not exist depends on the signs of these parameters' estimates. [83] has a well known book which purports to design strategies at automatically identifying causes. [27] investigate design of experiments which are said to lead to causal identification. It will be argued below that these are vain hopes, as algorithms cannot understand cause.

What's hidden in these works is the tacit assumption, shared by frequentist, Bayesian, and computer modeling efforts, that cause and effect can always be quantified, or quantified with at enough precision to allow cause to be gleaned from models. This unproved and really quite astonishingly bold assumption doubtless flows the notion that to be scientific means to be measurable; see [28]. It does not follow, however, that everything *can* be measured. And indeed, since Heisenberg and Bell, [92], we have known that some things, such as the causes for certain quantum mechanical events, cannot be measured, even though some of the effects might and are. We therefore know of cases where knowledge of cause is impossible. Let us see whether these cases multiply.

Now *cause* is a philosophical, or metaphysical concept. Many scientists tend to view philosophy with a skeptical eye; see e.g. [94]. But even saying one has no philosophy is a philosophy, and the understanding of cause or even the meaning of any probability requires a philosophy, so it is well to study what philosophers have said on the subject of cause, and see how that relates to probability models.

The philosophy we adopt here is probabilistic realism, which flows from the position of moderate realism; see [82, 32]. It is the belief that the real world exists and is, in part, knowable; it is the belief that material things exist and have form, and that form can exist in things or in minds (real triangles exist, and we know the form of triangle in the absence of real triangles). In mathematics, this is called the Aristotelian Realist philosophy, see [35] for a recent work. This is in contrast to the more common Platonic realism, which holds numbers and the like exist as forms in some untouchable realm, and nominalism, which holds no forms exists, only opinion does; see [89] for a history. The moderate realist position is another reason we call the approach in this paper reality-based probability. Probability does not exist as a thing, as a Platonist would hold, but as an idea in the mind. Probability is thus purely epistemological. This is not proved here, but [11] is an essential reference.

What is cause? [56] opens his article on probabilistic causation by quoting Hume's *An Enquiry Concerning Human Understanding*: "We may define a cause to be *an object, followed by another, and where all the objects similar to the first, are followed by objects similar to the second.*" This seemingly straightforward theory—for it is a theory—led Hume through the words *followed by another* ultimately to skepticism, and to his declaration that cause and event were "loose and separate". Since many follow Hume, our knowledge of cause is often said to be suspect. Cause and effect are seen as loose and separate because that *followed by* cut the link of cause from effect. The skepticism about cause in turn led to skepticism about induction, which is wholly unfortunate since our surest knowledge, such as that about mathematical axioms, can only come from inductive kinds of reasonings; there are at least five kinds of induction. The book by [48] is an essential reference. Skepticism about induction led, via a circuitous route through Popper and the logical positivists, to hypothesis testing, and all it associated difficulties; see the histories in [20, 8]. However, telling that story would take us too far afield; interested readers can consult [11] (Chapter 4), [95, 104] about induction, and Briggs (Chapter 5) and [13] about induction and its relations to hypothesis testing.

In spite of all this skepticism, which pervades many modern philosophical accounts of causation and induction, scientists retained notions of cause (and induction). After all, if science was not about discovering cause, what was it about? Yet if scientists retained confidence in cause, they also embraced Hume's separation, which led to curious interpretations of cause. Falsification, a notion of Popper's, even though it is largely discredited in philosophical circles ([94, 96]), is still warmly embraced by scientists, even to the extent that models that are said not to be falsifiable are not scientific.

It's easy to see why falsification is not especially useful, however. If for example a model on a single numerical observable says, as many probability models do say, an observable can take any value on the real line with a non-zero probability, no matter how small that probability (think of a normal model), then the model may *never* be falsified on any observation. Falsification can only occur where a model says, or it is implied, that a certain observable is impossible—not just unlikely, but *impossible*—and we subsequently see that observable. Yet even in physical models when this happens in practice, which is rare, the actual falsification is still not necessarily accepted because the model's predictions are accompanied by a certain amount of "fuzz" around its predictions, [23]; that is, the predictions are not believed to be perfectly certain. With falsification, as with testing, many confuse probability with decision.

Another tacit premise in modern philosophies is that cause is limited to efficient causality: described loosely as that which makes things happen. This limitation followed from the rejection of classical, Aristotelian notions of cause, which partitioned cause into four parts: (1) the formal or form of a thing, (2) the material or stuff which is causing or being affected, (3) efficient cause, and (4) final cause, the reason for the cause, sometimes called the cause of (the other) causes. See [31] for a general overview. For example, consider an ashtry: the formal cause is the shape of the ashtray, the material cause is the glass comprising it, the efficient cause the manufacturing process used to create it, and the final cause the purpose, which is to hold ashes. The reader would benefit from thinking how the fullness of cause explains observables of interest to him.

Final causation is teleological, and teleology is looked on askance by many scientists and philosophers; biologists in particular are skittish about the concept, perhaps fearing where embracing teleology might lead; e.g. [68]. Whatever its difficulties in biology, teleology is nevertheless a crucial element in assessing causes of willed actions, which are by definition directed, and which of course include all economic actions, e.g. [105]. Far from being rarefied, all these distinctions are of the utmost importance, because we have to know just which part of a cause is associated with what parameter in a probability model, as discussed momentarily. This is an area of huge research opportunity. For instance, is the parameter representing the efficient cause, or the material? Or the formal or final. Because cause is believed in modern research to be one thing, over-certainty again arises.

The modern notion of cause, as stated above, is that a cause, a power of some kind, acts, and then at some future point an effect occurs. The distance in time of this separation of cause and effect is never exactly specified, which should raise a suspicion that something from the description is missing. It has led to the confusion that time series (in the formal mathematical sense) might be causal. In any case, it is acknowledged that efficient causes operate on material things which possess something like a form, though the form of the material thing is also allowed to be indistinct, meaning that it is accepted that the efficient cause may change the form of a thing, which still nevertheless still remains the same thing, at least in its constituent parts. The inconsistency is that the form of the thing describes its nature; when a form is lost or changed to a new form, the old form is gone.

Contrast this confusing description with the classical definition of substantial form; essential references are [34, 82, 32]. Substances, i.e. things, are composed of material plus form. A piece of glass can take the form of a window or an ashtray. Substances, which are actual, also have or possess potentiality; to be here rather than there, to be this color rather than that, to not exist, and so forth. Potentiality is thus part of reality. Potentiality becomes actuality when the substance is caused to change. The principle of (efficient) causality (accepted by most philosophers) states that the reduction of potency to actuality requires something actual. A change in potentiality to actuality is either in essence, when something comes to be or passes out of existence, or in accident, when something about a thing changes, such as position or in some other observable component, but where the substance retains its essence (a piece of glass moved is still a piece of glass, unless it is broken or melted, then its essence has changed). A probability model quantifies potentiality in this sense. This is an active area of research in quantum mechanics and probability; see [63, 90].

Cause is ontological: it changes a thing's being or accidents, which are defined as those properties a substance has which are not crucial for its essence, i.e. what it is. A red house painted white is still a house. It is everywhere assumed the principle of sufficient reason holds, which states that every thing that exists has a reason or explanation for its existence. In other words, events do not happen for "no reason"; the idea that things happen for "no reason" is, after all, an extremely strong claim. Now we can be assured that there are sufficient reason for a thing's existence, but this in no way is an assertion that anybody can know what those reasons always are. And indeed we cannot always know a thing's cause, as in quantum mechanics.

Knowledge of cause is epistemological. As with anything else, knowledge can be complete, of truth or falsity, or incomplete, and of a probabilistic nature. If cause is purely efficient, then uncertainty of cause is only of efficient causes; indeed, as we'll see below this is the way most models are interpreted. There is an unfortunate ambiguity in English with the verb *to determine*, which can mean *to cause* or *to provide sufficient reason*. This must be kept in mind if cause is multifaceted, because a model make speak of any of four causes.

4.2 Cause in Models

Using a slightly different notation than above, most probability models follow the schema $y \sim f(x, \theta, \mathrm{M})$, where y is the observable of interest, and M the premises or evidence used to imply the form of f, i.e. the model. The function f is typically a probability distribution, usually continuous, i.e. $y \in \mathbb{R}$. Continuity of the observable is an assumption, which is of course impossible to verify, because no method of measurement exists that could ever verify whether y is actually infinitely graduated (in whatever number of dimensions): all possible measurements we can make are discrete and finite in scope. This may seem like a small point, but since we are interested in the cause of y, we have to consider what kind of cause can itself be infinitely graduated, which must be the case if y can take infinitely many values—where the size of the infinity has yet to be discovered. Is $y \in \mathbb{N}$ or is $y \in \mathbb{R}$ or is y in some higher infinity still? It should make us gasp to think of how a cause can operate on the infinite integers, let alone the "real" numbers, where measure theory usually stops. But if we think we can identify cause on $\mathbb{R}$, why not believe we can identify it on sets with cardinality larger than $\aleph_1$ (the cardinality of $\mathbb{R}$)? These are mind-boggling questions, but it is now perhaps clear the difference between potentiality and actuality is crucial. We can have a potentially infinite number of states of an observable, but only a finite number of actual states. Or we can have a potentially infinite number of observables, but only a finite number of actual observables: if any observable was infinite in actuality, that's all we would see out of our windows. Needless to say, how cause fits in with standard measure theory is an open area of research.

