Studies in Computational Intelligence 808

Vladik Kreinovich
Songsak Sriboonchitta *Editors*

Structural Changes and their Econometric Modeling

Springer

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

Songsak Sriboonchitta
Faculty of Economics
Chiang Mai University
Chiang Mai, Thailand

ISSN 1860-949X ISSN 1860-9503 (electronic)
Studies in Computational Intelligence
ISBN 978-3-030-04262-2 ISBN 978-3-030-04263-9 (eBook)
https://doi.org/10.1007/978-3-030-04263-9

Library of Congress Control Number: 2018960914

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

The Replacement for Hypothesis Testing

William M. Briggs[1(✉)], Hung T. Nguyen[2,3], and David Trafimow[2]

[1] New York, USA
matt@wmbriggs.com
[2] New Mexico State University, Las Cruces, USA
{hunguyen,dtrafimo}@nmsu.edu
[3] Chiang Mai University, Chiang Mai, Thailand

Abstract. Classical hypothesis testing, whether with p-values or Bayes factors, leads to over-certainty, and produces the false idea that causes have been identified via statistical methods. The limitations and abuses of in particular p-values are so well known and by now so egregious, that a new method is badly in need. We propose returning to an old idea, making direct predictions by models of observables, assessing the value of evidence by the change in predictive ability, and then verifying the predictions against reality. The latter step is badly in need of implementation.

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

1 The Nature of Testing

The plain meaning of *hypothesis testing* is to ascertain whether, or to what degree, certain hypotheses are true or false, or if a theory is good or bad, or useful or not. This is not, of course, what that phrase means in frequentist or Bayesian theory. Classical statistical philosophy has developed measures, such as p-values and Bayes factors, which are not directly related to the plain meaning. Yet the plain meaning is what all seek to know.

The relationship between a theory's truth or goodness and p-values is non-existent by design. The connection between a theory's truth and Bayes factors is more natural, e.g. Mulder and Wagenmakers (2016), but because Bayes factors focus on unobservable parameters, they exaggerate evidence for or against a theory (we demonstrate this presently). The predictive approach outlined below restores, and puts into proper perspective, the natural goals of modeling.

The two main goals of modeling physical observables are prediction and explanation, i.e. understanding the causes of the phenomenon of interest. Without delving too deeply into a highly complex subject, it should be obvious that if we knew the cause or causes of an observable, we would write these down and not need a probability model, see Briggs (2016). Probability models are only needed when causes are unknown, at least in some degree. Though there is some

V. Kreinovich and S. Sriboonchitta (Eds.): TES 2019, SCI 808, pp. 3–17, 2019.
https://doi.org/10.1007/978-3-030-04263-9_1

disagreement on the topic, e.g. Hitchcock (2016), Breiman (2001), and though the reader need not agree here, we suggest that there is no ability for a wholly statistical model to identify cause. Everybody agrees models can, and do, find correlations. And because correlations are not causes, hypothesis testing cannot find causes, nor does it claim to. At best, hypothesis testing highlights possibly interesting relationships.

Now every statistician knows these arguments, and agrees with them to varying extent (most disputes are about the nature of cause, e.g. Pearl (2000)). But the "civilians" who use the tools statisticians develop have not well assimilated the arcane philosophy behind those tools. Civilians all too often assume that if a hypothesis test has been "passed", a causal effect—or something very like it, like a "link" (a word nowhere defined)—has been confirmed. This is only natural given the name: hypothesis *test*. This explains the overarching desire for p-value hacking and the like. The result is massive over-certainty and a reproducibility crisis, e.g. see among many others Begley and Ioannidis (2015); see too Nosek et al. (2015).

This leaves prediction. Prediction makes sense and is understandable to everybody, and best of all opens all models to verification, to real testing. A hard check against reality is not the usual treatment statistical models receive. This is a shame. The many benefits of prediction are detailed below.

There is not much point here adding to the critiques of p-values. Not every argument against them is well known, but enough are in common circulation that even their most resolute defenders are given pause, e.g. Nguyen (2016), Trafimow and Marks (2015). The only good use for p-values is the one for which they are designed. Calculating the probability that certain functions of data will exceed some value supposing a specified probability model holds. About whether that, or *any* other, model is good, true, or useful, the *p-value is utterly silent*. It's funny, then, that the only uses to which p-values are put are on questions they can't answer.