The model f (given by M) of the uncertainty of y is also typically conditioned on other measures x, which are usually the real point of investigation. These measures x are again themselves also usually related with parameters θ, themselves also thought continuous. As said, there are a host of assumptions, many usually implicit, or implicit and forgotten, in M, which are those premises which justify the model and explain its terms. This is so even if, as if far from unusual, M is the excuse for an *ad hoc* model. Most models in actual use are *ad hoc*, meaning the they were deduced from first principles.

The stock example of a statistical model is regression, though what is said below applies to any parameterized model with (what are called) covariates or variables. Regression begins in assuming the uncertainty in the observable y is characterized by a parameterized distribution, usually the normal, though this is expanded in generalized linear regression. The first parameter μ of the normal is then assumed to follow this equation:

$$\mu \sim \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p. \quad (13)$$

The description of the uncertainty in this form is careful to distinguish the probabilistic nature of the model. Equation (13) says nothing about the causes of y. It is entirely a representation of how a parameter representing a model of the uncertainty in y changes with respect to changes in certain other measures, which may or may not have anything to do with the causes of y. The model is correlational, not causal. The parameters are there as mathematical "helpers", and are not thought to exist physically. They are merely weights for the uncertainty. And they can be got rid of, so to speak, in fully correlational predictive models; i.e. where the parameters are integrated out. For example, Bayesian posterior predictive distributions; see above and [6]. In these cases, as above, we (should) directly calculate

$$\Pr(y \in s|\mathrm{XD}_n\mathrm{M}) \qquad (14)$$

We remind the reader that M contains *all* premises which led to the form (13), including whatever information is given on the priors of the parameters and so forth. Again, no causation is implied, there are no parameters left, and everything, for scientific models, is measurable. Equation (14) shows only how the (conditional on D and M) probability of $y \in s$ changes as each x_i does. Of course, the equation will still give answers even if no x_i has any causal connection to y in any way. Any x inserted in (14) will give an answer for the (conditional) probability of $y \in s$, even when the connection between any x and y is entirely spurious. The hope in the case of spurious x_i is that the x_i will show low or no correlation with y and that

$$\Pr(y \in s|\mathrm{XD}_n\mathrm{M}) \approx \Pr(y \in s|\mathrm{X}_{-i}\mathrm{D}_n\mathrm{M})\ \forall s \qquad (15)$$

Indeed, if the equality is strict, then x_i is said to be as above, using the wording on [64], *irrelevant* for the understanding of the uncertainty of y. If the equality is violated, x_i is relevant. Relevance does not imply importance or that a cause between x and y has been demonstrated.

Another way of writing regression, which is mathematically equivalent but philosophically different, is this:

$$y = \theta_0+\theta_1x_1+\theta_2x_2+\cdots+\theta_px_p+\epsilon. \qquad (16)$$

This form is taken as directly or vaguely causal, depending on the temperament of the reader. This is a model *of* y itself and not the *uncertainty* of y as in (13). The last term ϵ is said, by some, *to be* normal (or some other distribution). Now *to be* is, or can be, an ontological claim. In order for ϵ to ontologically *be* normal, probability has to be real, a tangible thing, like mass or electron charge is. The ϵ is sometimes called an "error term", as if y could be predicted perfectly if it were not for the introduction of this somewhat mysterious "exogenous" cause or causes. The x_i are interior to the model, meaning they pertain to y in some, usually undefined, causal way.

The indirect causal language used for models in this form is vague in practice. This form of the model says that the x_i are causes of y *in combination* with the θ_i, which may be moderators of causes or the effects of causes due to the x. What part of cause—formal, material, efficient, final—is never said, though efficient causation is frequently implied. Often the θ_i are called *effect sizes*, which is causal language, meaning that unit increases in x_i *cause* y to change by θ_i, which is the measurable effect. But if this is true, then this cause must always operate in just the same way, unless blocked. More on this implications of this in a moment.

Many physics or deterministic models are written in a way similar to (16), but in those cases cause is better understood. In probability models, the ϵ is assumed to be filled with causes, though small ones, like flea bites. The causes x_i are linear in effect, or near enough so, with negligible forces outside linearity; if there are such forces they are deposited into ϵ or into functions of the x_i. The error term is causal but in such a way the effects are constrained not to be linear individually, but of a certain distributional shape. Causes outside this shape may exist, as might causes outside the x_i, but they are not captured by the model. Perfect predictability is rarely claimed, which is an admission all causes have not been identified.

Now all this is very confusing; worse, the amorphic causal language is bolstered when null hypothesis significance testing is used to decide which x_i "belong" to (16) and which do not. A so-called null hypothesis says typically that some $\theta_j = 0$, which is taken as meaning that x_j is not causal, which is an odd way to put it, as if θ_j itself is the true cause. The "alternate hypothesis" is that $\theta_j \neq 0$, which is everywhere taken as meaning x_j is causal, or that it is something like a cause, such as a "link". When a p-value associated with θ_j is greater than the magic number, x_j is sometimes thought not to be a cause; it is dismissed as if it has been proven not to be cause, or that it might be a cause but of great weakness. But then sometimes x_j, even with a large p-value is still believed to be a cause though that there was not yet enough evidence to "prove" x_j was a cause—or a "link".

Link is nowhere defined in the literature. It may be that x_j is a direct cause or some indirect cause; it is never quite specified, nor is the part of cause identified. It may be that a third, or a fourth, or some number down the chain, measure changes in response to y, which in turn back up the chain causes x_j to change. Now whatever the true cause is, it either happens to y or it happens to μ. But a cause can only happen to a parameter if that parameter materially exists. There does not exist, as far as I can discover, any explanation of how any causal power may change the value of a parameter: by what mechanism? Instead, it appears to be believed that observables y themselves *have* distributions. We see this in language that y are "drawn" from distributions, which must therefore exist in some Platonic realm (e.g. "y *is* normal"). Frequentists are, of course, committed to believing this, though most have not thought out the implications of this assertion; again, see [51, 52] for criticisms. Finally, if observables really do have distributions in an ontological sense, then it must be that causes operating on observables really do cause changes in their immaterial parameters—which always then must exist as real things in subsequences of sequences going to the limit (the definition of frequentism). Frequentism is Platonism. Even if any of this is true, it should at least be clear to the reader that all is not well thought out. The alternative is to forgo frequentism and move to more logically consistent theories of probability where cause is constrained to observables and tangible measures; i.e. a theory that is reality-based.

Regression is the paradigmatic parameterized probability model. There are other kinds of models, though, which are not parameterized but which are still discussed as if they are causal but yet have probabilistic interpretations. Neural nets and the like fall into this class of probability model. Classification and regression trees, random forests and other various so-called machine learning algorithms do, too; ;see [7]. Without getting into too much detail, artificial intelligence and generic machine learning models produce strings of decision rules like this:

$$\text{If } (x_{12} < a \text{ and } b \leq x_7 < c) \text{ or } (x_{25} = \text{Yellow ...}) \text{ then } y \in s. \qquad (17)$$

These pronouncements can and usually are made probabilistic by grouping rules and permutations of rules together, such that when the algorithm is given a set of x, it produces a probability of $y \in s$ (or something which is treated like a probability). Since these algorithms work with actual observations, they usually maintain a discrete and finite stance. This is in their favor since no *ad hoc* assumptions about formal probability models are made. When any formal probability is overlaid upon them, they can be treated like other parameterized probability models.

Various indirect measures of importance for each x_i are made in these models, such as examining how accuracy (defined relative to *in*-sample predictions of y) increases or decreases if x_i if removed from the model. If a particular x_j is never or rarely selected in a rule, it is thought not to be causal, or of low causal power; otherwise it is seen in the same senses as with regression above. The ascription of cause is as with parameterized models: the language is vague and claims of cause are shifting.

The hope of these non-parameterized models, especially of human behavior or biology, is that if only enough *different* x_i are collected y can be predicted with certainty or with near-certainty. Even if it is acknowledged that this ultimate-data-collection hope will not be realized in practice, it still exists as a hope. When it does, it is a tacit acknowledgement the models are thought causal, at least partially. This hope also assumes all causes are measurable, and therefore are material. These are very strong claims. In physics, this hope is known to be false, as said. Perhaps a Bell-like theorem of the unknowability of cause in human behavior will be somebody be found.