The majority—which includes all users of statistical models, not just careful academics—treat p-values like ritual, e.g. Gigerenzer (2004). If the p-value is less than the magic number, a theory has been proved. It does not matter that frequentist statistical theory insists that this is not so. It is what everybody believes. And the belief is impossible to eradicate. For that reason alone, it's time to retire p-values.

As stated, Bayes factors come closer to the mark, but since they are stated in terms of unobservable parameters, their use will always lead to over-certainty. This is because we are always more certain of the value of parameters than we are of observables. This is obvious since the posterior of any parameters feeds into the equations for the predictive posterior of observables. Take an easy example. Suppose we characterize the uncertainty in the observable y using a normal with known parameters. Obviously, we are more uncertain of the observable than the parameters, which are known with certainty. If we then suppose there is uncertainty in the parameters (perhaps supplied by a posterior, or by guess), we have to integrate out this new uncertainty in the parameters, which increases the

uncertainty in the observable. For these reasons, we do not comment further on Bayes factors, though we do use what is usually considered an objective Bayes framework, suitably understood, to produce predictions. Frequentist probability predictions can also be used, but with difficulties in interpretation.

We take probability to be everywhere conditional, and nowhere causal, in the same manner as Briggs (2016), Franklin (2001), Jaynes (2003), Keynes (2004). Accepting this is not strictly necessary for understanding the predictive position, but it is for a complete philosophical explanation. This philosophy's emphasis on observables and measurable values which inform observables is also important.

2 Predictive Assessment

All quantifiable probability models for observables y can fit this predictive schema:

$$\Pr(y \in s|\mathrm{X},\ \mathrm{D},\ \mathrm{M}) \tag{1}$$

where y is the observable of interest (the dimension can be read from the context), s a subset of concern, M is the evidence and premises that suggest the model form, D is optionally old (or assumed) measurements of (y, x) and X optionally represents new or assumed values of x. It is well to stress that probability, like logic, does not restrict itself to statements on observable propositions. But scientific models do revolve around that which can be measured. Thus, the only type of models we discuss here will be for observable, i.e. measurable, y.

It is also worth emphasizing M is usually a complex, compound proposition that includes *everything* used to judge the model. Statisticians have developed a shorthand that works well with mathematical manipulations of models, but which masks important model information. Since nearly all models in practical use are assigned *ad hoc*, the masking emboldens the false belief the model used in an application is *the* correct model, or at least one "close enough" to the true one. This over-emphasizes the importance of hypothesis testing, leading to over-certainty that causal, or semi-causal "links", have been properly identified. And this in turn has led to a most unfortunate *non*-practice of model verification. It is rare to never that the vast army of published models ever undergo testing against the real world. About that subject, more below.

The majority of probability models follow one of two basic forms. Paradigmatic examples:

$\mathrm{M}_D =$ "A 6-sided object with sides labeled 1–6, which will be tossed, after which one side must show". The observable y is the side, with $s = 1 \cdots 6$. Then $\Pr(y = i|\mathrm{M})1/6, \forall i$. About why this deduction holds, and about why we believe we can deduce probability and why we do not believe probability is subjective, we relegate to Briggs (2016).

$\mathrm{M}_{temp} =$ "The uncertainty of tomorrow's high temperature quantified by a normal distribution, whose central parameter μ is a function of yesterday's high and an indicator of precipitation"; i.e. a standard regression.

M_D has no parameters and requires no old observations. Its general form is $\mathrm{M}_P = \mathrm{P}_1\mathrm{P}_2 \cdots \mathrm{P}_m$, where each P is a premise as in a logical argument, and the

model itself is a conjunction of these premises. Each of the P may be arbitrarily complex.

M_{temp} is a parameterized model typically requiring old observations, and in Bayesian analysis evidence on the uncertainty of the parameters, i.e. prior distributions. The evidence suggesting the priors is assumed to be part of M_{temp}. Of course, there may, and even must, be some number of premises P included in parameterized models. The one that must be present is the one identifying the parameterized model. E.g. P = "Uncertainty in the observable will be quantified with a normal distribution". This P is almost always *ad hoc*. This does not mean *not useful*.