This is very important for economics since all measures and observables are caused ultimately by human behavior, so it is worth emphasizing. For some physical processes in quantum mechanics, it is well known that perfect, i.e. certain, prediction is impossible. Measures x_i cannot be found that will bring certainty. This is why, besides Bell's proofs, Heseinberg's Uncertainty Principle has such an apt name. Free will, or free choice, which of course is implicit in all observables about human behavior, is also not predictable in an analogous way. Some who use a computer-as-brain metaphor believe that all human behavior is in principle deterministic, i.e. it has distinct material causes, and that free will is an "illusion" (it is never stated who is having the illusion). But these expressions are hopeful only, because no causal mechanism is known that can cause a being to believe it is conscious and has experiences of choice. Indeed, some philosophers believe human intellect and will to be non-material ([33]), and therefore that which is causing changes in will cannot be measured (*effects* of will can, however, be measured). Others argue against the computer-as-brain metaphor, e.g. [75]. These counter-arguments (to the hope that all life is strictly deterministic) collectively can be taken, at least in a metaphorical sense, as the start of a Bell-like theorem that cause of will cannot with certainty be known. This means models which take any account of human behavior, which includes all economic and financial observables, must in the end remain in the correlational realm with regard to behavior.

4.3 When Cause is not a Cause

It cannot be denied that cause can be known, and that some causal x can be put into uncertainty models. Indeed, when this happens, which it does especially in those models which are frequently tested against reality, it is discovered testing (and p-values) "verifies" these causes. Yet they were known, or suspected, before testing. And indeed must have been, as we shall see. How cause is known is discussed presently. We must first understand cause is not simple. We remind the reader of a simple example. Consider an experiment which measures exposure or not to some dread thing and whether or not people developed some malady. Hypothesis testing might "link" the exposure to the disease; and, indeed, everybody would act as if they believed the exposure has caused the malady. But the exact opposite can be true.

It will almost surely be the case that not everybody in the exposed group will have developed the malady; and it will also be that some people in the not-exposed group will also have the malady. It thus cannot be that the people in not-exposed group had their disease caused by the exposure, for of course they were not exposed. It then necessarily follows that their malady was caused by something other than the exposure. This, then, is definitive proof that at least one more cause than the supposed cause of the exposure exists. There is no uncertainty in this judgment.

We still do not know if the exposure was a cause. It could be that every person in the exposed group had their disease caused by whatever caused the disease in the not-exposed group—or there could even be other causes that did not affect anybody in the not-exposed group but that, somehow, caused disease in the exposed group. It could be that exposure caused some disease, but there is no way to tell, without outside assumptions, how many more maladies (besides the known other cause(s)) were caused by the exposure.

It's worse still for those who hold uncertainty models can discover cause. For how do we explain those people in either group who did not develop the disease? Even if exposure causes disease sometimes, and the other (unknown-but-not-exposure) cause which we know exists only causes disease sometimes, we still do not know why the exposure or non-exposure causes disease only sometimes. Why did these people develop the malady and these not? We don't know. We can "link" various correlations as "cause blockers" or "mitigators", but we're right back where we started from. We don't know, from the data alone, what is a cause and what is not, and what blocks these (at least) two causes sometimes but not in others.

Cause therefore, like probability, is known conditionally. In our most rigorous attempts at discovering cause, we design an experiment to demonstrate a causal connection or effect: x is believed to cause a change in y. Every known or believed cause or moderator of y or of x is controlled, to the best of the experimenter's ability. Then x is deployed or varied and the change in y is measured as precisely as possible. If after x changes we see y change, or we see a change of a certain magnitude and the like, we say x is indeed a cause of y (in some form).

But this judgment supposes we have correctly identified *every* possible cause or moderator of y or of x. Since in science we deal with the contingent, y will be contingent. This means that even in spite of our certainty that x is a or is the cause of y, there always exists the possibility (however remote) that something unknown was responsible for the observations, or that blocked or prevented x from using its powers to change y; [29, 87, 88]. This possibility is theoretical, not necessarily practical. In practice we always limit the number of possible causes x to a finite set. If we design an experiment to deduce if $x =$ "gravity" caused the $y =$ "pencil to drop", we naturally assume, given our background knowledge about gravity and its effects on things like pencils, that the pencil will drop be*cause* of gravity. Yet if the pencil does drop, it remains a theoretical possibility that something else, perhaps a mysterious pencil-pulling ray activated by pranksters, caused the pencil drop and not gravity. There is no way to prove this is not so. Yet we *implicitly* (and rightly!) condition our judgment on the non-existence of this ray and on any other freakish cause. It is that (proper) conditioning which is key to our understanding of cause.

This discussion might seem irrelevant, but it is not. It in fact contains the seed of the proof that algorithms automated to discover cause of observables must contain *in advance* at least the true cause of y. And so it must be that cause, or rather knowledge of cause, is something that can only be had in the mind. No algorithm can determine what was a cause or was not, unless that algorithm was first "told" which are causes and which not. This is proved shortly. The point of this exercise is to exhort researchers from preaching with too much vigor and too much certainty about their results, especially in those instances where a researcher claims a model has backed up his claims. All models only do what they were told; that a model fits data is not independent verification of the model's truth.

A true cause x of y, given *identical* conditions and where *identical* is used in its strictest sense, will always lead to perfect correlations (not necessarily linear, of course) of an observable. An algorithm can certainly note this perfect correlation, and it can be "told" to say things like "If a perfect correlation is seen at least so-many times, a cause exists." But perfect correlations are not always indicative of cause. Samples, even though thought large, can be small, and the correlation spurious. The direction of causality can be backwards, where it's not x causing y, but y causing x. Third, and of even greater importance, removed measures might be causing the observations: e.g. w is causing v and z which are in turn causing y and x, which is all we measure. These kinds of remote-cause scenarios multiply. In the last, if w is not measured, or v or z are not, then the algorithm has no way to identify cause. If they are measured and in the algorithm, it must be because the cause was already suspected or known.

4.4 Cause is in the Mind, not the Data

Suppose you fed into the algorithm a series of numbers, starting at 1, then 2, and so on. The machine discovers the rule that for any three of these numbers x, y, and z "If $x = y$ and $y = z$, then $x = z$. " The rule is true for all the numbers fed into the algorithm. But is it *always* true, i.e. true for numbers not yet fed into the algorithm? The algorithm cannot tell us—unless, again, it were pre-programmed to announce that after so-many examples the universal is true. By "universal" we mean the proposition holds for an infinity of natural numbers. This example is, of course the second of Peano's axioms for mathematics, believed by all (who consider it) to be true. But that is because humans have the ability to extract this universal from data, whereas an algorithm cannot. In this case, no algorithm can ever consider an infinity of numbers; whereas, we can. Consider that the universal is not necessarily true in this case, because the algorithm may have been fed numbers from some process which after a point sees the sequence become intransitive, at which point the rule breaks down. The algorithm cannot know what is beyond what it sees. The difference from the universal and the eventually intransitive sequence is conditioning. When we judge the axiom true, we condition on the idea it applies to numbers not tied to any process, except their progression. But the algorithm cannot know this; it cannot *know* anything. This is why it can be fooled if fed limited sequences. People can be fooled, too, of course, but a person and not an algorithm would understand whether the input was part of a contingent process or was purposely the sequence of natural numbers.

Now that is an obscure and philosophical answer to the question whether algorithms can discover causes. It is also contra to some who argue algorithms can indeed mimic human thinking; perhaps not now, but eventually; see [83]. So here is simple proof that cause is in the mind and not the data. A silly thought experiment: Pick something that happened. It doesn't matter what it is, as long as it happened. Something caused this thing to happen; which is to say, something actual turned the potential (of the thing to happen) to actuality. All four facets of cause were involved, of course. I will take, as my example, the death of Napoleon. One afternoon he was spry, sipping his Grand cru, and planning his momentous second comeback, and the next morning he was smelling like week-old Brie.

Next we want to design an algorithm to discover the cause of this thing (in all four aspects of cause, or even just the efficient cause). This can be a regression, machine learning routine, neural net, deep learning algorithm, artificial intelligence routine, anything you like. Plug into this algorithm, or into a diagram in the computer, or into whatever device you like, THE EVENT. Then press "GO" or "ACTIVATE" or whatever it is that launches the algorithm into action. What will be the result?

Nothing, of course. This lifeless algorithm cannot discover cause, because it has left out "data". There is nothing for the algorithm to process: no data to work on. There are no "x" to tie to the "y", i.e. the event. So which data should we put in? We have to choose. There are an infinite number of measures (x) available, and any machine we design will have finite capacity. We must do some kind of winnowing to select only certain of these infinite x. Which? How about these (keep in mind my event): (x_1) The other day I was given a small bottle of gin in the shape of a Dutch house in delft blue. (x_2) You weren't supposed to drink the gin, but I did. (x_3) In my defense, I wasn't told until after I drank it that I shouldn't have. (x_4) It wasn't that good.

Now if these x seem absurd to you,

you have proven that cause is in the mind and not the data, that it is you who are extracting cause from data and not algorithms. For you have used your knowledge of cause to discern there could not be any possible connection between a novelty bottle of gin and the death of a tyrant two centuries previously. We can't let an algorithm figure out the cause if we do not first feed the algorithm x which we believe or suspect are in the causal path of y, and tell it which measure of relation between an x and y is appropriate to use in judgment and whether this measure has crossed the admitted threshold. And even this does solve the full causal path problem, where other measures may be causing the x and y.

Again, there are an infinite number of measures x. Everything that's ever happened, in the order it happened, is data. That's a lot of data. How can any algorithm pick cause out of all that to tell us the cause of any event? That tall order is thus not only tall, but impossible, too, since everything that's ever happened wasn't, for the most part, measured. And even it if it was, no device could store all this data or manipulate it.

We could argue that only data related to y in some way should be input into the algorithm, perhaps the relations discovered by previous algorithms. But *related* must mean those measures which are the cause of the event, or which are not the direct causes, but incidental ones, perhaps measures caused by the event itself, or measures that caused the cause of the event, and that sort of thing. Those measures which are in (we can call it) *the causal path.* Any previous algorithm used to winnow x down to a suitable subset of related items must itself have been told which x of the infinite choices to start with. And so on for any algorithms "upstream" of ours. There thus must come a point where human intelligence, which has the ability to extract universals like cause from data, albeit imperfectly, comes into play and does the choosing. Algorithms are thus only good for automating the tedious tasks of computation.

The best any algorithm can do is to find prominent correlations, which may or may not be directly related to the cause itself (and some may be spurious), using whatever rules of "correlation" we pre-specify. These correlations will be better or worse depending on our understanding of the cause and therefore of what "data" we feed our algorithm. The only way we know these data are related to the cause, or are the cause, is because we have a power algorithms can never have, which is the extraction of and understanding of universals. To wrap up: cause is hard, and we're better off claiming nothing more than correlation in most instances.

4.5 Tests of Stationarity and Cause

The notion of stationarity is ubiquitous in time series modeling, e.g. [18]. There is often great concern that a given series is not stationary, a concern which has given rise to a suite of tests, like the unit-root or Dickey Fuller, e.g. [30]. We need not explore the mathematical details. The idea is that the probability model for a non-stationary series for observable y_t has parameters, such as those representing covariance, that change in time, and that if this is so, ordinary methods of estimation will fail when making statements about other parameters, such as those for auto-regression. This is true mathematically, but not causally.

Now all y_t in a series are caused: it cannot be, in ordinary time series models, that any y_{t-j} for any j and t caused y_t, in any efficient sense. For instance, last year's GDP did not cause this year's GDP to take the value it did. They may share common causes, of course, and they also may not. Since probability is not ontic, i.e. does not exist materially, it cannot be the series itself that is not stationary, but the causes underlying it are not constant. Cause can change in all the ways indicated above: by form, material, efficient power, or final. We have simple proof that at least one cause changed or a new one intervened whenever $y_t \neq y_{t-1}$. So when a series is said not to be stationary, it only means that the causes of the observable have changed enough so that the distribution representing uncertainty in the observable's variables must be changed at a point or points.

Explanation is, as we saw, a model goal, and knowledge of cause is the best explanation, and this is so in time series as in any other data. There are thus two goals in time series modeling: to say when a thing like a "change point" (change in the nature of causes) occurred, and prediction. Both goals are predictive.

If we seek to say when a change point happened, then we must posit a model of change, which we treat predictively like any other model. It will issue predictions $\Pr(\text{Change at } t|\text{D}_n\text{Y}) = p_t, t = 0, 1, \ldots, T$ (we don't need the X here). To pick which t was the culprit requires having a decision rule, just as we must have a decision rule to move from probability to point in any model.

Yet often we don't care when the change was, because our interest is in prediction. In that case, we don't have to pick which t was the change point, and we simply make predictions of $y_{t+k}, k > 0$ weighted (in the suitable and obvious way) by every possible change point t. The change point is integrated out like any parameter. In any case, testing is not needed.

4.6 The End of Cause

There is much more to this discussion (see [11]), but I have taxed the reader too much already, with much of the discussion seemingly arcane. The hope is that if the reader does not see the arguments above as definitive proof, then he see there is at least good evidence that our traditional means of ascribing cause are at the very least too certain, if not often in downright error.

Consider, lastly, a typical (but fictional) news headline, drawn from a scientific statistical study: "New Study Reveals Those Who Read At Least Five Books A Year Make More Money." The implication is, of course, that reading (at least five but not *four* books) causes higher incomes. News outlets will have none of the caveats we presume the conscientious researchers will put into their paper; the media (and many government agencies) will assume cause has been definitively proven. It is the fault of researchers everywhere for not correcting these exceedingly common overstatements.

Yet even if researchers admit to limitations, they will make the same mistakes as the press. We are in the same situation of the exposure and not-exposure example causing disease. Even if it is true that reading at least five (and not four or three) books causes higher incomes, it won't always do so, yet in the Discussion section of their paper the researchers will have surely launched into great theories of how the reading caused the higher incomes, implying the cause happened to all readers and with the same force. Even if the differences between (what are called) readers and non-readers is small, success will be claimed, when it is far from clear the direction of the cause is consistent, whether a third cause intervened, and whether the claimed cause is even real or reproducible.

Consider this more mundane headline that might be generated by the reality-based, predictive method using the same data: "There Is A 6% Chance Those Who Read At Least Five Books A Year Will Make At Least $400 More Annually, But Only If The Following Conditions Hold." This is both predictive (6%), and verifiable ($400+, the following conditions). Yet that kind of announcement, were it a university press release, while being vastly more honest, is not the sort of information that will get a researcher's (or the university's) name into the paper.

5 TRUST BUT VERIFY

However many models are left in consideration at the end of the modeling process, unless those models were deduced from first principles (see [11] Chapter 8 for examples), there will or should be uncertainty whether the model or models are of any use in reality. A tremendous weakness of hypothesis testing is that it certified, if you like, a model's goodness by requiring only that it evince at least one wee p-value. This is an absurd situation when we recall both how easy it is to produce wee p-values, and that the vast majority of models in use are *ad hoc*; regression being the largest example.

5.1 The Intrusion of Reality

Scarcely any who use statistical models ever ask *does the model work?* Not works in the sense that data can be fit to it, but works in the sense that it can make useful predictions of reality of observations never before seen or used in any way. Does the model verify? Verification is a strict test, a test models in physics, chemistry, meteorology, and all engineering fields must pass to be considered as viable. In those fields, models are not just proposed, but they are proposed and tested. Would you strap yourself into brand new a flying car built on speculative theoretical principles but was never tested in any way, except for how good the model looked on a computer?

Some fields never verify their models. They are content with hypothesis testing and the many opportunities for theorizing (some might say "pontificating") that method provides. There is thus no real way to know how good the models in these areas really are. Overconfidence must be the rule, however, else we would not have the replication crisis mentioned above.

Verification in economic data is not uncommon in time series models. Time series models are set in a naturally predictive form, where predictions are (or should be) a matter of course. Many formal verification methods have accordingly came from this research; e.g. [2]. Another fecund field is (readers might be surprised to learn) meteorology. Weather and climate forecasts appear with regularity and the demand for accuracy is keen. Great strides in mathematical verification methods have accordingly arisen in these areas; see [103].

The process of verification in its ideal form is simplicity itself: (1) create using old observations the model or models, (2) make probabilistic predictions using them, (3) wait for new observations to accrue, (4) and then score those models with respect to the decisions that were made using the predictions. This is exactly how you would assess driving a new flying car.

Scientists, economists, and other researchers are often impatient about step (3). New observations typically have not arrived by the time papers must be published, and, as everybody knows with absolute certainty, it really is publish or perish. Ideally, researchers would wait until they have accumulated enough new, never-before-used-in-any-way observations so that they could prove their proposed models have skill or are useful. The rejoinder to this is that requiring actual verification would slow research down. But this is a fallacy, because it assumes what it seeks to prove; namely, that the new research is worthy. The solution, which ought to please those in need of publications, is two-fold: use verification methods to estimate model goodness using the old data, which itself is a prediction, and then when new observations finally do become available, perform actual verification (and write a second paper about it). Of course, this last step might too often lead to disappointment as it can reveal depressing conclusions for those who loved their models too well.