Classical hypothesis testing in frequentist or Bayesian terms is usually applied to parametric models, with the goal of model selection, a potentially confusing term, as we shall see. The general idea is simple. In its most basic form, two models are proposed, parameterized or not, both identical except one will have one less premise or parameter. For example:

$$\mathrm{M}_{P_a} : \mathrm{P}_1\mathrm{P}_2 \cdots \mathrm{P}_{m-1}\mathrm{P}_m \tag{2}$$
$$\mathrm{M}_{P_b} : \mathrm{P}_1\mathrm{P}_2 \cdots \mathrm{P}_{m-1} \tag{3}$$

$$\mathrm{M}_{\theta_1} : \mu = \theta_0 + \theta_1 x_1 + \theta_2 \mathrm{I}(x_2) \tag{4}$$
$$\mathrm{M}_{\theta_2} : \mu = \theta_0 + \theta_1 x_1 \tag{5}$$

where in the first set of comparisons M_{P_b} has one fewer premise than does M_{P_a}. In the second set of comparison x_1 might be, from the example above, yesterday's high temperature, and $\mathrm{I}(x_2)$ the indicator of precipitation. The ordering of more to less complex models does not, of course, matter.

Predictive selection for premise-based models is simplicity itself. But don't let its simplicity fool you. It contains the very basis of how models are actually built. Calculate

$$\Pr(y \in s|\mathrm{X},\, \mathrm{D},\, \mathrm{M}_{P_a}) = p + \delta \tag{6}$$
$$\Pr(y \in s|\mathrm{X},\, \mathrm{D},\, \mathrm{M}_{P_b}) = p \tag{7}$$

Using the nomenclature of Keynes, premise P_m is *relevant* to y at s if $\delta \neq 0$ (the obvious restrictions on the values of p and δ apply); otherwise it is relevant. Using the example above with M_D remaining the same, and letting $\mathrm{M}_{D+1} = \mathrm{M}_D$ & "Candy canes have peppermint flavoring." Then

$$\Pr(y \in s|\mathrm{M}_{D+1}) = 1/6 + 0 = 1/6, \forall s \tag{8}$$
$$\Pr(y \in s|\mathrm{M}_D) = 1/6, \forall s. \tag{9}$$

Obviously, the flavoring of candy canes is irrelevant to knowing which side of a die will show. At no value of s was δ non-zero. The premise is therefore rejected.

The example is silly, but it highlights an important truth. *All* models are built like this. Scores of irrelevant premises are rejected at the outset, with little or no

thought. This is the right thing to do, too. Yet it is the reason the premises are rejected that is important. Model builders reject premises because they know the probability of the observable y at some measurable x *will not change*. If you like, we can say that the hypothesis that the premise is relevant has been rejected—and rejected absolutely.

Hypothesis testing, then, begins well before any p-value is calculated or even data collected. It does not reach any level of formality until well down the road. This is interesting because if people were truly serious about the theory behind p-values, to remain consistent with that theory, p-values (and Bayes factors) should be used to rule out *every* hypothesis not making it into the final model. Now *every* is a lot; indeed, it is infinite. Since any hypothesis not making it into the final model must be rejected in the formal way, true p-value and Bayes factor believers would thus never finish testing. No model would ever get built in finite time.

What we are proposing is an approach which is everywhere consistent. And which produces no paradoxes.

In the case of comparing parameterized probability models, there is uncertainty in which model is "better". But there is no uncertainty in calling any model *true*, if that word is meant in the causal sense. None but the strictly causal (perfectly predictive) model is true. If we knew the actual cause of y, or what determines the value of y, then we would not need a probability model. Causal models are *not* impossible, or even rare. Physics is awash in causal and deterministic models (to know the cause is greater than to know what determines a value).

Most, or even all, statistical models are *ad hoc*. In the temperature example, it is obvious many other parameterized, and even unparameterized, models could have been used to express uncertainty in y. Not just in the sense that extra terms could be added to the right hand side of the regression, but entirely different model structures. Normal distributions do not have to be used, for instance. The model need not be linear in the parameters. The possibilities for *ad hoc* models are limitless.

That is what makes talk of "true" values of the parameters curious. Since statistical models are *ad hoc* and not true in any causal sense, and since nearly all models do not specify the precise and total circumstance of an observable (i.e. all auxiliary premises, see Trafimow (2017)), it is vain to search for "true" values of parameters. Even at a hypothetical, never-will-be-reached limit. Again, physics comes closest to an apt understanding of true values of parameters, because there carefully controlled experiments can be run that delineate all the (known) possible causal factors. In these limited circumstances, it makes more sense to speak of true parameter values. Parameters in this sense often have physical meaning, at least by proxy. But, again, this does not hold for the vast majority of probability models.

Predictive selection for parametric models is as easy as above. Calculate

$$\Pr(y \in s|\mathrm{X},\, \mathrm{D},\, \mathrm{M}_1) = p + \delta \tag{10}$$

$$\Pr(y \in s|\mathrm{X},\, \mathrm{D},\, \mathrm{M}_2) = p \tag{11}$$

Assume M_1 is the model with the greater number of parameters. Again, we assume the obvious numerical restrictions of p and δ. If at s, and given X and D, $\delta = 0$, the parameter(s) in M_1, and therefore the measurements associated with those parameters, are *irrelevant* to the uncertainty of y. These X, and these parameters, are therefore not needed in the model. Removing them does not change the probability. The models in (10) are predictive, meaning the uncertainty in the parameters given by priors is integrated out. Yet even frequentists can use this method, as long as probability predictions can be made from the frequentist model.

If at any s, for the given X and D, $\delta \neq 0$, then the X and its parameters are *relevant*. Whether to keep the extra parameters becomes a standard problem in decision analysis. A relevant parameter important to one decision maker can be unimportant to another. There can be no universal value of δ useful in all situations, like there is with the magic number for p-values. As should be clear, relevance depends on s and on everything on the right hand side of the probability equation. That means any change on the right hand side might change the measure of relevance. That accords with common sense: change your information, change your basis of judgment.

In practice, on a per-model, per-decision basis, a δ is chosen, which may depend on s, below which measurements are decided to be unimportant, and above which are important. Measurements, and their associated parameters, are kept or discarded accordingly.

An additional advantage of this approach is that no parameter estimates are needed, or even desired. Parameters are not in any case observable. The models are already *ad hoc* anyway, so focusing on parameter estimates, either as a Bayesian posterior or a frequentist point estimate with confidence interval, produces over-certainty in any X's importance. The predictive approach thus unifies testing and point estimation.

Not only can (10) be used in intra-model selection, but it is ripe for estimating the probabilistic importance of each X. It will often be found that a model with multiple parameters will show a wee p-value and large (relative) point estimate for one parameter, and a non-publishable p-value and small point estimate for the second parameter. But when (10) is employed, the order of importance is inverted. Changing the value of X for the classically "weaker" parameter will produce larger variations in probability of $y \in s$, especially for values of s thought crucial in the problem at hand.

3 Examples

3.1 Example 1: Product Placement Recall

We begin for the sake of clarity with the simplest of examples. Results of a survey to relate ability to recall product placement in theater films by movie genre (Action, Comedy, Drama) and sex were asked on 137 people, each giving a response (a score) with the number of correct recalls in the discrete interval 0–6, Park and Berger (2010). The data were initially analyzed using null hypothesis significance testing. The conclusion of the authors was "Results suggest that brand recognition is more common in drama films."

An ordinary regression in R on the score by sex (M = 1, or 0) and movie genre was run, producing the following ANOVA table (*sans* hyperventilating asterisks).