The point about the observations used in verification having never been used in any way cannot be understressed. Many methods like cross validation use so-called verification data sets to estimate model goodness. The problem is that the temptation to tweak the original model so that it performs better on the verification set is too strong for most to resist. I know of no references to support this opinion, but having felt the temptation myself (and given in to it), I am sure it is not uncommon. Yet when this is done it in essence unites the training and validation data sets so that they are really one, and we do not have a true test of model goodness in the wild, so to speak.

Yet we do have to have some idea of how good a model might be. It may, for instance, be expensive to wait for new observations, or those observations may depend on the final model chosen. So it is not undesirable to have an estimate of future performance. This requires two elements: a score or measure of goodness applied to old observations, and a new model of how that score of measure will reproduce in new observations. As for scores and measures, there are many: years of research has left us well stocked with tools for assessing model predictive performance; e.g. [41, 16, 73, 74, 85, 55, 15]. A sketch of those follows presently. But it is an open question, in many situations, how well scores and measures predict future performance, yet another area wide open for research. In order to do this well, we not only need skillful models of observables Y, but also of the measures X, since all predictions are conditional on those X. The possibilities here for new work are nearly endless.

5.2 Verification Scores

Here is one of many examples of a verification measure, the continuous ranked probability score (CRPS), which is quite general and has been well investigated, e.g. [40, 55]. We imagine the CRPS to apply to old observations here for use in model fit, but it works equally well scoring new ones.

Let $F_i(s) = \Pr(\mathrm{Y} < s|\mathrm{X}_i\mathrm{D}_n\mathrm{M})$, i.e. a probabilistic prediction of our model for past observation X_i. Here we let s

vary, so that the forecast or prediction is a function of s, but s could be fixed, too. Let Y_i be the i-th observed value of Y. Then

$$\mathrm{CRPS}(F, \mathrm{Y}) = \sum_i \left(F_i - \mathrm{I}\{s \geq \mathrm{Y}_i\}\right)^2 \tag{18}$$

where I is the indicator function. The score essentially is a distance between the cumulative distribution of the prediction and the cumulative distribution of the observation (a step function at Y_i). A perfect forecast or prediction is itself a step function at the eventually observed value of Y_i, in which case the CRPS at that point is 0. Lower scores in verification measures are better (some scores invert this). The "continuous" part of the name is because (18) can be converted to continuity in the obvious way; see below for an example. If F is not analytic, numerical approximations to CRPS would have to suffice, though these are easy to compute. When $F_i = p_i$, i.e. a single number, which happens when Y is dichotomous, the CRPS is called the Brier Score.

Expected scores are amenable to decomposition, showing the constituent components of performance; e.g. [17]. One is usually related to the inherent variability of the observable, which translates into an expected non-zero minimum of a given score (for a certain model form); for a simple example of the Brier scores, see [72]. This expected minimum phenomenon is demonstrated below in the continuing example. Model goodness is not simply captured by one number, as with a score, but in examining calibration, reliability, and sharpness. Calibration is of three dimensions: calibration in probability, which is when the model predictions converge to the prediction-conditional relative frequency of the observables, calibration in exceedance and calibration in average; these are all mathematically defined in [41]. There is not the space here to discuss all aspects of scoring, nor indeed to give an example of validation in all its glory. It is much more revealing and useful than the usual practice of examining model residuals, for a reason that will be clear in a moment.

CRPS is a proper score and it is sensitive to distance, meaning that observations closer to model predictions score better. *Proper scores* are defined conditional (as are all forecasts) on X, D_n and M; see [91] for a complete theoretical treatment of how scores fit into decision analysis. Given these, the proper probability is F, but other probabilities could be announced by scheming modelers; say they assert $G \neq F$ for at least some s, where G is calculated conditional on tacit or hidden premises not part of M. Propriety in a score is when

$$\sum_i \mathrm{S}(G_i, \mathrm{Y}_i) F_i \geq \sum_i \mathrm{S}(F_i, \mathrm{Y}_i) F_i. \tag{19}$$

In other words, given a proper score the modeler does best when announcing the full uncertainty implied by the model and in not saying anything else. Propriety is a modest requirement, yet it is often violated. The popular scores RMSE, i.e. $\sqrt{\sum_i (\hat{F}_i - \mathrm{Y}_i)^2/n}$, mean absolute deviation, i.e. $\sum_i |\hat{F}_i - \mathrm{Y}_i|/n$, where $\hat{F}_i$ is some sort of point forecast derived as a function of F_i, are not proper. The idea is information is being thrown away by making the forecast into a point, where we should instead be investigating it as a full probability: a point is a decision, and not a probability. Of course, points will arise when decisions must be made, but in those situations actual and not theoretical cost-loss should be used to verify models, the same cost-loss that led to the function that computed the points. Similarly, scores like R^2 and so on are also not proper.

If F (as an approximation) is a (cumulative) normal distribution, or can be approximated as such, then the following formula may be used (from [40]):

$$\begin{aligned} &\mathrm{CRPS}(\mathrm{N}(m, s^2), \mathrm{Y}) = \\ &s\left(\frac{1}{\sqrt{\pi}} - \frac{\mathrm{Y}-m}{s}\left(2\Phi\left(\frac{\mathrm{Y}-m}{s}\right) - 1\right)\right) \\ &- s\left(2\phi\left(\frac{\mathrm{Y}-m}{s}\right)\right) \end{aligned} \tag{20}$$

where ϕ and Φ are the standard Normal probability density function and cumulative distribution function, and m and s are *known* numbers given by our prediction. These could arise in regression, say, with conjugate priors. Estimates of (20) are easy to have in the obvious way.

CRPS, or any score, is calculated per prediction. For a set of predictions, the sum or average score is usually computed, though because averaging removes information it is best to keep the set of scores and analyze those. CRPS_i can be plotted by Y_i or indeed any x_i. Here is another area of research about how best to use the information given in verification score.

Next we need the idea of *skill*. Skill is had when one model demonstrates superiority over another, given by and conditional on some verification measure. Skill is widely used in meteorology, for example, where the models being compared are often persistence and the fancy new theoretical model proposed by a researcher. This is a highly relevant point because persistence is the forecast that essentially says "tomorrow will look exactly like today", where that statement is affixed with the appropriate uncertainty, of course. If the new theoretical model cannot beat this simple, naive model, it has no business being revealed to the public. Economists making time series forecasts are in exactly the same situation. Whatever model they are proposing should *at least* beat persistence. If it can't, why should the model be trusted when it isn't needed to make good predictions?

It is not only times series models that benefit by computing skill. It works in any situation. For example, in a regression, where one model has p measures and another, say, has $p + q$. Even if a researcher is happy with his model with p measures, it should at least be compared to one with none, i.e. where uncertainty in the observable is characterized by the distribution implied by the model with no measures. In regression, this would be the model with only the intercept. If the model with the greater number of measures cannot beat the one with fewer, the model with more is at least suspect or has no skill. Because of the potential for over-fitting, it is again imperative to do real verification on brand new observations. How skill and information theoretic measures are related is another open area of investigation.

Skill scores K have the form:

$$\mathrm{K}(F, G, \mathrm{Y}) = \frac{\mathrm{S}(G, \mathrm{Y}) - \mathrm{S}(F, \mathrm{Y})}{\mathrm{S}(G, \mathrm{Y})}, \tag{21}$$

where F is the prediction from what is thought to be the superior or more complex model and G the prediction from the inferior. Skill is always relative. Since the minimum best score is $\mathrm{S}(F, \mathrm{Y}) = 0$, and given the normalization, a perfect skill score has $\mathrm{K} = 1$. Skill exists if and only if $\mathrm{K} > 0$, else it is absent. Skill like proper scores can be computed as an average over a set of data, or individually over separate points. Plots of skill can be made in an analogous way. Receiver operating characteristic (ROC) curves, which are very popular, are not to be preferred to skill curves since these do not answer questions of usefulness in a natural way; see [16] for details.

We should insist that no model should be published without details of how it has been verified. If it is has not been verified with never-before-seen observations, this should be admitted. Estimates of how the model will score in future observations should be given. And skill must be demonstrated, even if this is only with respect to the simplest possible competitive model.

5.3 Example Continued

We now complete the housing price example started above. Fig. 2 showed all predictions based on the assumption that old X were likely to be seen in the future, which is surely not implausible. Individual predictions with their accompanying observations can also be shown, as in Fig.2, which shows the four observations of the data (picked at random), and with predictions assuming their X are new.