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.4994     0.1930  18.134   <2e-16
M1            0.3952     0.2489   1.588   0.1147
GenreComedy   0.4087     0.2712   1.507   0.1342
GenreDrama    0.7077     0.2792   2.535   0.0124
```

The p-values for sex (difference) and Comedy were larger than the magic number. Some authors would at this point remove sex from the model. The p-value for Drama was publishable, hence the conclusion of the authors.

Predictive probabilities of the full model were calculated, assuming standard out-of-the-box "flat" priors. Posteriors on the parameters were first calculated, then these were integrated out to produce the predictive posterior of the observable score, see Bernardo and Smith (2000). The results would, of course, change with a different prior; but so would they change with a different model. We are *not* recommending this model, and certainly not recommending flat priors; we are only showing how the predictive approach works in a common situation.

There is a bit of difficulty in creating predictive probabilities, because the scores can only take the values 0–6, but the standard normal regression model produces predictive probability densities along the continuum. Indeed, the model produces predictions of positive probabilities for values of scores less than 0 and greater than 6, scores that will never be seen (they are impossible) in any repeat of the experiment. We elsewhere call the assignment of positive probability to impossible events *probability leakage*, Briggs (2013). It usually shows up when regression models do not make good approximations and when the observable lives in a limited range, or when the observable's discreteness is stark.

In this case, for males, the predictive probabilities for scores greater than 6 are 0.06 for Action, 0.1 for Comedy, and 0.15 for Drama (these are probabilities for known impossible values). In other words, given the person is a male assessing a Drama, the model predicts a probability of 0.15 for new scores greater than 6. For females, the numbers are 0.03, 0.06, and 0.09 respectively. Not small numbers. For scores less than 0, the predictive probabilities are for men are all

less than 0.001; for women the largest is 0.003. Whether any of these numbers is important depends on the decisions to which the model is put, and not on whether any statistician thinks them small or large. About these decisions, we are here agnostic.

The next decision is how to turn the predictions which are over the real line to predictions of discrete observable scores. One way of doing this, which is not unique, is to calculate the predictive probability for being between 0 and 0.5, and assign that to a predictive probability of score = 0; next calculate the predictive probability for being between 0.5 and 1, and assign that to a predictive probability of score = 1; and so on. The probability of 5.5 to 6.5 can be assigned to score = 6, with the remainder being left to leakage, or everything greater than 5.5 can be assigned to score = 6; correspondingly, everything less than 0.5 can be assigned score = 0. Now all this rigmarole would not have been necessary if a model which only allowed scores 0–6 were used (perhaps a multinomial regression). But our purpose here is not to find terrific or apt models; we only want to explain how to use the predictive approach for models people routinely use.

It is crucial to understand that in creating predictive probabilities, as in Eq. (6), the model must be fully specified in each prediction. In other words, we created a model of sex and genre because we thought these measurements would change the uncertainty in the score, therefore for *each and every* prediction we make, we must specify a value of sex and genre.

Figure 1 shows the predictive probability for men for each genre. Clearly, the difference in these probabilities are non-zero, hence $\delta \neq 0$; so, genre is relevant to uncertainty in score. The differences in probabilities clearly depends on the level of score (the s), ranging from about 0.001 (in absolute values) for $s = 1$, up to 0.14 for $s = 6$. Again, whether these differences are important depends on the decisions to which the model will be put. Supposing for the sake of argument a $\delta = 0.05$ (a familiar number!) to indicate importance, then there is no important differences in probabilities between Action and Comedy for scores of 0–2 and 4–5 but there are for scores of 3 and 6. The p-value would lead to the decision of no difference between Action and Comedy. But with our chosen δ, there is a clear difference in importance.

Now the same plot (or calculations: visual inspection is not necessary) should be done for females by genre, and the differences assessed there too. We skip that step, noting that the important differences exist here, too, and for different scores for the genres. We instead show Fig. 2, the differences in sex at the Drama genre. The differences (in absolute value) are between 0.002 and 0.08. The importance δ is exceed at scores of 3 and 6.

Again, the p-value for sex was not wee, and sex might have been dropped from the model. The important differences noted for Drama were also found for Comedy, but not for Action, though these were not noted by the p-values.

This level of detail in an analysis won't always be needed. Instead, tables like the following can *and should* be presented. Plots and summaries may of course be better, depending on the situation. Here there are two *different* regression

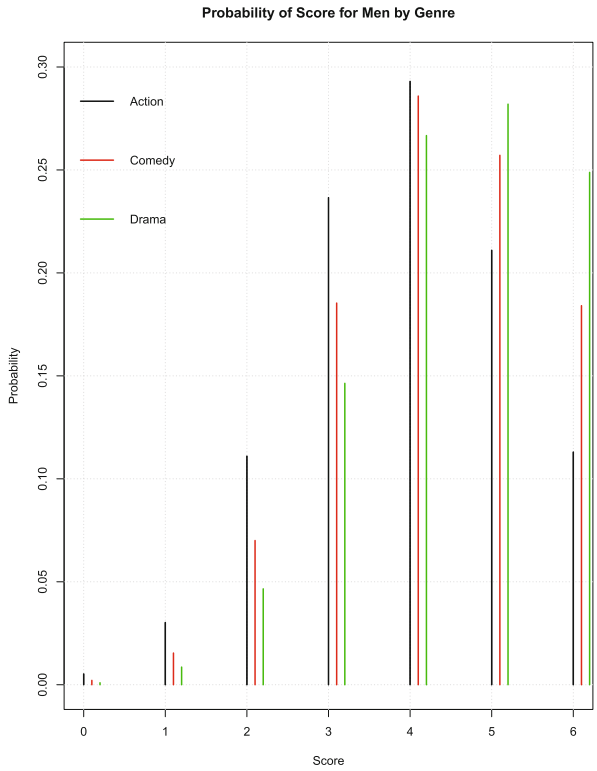

Fig. 1. The days on which the interested events occur for DTAC

Table 1. Probabilities (rounded to nearest hundredth) for scores 0–6 for the genre Drama, with and without considering sex, in two separate regression models.

Sex	$s = 0$	$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$	$s = 6$
Either	0.00	0.01	0.06	0.17	0.28	0.27	0.20
Male	0.00	0.01	0.05	0.15	0.27	0.28	0.25
Female	0.00	0.02	0.08	0.20	0.29	0.25	0.16

models, the first without sex and the second with. Readers are free to make decisions based on their own δs, which might differ from the authors'.

3.2 Example 2: Professor's Salaries

This next example shows the flexibility of the predictive method, and its potential for partial automation. Full automation of analysis is not recommended for any model, except in special circumstances. Automation can cause one to forget limitations.

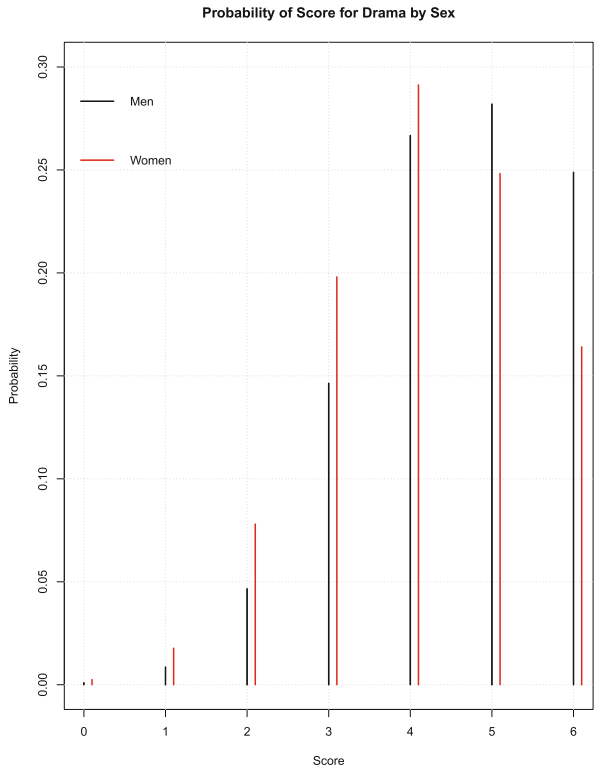

Fig. 2. Predictive probability of score for men and women for the Drama genre.

Nine-month salaries for 2008–2009 were collected on 397 academics at various ranks for a college in the USA for two departments A and B "roughly corresponding to theoretical disciplines and applied disciplines, respectively", quoted from Fox and Weisberg (2011). Faculty sex, years since PhD and years of service were also measured. The minimum measured salary was $57,800 and the maximum was $231,5000, proving at least one of us is in the wrong job.

Obviously, we use this data to make predictions of people not in this data set, because we already know all we can about the salaries of people we have already measured.

That is, we desire naturally to make predictions.

Here is the ordinary ANOVA table:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    78.8628     4.9903  15.803  < 2e-16
rankAsstProf  -12.9076     4.1453  -3.114  0.00198
rankProf       32.1584     3.5406   9.083  < 2e-16
disciplineB    14.4176     2.3429   6.154 1.88e-09
```

```
yrs.since.phd   0.5351     0.2410   2.220  0.02698
yrs.service    -0.4895     0.2119  -2.310  0.02143
sexMale         4.7835     3.8587   1.240  0.21584
```