These plots are in the right format for computing the CRPS, which is shown in Figs. 4 and 5. The first shows the normalized (the actual values are not important, except relatively) individual CRPS by the observed prices. Scores away from the middle prices are on average worse, which will not be a surprise to any user of regression models, only here we have an exact way of quantifying "better" and "worse." There is also a distinct lowest value of CRPS. This is related to the inherent uncertainty in the predictions, conditional on the model and old data, as mentioned above. This part of the verification is most useful in communicating essential limitations of the model. Perfect predictions, we project, assuming the model and CRPS will be used on genuinely new data, are not possible. The variability has a minimum, which is not low. An open question is whether this lower bound can be estimated in future data from assuming the model and data; equation (20) implies the answer is at least yes sometimes (it can be computed in expected value assuming M's truth).

Now plots like Fig. 4 can made with CRPS by the Y or X, and it can be learned what exactly is driving good and bad performance. This is not done here. Next, in Fig. 5 the individual CRPS of both models, with and without nox, are compared. A one-to-one line is overdrawn. It is not clear from examining the plot by eye whether adding nox has benefited us.

Finally, skill (21) is computed, comparing the models with and without nox. A plot of individual skill scores as related to nox is given in Fig. 6. A dashed red line at 0 indicates points which do not have skill. Similar plots for the other measures may also be made.

The full model does not do that well. There are many points at which the simpler model bests the more complex one, with the suggestion that for the highest values of nox the full model does worst. The overall average skill score was K = -0.011, indicating the more complicated model (with nox) does not have skill over the less complicated model. This means, as described above, that if the CRPS represents the actual cost-loss score of a decision maker using this model, the prediction is that in future data, the simpler model will outperform the more complex one.

Whether this insufficiency in the model is due to probability leakage, or that the CRPS is not the best score in this situation, remain to be seen.

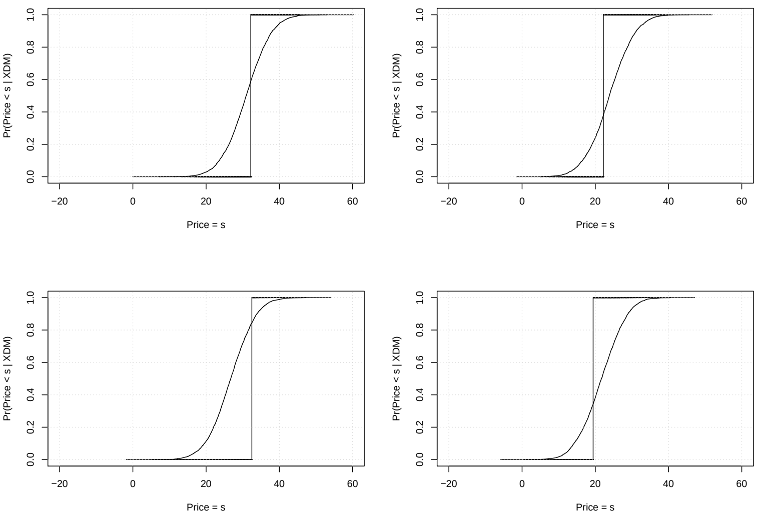


Fig. 3 The probability prediction of housing prices at four old X, assumed as new. An empirical CDF of the eventual observation at each X is over-plotted.

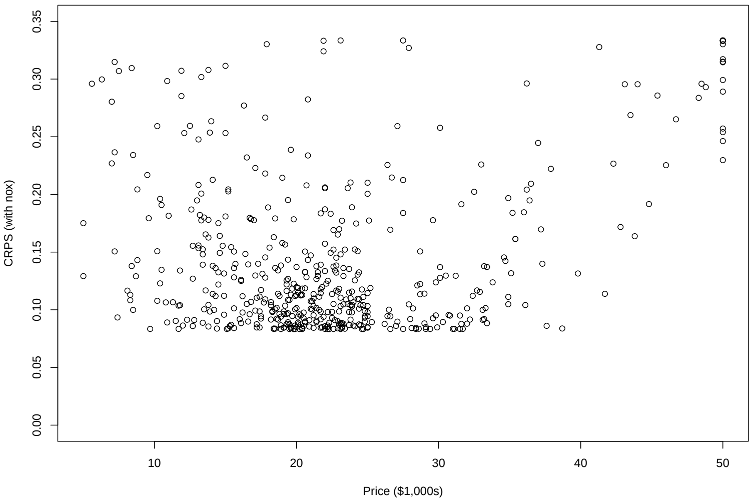


Fig. 4 The individual CRPS scores (normalized to 1) by the price of houses (in $1,000s). Lower scores are better. Not surprisingly, scores away from the middle prices are on average worse.

We have thus moved from delightful results as indicated by p-values, to more sobering results when testing the model against reality—where we also recall this is only a guess of how the model will actually perform on future data: nox is not useful. Since the possibility for over-fitting is always with us, it is the case that future skill measures would likely be worse than those seen in the old data.

6 THE FUTURE

As the example suggests, the failure to generate exciting results might explain why historically the predictive method was never adopted. Reality is

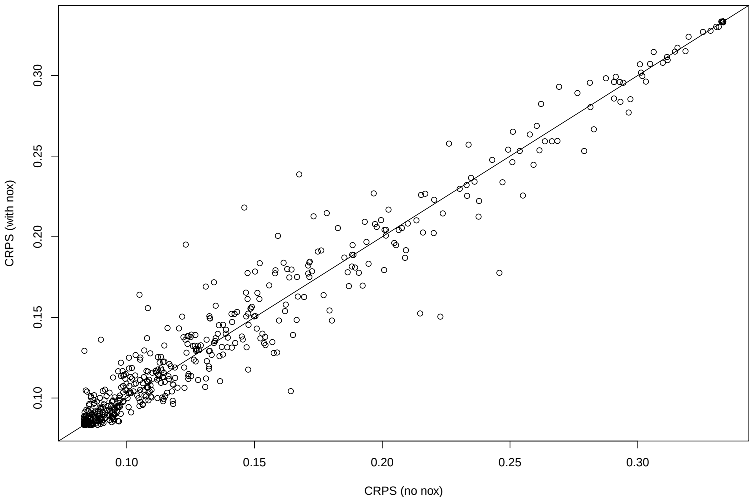


Fig. 5 The individual CRPS scores of the full model, with now, by the CRPS of the model removing nox. A one-to-one line has been overdrawn. There does not seem to be much if any improvement in scores by adding nox to the model.

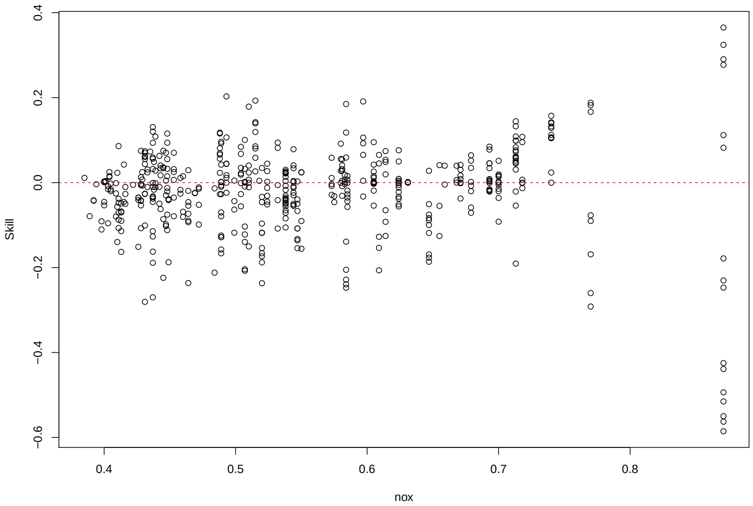


Fig. 6 The individual skill scores comparing models with the old X with and without nox, as related to nox. A dashed red at 0, indicating no skill, has been drawn.

a harsh judge. Yet above if it was considered verification was judgmental in-sample, imagine the shock when it will be learned verification is downright cruel out-of-sample, i.e. in new, never before seen observations. The reality-based approach is therefore bound to lead to disappointment where before there was much joy. Such is often the case in science when, as the saying goes, a beautiful theory meets ugly facts.

There should by now at least be more suspicion among researchers that models have truly identified cause. We have seen the many confusing and disparate ideas about cause which are common in uncertainty models, a class which includes so-called artificial intelligence and machine learning algorithms. We don't often recognize that nothing can come out any algorithm which was not placed there, at the beginning, by the algorithm creator. Algorithms are useful for automating tedious processes, but coming to knowledge of cause requires a much deeper process of thinking. Every scientist believes in confirmation bias; it's just they always believe it happens to the other guy.