The standard ANOVA tells us little about predictions. That is easily remedied in Table 2, which we label the predictive "ANOVA" table. It uses the same regression model with (again) "flat" out-of-the-box priors. It shows the central (most likely) estimate for the condition noted, and which holds all other measurements fixed at their observed median values or base levels (to be defined below). The categorical variables are stepped through their levels, while the others step through the first, second, and third observed quartiles. Any other values of special interest may of course be substituted, but we leave these to demonstrate how an automatic analysis might look.

Table 2. Predictive "ANOVA" table for salaries.

Variable	Level	Central Salary ($1,000s)	Pr(Salary > base level)
rank	AssocProf	101	0.5
rank	AsstProf	88.6	0.343
rank	Prof	134	0.844
discipline	A	119	0.5
discipline	B	134	0.675
yrs.since.phd	5	125	0.5
yrs.since.phd	21	134	0.606
yrs.since.phd	40	144	0.719
yrs.service	3	140	0.5
yrs.service	16	134	0.421
yrs.service	37	123	0.302
sex	Female	129	0.5
sex	Male	134	0.56

This Table also shows the predicted probability that a person holding these attributes would have a higher salary than a "base level" person. The base level is not unique and can be user specified as a particular level of interest. Here we take the first level of all other categorical measures as ordered (alphabetically) by R. The first level of rank is "AssocProf", with "AsstProf" coming after, alphabetically. The non-categorical measures take as base their observed first quartile values.

For example, the predicted most likely salary for an Associate Professor in discipline B (the median), and male (also median), with 21 years since PhD and 16 years of service is $101 thousand. The probability another person at the base level, which in this case is a person with the *same attributes*, is, as expected, 0.5 (in this model, the posterior predictive distributions are all symmetric around

the central value). We next hold all these attributes constant, but change the rank, so that we now have a new male Assistant professor in discipline B with 21 years since PhD and 16 year service. The probability this new man has a higher salary is 0.34, meaning, of course, a man with the higher rank has a probability of 0.66 of having a higher salary.

These tables take only a little getting used to, and they are easily modified, as a standard ANOVA is not, for questions interesting to decision makers. Relevance can be picked off the table: any probability differing from 0.5 shows relevance, at least for the levels specified. Direct information about the observable is also prominent.

This table does not obviate a fuller analysis, as was done above in the first example. Plots and tables of the same sort can and should be made. For example, as in Fig. 3.

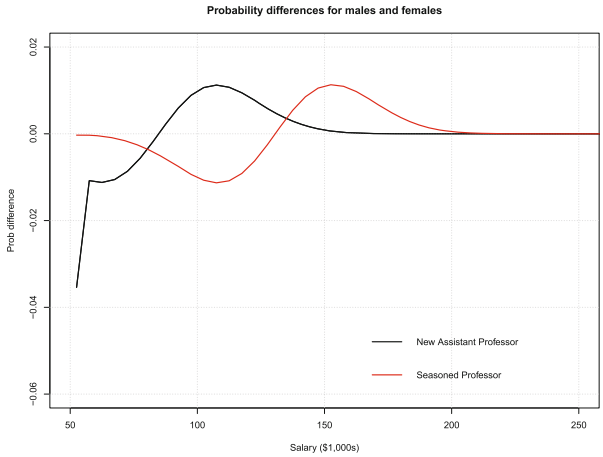

Fig. 3. Predictive probability differences between men and women in discipline B for new Assistant Professors in black (0 years of service, 1 year since PhD) and for seasoned Professors in red (24 years of service, 25 years since PhD). Probabilities are calculated every $5,000.

This shows the predictive probability differences between men and women in discipline B for new Assistant Professors in black (0 years of service, 1 year since PhD) and for seasoned Professors in red (24 years of service, 25 years since PhD). Probability differences are calculated every $5,000. Most of these differences are 0.01, or less. The largest difference was for new hires at a salary lower than was observed. This implication is that while there were observed differences in salaries between men and women, the chances are not great for seeing them persist in new data. At least, not for individual salaries. Calculating the differences over larger "block" sizes of salaries, say, every $10 or $15 thousand would show larger differences.

4 The Conclusion Lies in Verification

The predictive approach does not solve all modeling ills. No approach will. It reduces some, but only some, of the excesses in classical hypothesis testing. Although we advise against a universal, one-size-fits-all value of δ, all experience shows such a value *will* be picked. Doing so makes model selection and presentation automatic. People prefer less work to more. The predictive approach clearly entails more work than standard hypothesis testing in every aspect. As such, there will be reluctance to use it. It also does not provide answers that are as sharply defined as hypothesis testing. And people crave certainty—even when this certainty is exaggerated, as it is with classical hypothesis testing. Every statistician knows how easy it is to "prove" things with p-values.