Creators of models and algorithms are one class, users are another. The chance of over-certainty increases in use and interpretations of models in this latter class because a user will not on average be as aware of the shortcomings, limitations, and intricacies of models are creators are. All common experience bears out that users of models are more likely to ascribe cause to hypotheses than more careful creators of models. The so-called replication crisis can in part be put down to non-careful use of models, in particular the unthinking use of hypothesis testing; e.g. [3, 99].

The situation is even worse than it might seem, because beside the formal models considered here, there is another, wider, and more influential class, which we might call media models. There is little to no check on the wild extrapolations that appear in the press (and taken up in civic life). I have a small collection of headlines reporting on medical papers, each contradicting the other, and all trumpeting that causes have been discovered (via testing); see [10]. One headline: "Bad news for chocoholics: Dark chocolate isn't so healthy for you after all," particularly not good, the story informs, for heart disease. This was followed by another headline three short months later in the same paper saying"Eating dark chocolate is good for your heart." Similar collections for economics studies could easily and all too quickly be compiled.

It could be argued the ultimate responsibility is on the people making the wild and over-sure claims. This holds some truth, but the appalling frequency that this sort of thing happens without any kind of corrections from authorities (like you, the reader) implies, to the media, that what they are doing is not wrong.

Bland warnings cannot take the place of outright proscriptions. We must ban the cause of at least some of the over-certainty. No more hypothesis testing. Models must be reported in their predictive form, where anybody (in theory) can check the results, even if they don't have access to the original data. All models which have any claim to sincerity must be tested against reality, first in-sample, then out-of-sample. Reality must take precedence over theory.

References

[1] Amrhein, V., Korner-Nievergelt, F. and Roth, T. (2017). The Earth is flat ($p > 0.05$): Significance Thresholds and the Risis of Unreplicable Research. *PeerJ*, 5:e3544.

[2] Armstrong, J. S.(2007). Significance Testing Harms Progress in Forecasting (with discussion). *International Journal of Forecasting*, 23, 321-327.

[3] Begley, C. G. and Ioannidis, J. P. (2015). Reproducibility in Science: Improv-

ing the Standard for Basic and Preclinical Research. *Circulation Research*, 116, 116-126.

[4] Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R. and et al (2018). Redefine Statistical Significance. *Nat. Hum. Behav.*, 2, 6-10.

[5] Berger, J. O. and Selke, T. (1987). Testing a Point Null Hypothesis: The Irreconcilability of P-values and Evidence. *JASA*, 33, 112-122.

[6] Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory*. Wiley, New York.

[7] Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science, 16(3), 199-215.*

[8] Briggs, W. M. (2006). Broccoli Reduces the Risk of Splenetic Fever! The use of Induction and Falsifiability in Statistics and Model Selection. arxiv.org/pdf/math.GM/0610859.

[9] Briggs, W. M. (2013). On Probability Leakage. *arxiv.org/abs/1201.3611.*

[10] Briggs, W. M. (2014). Common Statistical Fallacies. *Journal of American Physicians and Surgeons*, 19(2), 678-681.

[11] Briggs, W. M. (2016). *Uncertainty: The Soul of Probability, Modeling & Statistics*. Springer, New York.

[12] Briggs, W. M. (2017). The Substitute for P-values. *JASA*, 112, 897-898.

[13] Briggs, W. M. (2019). *Everything Wrong with P-values under one Roof. In Kreinovich, V., Thach, N. N., Trung, N. D. and Thanh, D. V. editors, Beyond Traditional Probabilistic Methods in Economics, 22-44.* Springer, New York.

[14] Briggs, W. M., Nguyen, H. T. and Trafimow, D. (2019). *The Replacement for Hypothesis Testing. In V. Kreinovich and S. Sriboonchitta, editors, Structural Changes and Their Econometric Modeling, 3-17.* Springer, New York.

[15] Briggs, W. M. and Ruppert, D. (2005). Assessing the Skill of yes/no Predictions. *Biometrics*, 61(3),799-807.

[16] Briggs, W. M. and Zaretzki, R. A. (2008). The Skill Plot: A Graphical Technique for Evaluating Continuous Diagnostic Tests. *Biometrics*, 64, 250-263. (with discussion).

[17] Brocker, J. (2009). Reliability, Sufficiency, and the Decomposition of Proper Scores. *Quarterly Journal of the Royal Meteorological Society*, 135, 1512-1519.

[18] Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods.* Springer, New York, NY.

[19] Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Johannesson, M., Kirchler, M., Almenberg, J. and Altmejd, A. (2016). Evaluating Replicability of Laboratory Experiments in Economics. *Science*, 351, 1433-1436.

[20] Campbell, S. and Franklin, J. (2004). Randomness and Induction. *Synthese*, 138, 79-99.

[21] Carroll, R. J, Ruppert, D., Stefansky, L. A. and Crainiceanu, C. M. (2006). *Easurement Error in Nonlinear Models: A Modern Perspective.* Chapman and Hall, London.

[22] Chang, A. C. and Li, P. (2015). *Is Economics Research Replicable? Sixty Published papers from Thirteen Journals Say 'Usually Not'.* Technical Report 2015-083, Finance and Economics Discussion Series, Divisions of Research & Statistics and Monetary Affairs, Federal Reserve Board, Washington, D.C.

[23] Cintula, P., Fermuller, C. G. and Noguera, C. (2017). *Fuzzy Logic. In Edward N. Zalta, editor, The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab*, Stanford University.

[24] Clarke, B. S. and Clarke, J. L. (2018). *Predictive Statistics.* Cambridge University Press.

[25] Cohen, J. (1995). The Earth is Round ($p < .05$). *American Psychologist*, 49, 997-1003.

[26] Colquhoun, D. (1979). An Investigation of the False Discovery Rate and the Misinterpretation of P-values. *Royal Society Open Science*, 1, 1-16.

[27] Cook, T. D. and Campbell, D. T. (1979). *Quasi-Experimentation: Design & Analysis Issues for Field Settings.* Houghton Mifflin, Boston.

[28] Crosby, A. W. (1997). *The Measure of Reality: Quantification and Western Society, 1250-1600.* Cambrigde University Press.

[29] Duhem, P. (1954). *The Aim and Structure of Physical Theory.* Princeton University Press.

[30] Einstein, A., Podolsky, P. and Rosen, N. (2001). Testing for Unit Roots: What Should Students be Taught? *Journal of Economic Education*, 32, 137-146.

[31] Falcon, A. (2015). *Aristotle on Causality. In Edward N. Zalta, editor, The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab*, Stanford University.

[32] Feser, E. (2010). *The Last Superstition: A Refutation of the New Atheism.* St. Augustines Press, South Bend, Indiana.

[33] Feser, E. (2013). Kripke, Ross, and the Immaterial Aspects of Thought. *American Catholic Philosophical Quarterly*, 87, 1-32.

[34] Feser, E. (2014). *Scholastic Metaphysics: A Contemporary Introduction.* Editions Scholasticae, Neunkirchen-Seelscheid, Germany.

[35] Franklin, J. (2014). *An Aristotelian Realist Philosophy of Mathematics: Mathematics as the Science of Quantity and Structure.* Palgrave Macmillan, New York.

[36] Freedman, D. (2005). *Linear Statistical Models for Causation: A Critical Review. In Brian S. Everitt and David Howell, editors, Encyclopedia of Statistics in Behavioral Science*, 2-13.Wiley, New York.

[37] Geisser, S. (1993). *Predictive Inference: An Introduction.* Chapman & Hall, New York.

[38] Gelman, A. (2000). Diagnostic Checks for Discrete Data Regression Models using Posterior Predictive Simulations. *Appl. Statistics*, 49(2), 247-268.

[39] Gigerenzer, G. (2004). Mindless Statistics. *The Journal of Socio-Economics*, 33, 587-606.

[40] Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *JASA*, 102, 359-378.

[41] Gneiting, T., Raftery, A. E. and Balabdaoui, F. (2007). Probabilistic Forecasts, Calibration and Sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69, 243-268.

[42] Goodman, S. N. (2001). Of P-values and Bayes: A Modest Proposal. *Epidemiology*, 12, 295-297.

[43] Greenland, S. (2000). Causal Analysis in the Health Sciences. *JASA*, 95, 286-289.

[44] Greenland, S. (2017). The Need for Cognitive Science in Methodology. *Am. J. Epidemiol.*, 186, 639-645.

[45] Greenland, S. (2018). Valid p-values Behave Exactly as they Should: Some Misleading Criticisms of P-values and Their Resolution with S-values. *Am. Statistician.*.

[46] Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N. & Altman, D. G. (2016). Statistical tests, P values, Confidence Intervals, and Power: A Guide to Misinterpretations. *European Journal of Epidemiology*, 31(4), 337-350.

[47] Greenwald, A. G. (1975). Consequences of Prejudice Against the Null Hypothesis. *Psychological Bulletin*, 82(1), 1-20.

[48] Groarke, L. (2009). *An Aristotelian Account of Induction.* Mcgill Queens University Press, Montreal.

[49] Haber, N., Smith, E. R., Moscoe, E., Andrews, K., Audy, R., Bell, W., Brennan, A. T., Breskin A, Kane, J. C., Karra, M., McClure, E. S. and Suarez, E. A. (2018). *Causal Language and Strength of Inference in Academic and Media Articles Shared in Social Media (claims): A Systematic Review.* PLOS One.

[50] Hájek, A. (1997). Mises Redux - Redux: Fifteen Arguments Against Finite Frequentism. *Erkenntnis*, 45, 209-227.

[51] Hájek, A. (2007). *A Philosopher's Guide to Probability. In Uncertainty: Multi-disciplinary Perspectives on Risk*, Earthscan.

[52] Hájek, A. (2009). Fifteen Arguments Against Hypothetical Frequentism. *Erkenntnis*, 70, 211-235.

[53] Harrell, F. (2018). *A Litany of Problems with P-values*, Aug 2018.

[54] Harrison, D. and Rubinfeld, D. L. (1978). Hedonic Prices and the Demand for Clean Air. *Journal of Environmental Economics and Management*, 5, 81-102.

[55] Hersbach, H. (2000). Decompostion of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting*, 15, 559-570.

[56] Hitchcock, C. (2018). *Probabilistic Causation. In EdwardN. Zalta, editor,* The Stanford Encyclopedia of Philosophy. *Metaphysics Research Lab*, Stanford University.

[57] Gale, R. P., Hochhaus, A. and Zhang, M. J. (2016). What is the (p-) value of the P-value? *Leukemia*, 30, 1965-1967.

[58] Holmes, S. (2018). Statistical Proof? The Problem of Irreproducibility. *Bulletin of the American Mathematical Society*, 55, 31-55.

[59] Hubbard, R. and Lindsay, R. M. (2008). Why P values are not a Useful Measure of Evidence in Statistical Significance Testing. *Theory & Psychology*, 18, 69-88. Ion2005Ioannidis, J. P. A. (2005). *Why Most Published Research Findings are False.* PLoS Medicine, 2(8), e124.

[60] Ioannidis, J. P. A., Stanley, T. D. and Doucouliagos, H. (2017). The Power of Bias in Economics Research. *The Economic Journal*, 127, F236-F265.

[61] Johnson, W. O. (1996). *Modelling and Prediction: Honoring Seymour Geisser, Chapter Predictive Influence in the Log Normal Surival Model, 104-121.* Springer.

[62] Johnson, W. O. and Geisser, S. (1982). A Predictive view of the Detection and Characterization of Influence Observations in Regression Analysis. *JASA*, 78, 427-440.

[63] Kastner, R. E., Kauffman, S. and Epperson, M. (2017). Taking Heisenberg's Potentia Seriously. *arXiv e-prints.*

[64] Keynes, J. M. (2004). *A Treatise on Probability.* Dover Phoenix Editions, Mineola, NY.

[65] Konishi, S. and Kitagawa, G. (2007). *Information Criteria and Statistical Modeling.* Springer, New York.

[66] Lee, J. C., Johnson, W. O. and Zellner, A. (Eds.) (1996). *Modelling and Prediction: Honoring Seymour Geisser.* Springer, New York.

[67] Lord, F. M. (1967). A Paradox in the Interpretation of Group Comparisons. *Psychological Bulletin*, 66, 304-305.

[68] Mayr, E. (1992). The Idea of Teleology. *Journal of the History of Ideas,*, 53(1), 117-135.

[69] McShane, B. B., Gal, D., Gelman, A., Robert, C. and Tackett, J. L. (2017). Abandon Statistical Significance. *The American Statistician.*

[70] Meng, X. L. (2008). Bayesian Analysis. *Cyr E. M'Lan and Lawrence Joseph and David B. Wolfson*, 3(2), 269-296.

[71] Mulder, J. and Wagenmakers, E. J. (2016). Editor's Introduction to the Special Issue: Bayes Factors for Testing Hypotheses in Psychological Research: Practical Relevance and New Developments. *Journal of Mathematical Psychology*, 72, 1-5.

[72] Murphy, A. H. (1973). A New Vector Partition of the Probability Score. *Journal of Applied Meteorology*, 12, 595-600.

[73] Murphy, A. H. (1987). Forecast Verification: Its Complexity and Dimensionality. *Monthly Weather Review*, 119, 1590-1601.

[74] Murphy, A. H. and Winkler, R. L. (1987). A General Framework for Forecast Verification. *Monthly Weather Review*, 115, 1330-1338.

[75] Nagel, T. (2012). *Mind & Cosmos: Why the Materialist Neo-Darwinian Conception of Nature is Almost Certainly False.* Oxford University Press, Oxford.

[76] Neyman, J. (1937). Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philosophical Transactions of the Royal Society of London A*, 236, 333-380.

[77] Nguyen, H. T. (2016). *On Evidence Measures of Support for Reasoning with Integrated Uncertainty: A Lesson from the Ban of P-values in Statistical Inference. In Integrated Uncertainty in Knowledge Modelling and Decision Making, 3-15.* Springer.

[78] Nguyen, H. T., Sriboonchitta, S. and Thach, N. N. (2019). *On Quantum Probability Calculus for Modeling Economic Decisions. In Structural Changes and Their Econometric Modeling.* Springer.

[79] Nguyen, H. T. and Walke, A. E. (1994)r. *On Decision Making using Belief Functions. In Advances in the Dempster-Shafer Theory of Evidence*, Wiley.

[80] Nosek, B. A., Alter, G., Banks, G. C. and Others (2015). Estimating the Reproducibility of Psychological Science. *Science*, 349, 1422-1425.

[81] Nuzzo, R. (2015). How Scientists Fool Themselves - and How They Can Stop. *Nature*, 526, 182-185.

[82] Oderberg, D. (2008). *Real Essentialism.* Routledge, London.

[83] Pearl, J. (2000). *Causality: Models, Reasoning, and Inference.* Cambridge University Press, Cambridge.

[84] Peng, R. (2015). The Reproducibility Crisis in Science: A Statistical Counterattack. *Significance*, 12, 30-32.

[85] Pocernich, M. (2007). Verfication: Forecast Verification Utilities for R.

[86] Poole, D. (1989). Explanation and Prediction: An Architecture for Default and Abductive Reasoning. *Computational Intelligence*, 5(2), 97-110.

[87] Quine, W. V. (1951). Two Dogmas of Empiricism. *Philosophical Review*, 60, 20-43.

[88] Quine, W. V. (1953). *Two Dogmas of Empiricism.* Harper and Row, Harper Torchbooks, Evanston, Il.

[89] Russell, B. (1920). *Introduction to Mathematical Philosophy.* George Allen & Unwin, London.

[90] Sanders, G. (2018). *An Aristotelian Approach to Quantum Mechanics.*

[91] Schervish, M. (1989). A General Method for Comparing Probability Assessors. *Annals of Statistics*, 17, 1856-1879.

[92] Shimony, A. (2013). *Bell's Theorem.* The Stanford Encyclopedia of Philosophy.

[93] Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, 25, 289-310.

[94] Stove, D. (1982). *Popper and After: Four Modern Irrationalists*, Pergamon Press, Oxford.

[95] Stove, D. (1983). *The Rationality of Induction.* Clarendon, Oxford.

[96] Trafimow, D. (2009). The Theory of Reasoned Action: A Case Study of Falsification in Psychology. *Theory & Psychology*, 19, 501-518.

[97] Trafimow, D. (2016). The Probability of Simple Versus Complex Causal Models in Casual Analyses. *Behavioral Research*, 49, 739-746.

[98] Trafimow, D. (2017). Implications of an Initial Empirical Victory for the Truth of the Theory and Additional Empirical Victories. *Philosophical Psychology*, 30(4), 411-433.

[99] Trafimow, D. el. al (2018). Manipulating the Alpha Level cannot Cure Significance Testing. *Frontiers in Psychology*, 9, 699.

[100] Trafimow, D. and Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, 37(1), 1-2.

[101] Vigen, T. (2018). *Spurious Correlations.*

[102] Wasserstein, R. L. (2016). The ASA's Statement on P-values: Context, Process, and Purpose. *American Statistician*, 70, 129-132.

[103] Wilks, D. S. (2006). *Statistical Methods in the Atmospheric Sciences.* Academic Press, New York.

[104] Williams, D. (1947). *The Ground of Induction.* Russell & Russell, New York.

[105] Woodfield, A. (1976). *Teleology.* Cambridge University Press, Cambridge.

[106] Ziliak S. T. and McCloskey, D. N. (2008). *The Cult of Statistical Significance.* University of Michigan Press, Ann Arbor.