Any approach that does not add model verification to model selection is doomed to failure. Models *must* be tested against reality. It is not at all clear how to do this with classical hypothesis testing. As said above, the idea a "test" has been passed gives the false impression the model has been checked against reality and found good.

True verification is natural using the predictive approach. Models under the predictive approach are reported in probability form. Advanced training in statistical methods are not needed to understand results. The models reported in Table 1 require no special expertise to comprehend. These are the (conditional) probabilities of *new* scores that might be observed, perhaps depending on the sex of the participant. "Bets" (i.e. decisions) can be made using this table. Here the standard apparatus of decision analysis comes into play in choosing which probabilities are important, and which not. If the model is a good one, the probabilities will be well calibrated and sharp, when considered with respect to whatever bets or decisions that are made with it.

Anybody can check a predictive model (given they can recreate the original scenarios). The original data is not needed, nor the computer code used to generate it. The model is laid bare for all to see and test. Limitations and strengths, especially for controversial and "novel" research, will quickly become apparent.

How best to do verification we leave to outside authorities. This list is far from complete, but a good place to start is here: e.g. Gneiting and Raftery (2007), Briggs and Zaretzki (2008), Hersbach (2000), Wilks (2006), Briggs and Ruppert (2005), Briggs (2016), Gneiting et al. (2007). The idea is basic. Produce predictions and compare these using proper scores against observations never used or seen in any way before. This is the exact way civil engineers test models of bridges, or electrical engineers test models of cell phone capacity, etc. The "never used" is strict, and thus excludes cross validation and other approaches which reuse or "peek" at verification datasets when building a model. It's not that these methods don't have good uses, but that the will always inflate certainty in the actual value of a model.

Verification, like model building is not exact, and cannot be. We must guard against the idea that if a theory has passed whatever test we devise, we have the best or a unique theory. Verification is not proof. Quine and Duhem long ago showed theories or models besides the one under consideration and testing could

equally well explain any set of observed (contingent) data, Quine (1953), Duhem (1954). And when testing, the auxiliary assumptions (all implicit premises) of a model can be difficult or impossible to disentangle; see Trafimow (2009), Trafimow (2017) for a discussion. What can be said is that given past good performance of a model, and taking care the conditions in all explicit and implicit premises are also met, it is likely the model will continue to perform well.

References

Begley, C.G., Ioannidis, J.P.: Reproducibility in science: Improving the standard for basic and preclinical research. Circ. Res. **116**, 116–126 (2015)
Bernardo, J.M., Smith, A.F.M.: Bayesian Theory. Wiley, New York (2000)
Breiman, L.: Statistical modeling: the two cultures. Stat. Sci. **16**(3), 199–215 (2001)
Briggs, W.M.: On probability leakage. arxiv.org/abs/12013611 (2013)
Briggs, W.M.: Uncertainty: The Soul of Probability, Modeling & Statistics. Springer, New York (2016)
Briggs, W.M., Ruppert, D.: Assessing the skill of yes/no predictions. Biometrics **61**(3), 799–807 (2005)
Briggs, W.M., Zaretzki, R.A.: The skill plot: a graphical technique for evaluating continuous diagnostic tests. Biometrics **64**, 250–263 (2008). (With discussion)
Duhem, P.: The Aim and Structure of Physical Theory. Princeton University Press, Princeton (1954)
Fox, J., Weisberg, S.: An R Companion to Applied Regression, 2nd edn. SAGE Publications, Thousand Oaks (2011)
Franklin, J.: Resurrecting logical probability. Erkenntnis **55**, 277–305 (2001)
Gigerenzer, G.: Mindless statistics. J. Socio Econ. **33**, 587–606 (2004)
Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. JASA **102**, 359–378 (2007)
Gneiting, T., Raftery, A.E., Balabdaoui, F.: Probabilistic forecasts, calibration and sharpness. J. R. Stat. Soc. Ser. B Stat. Methodol. **69**, 243–268 (2007)
Hersbach, H.: Decompostion of the continuous ranked probability score for ensemble prediction systems. Weather Forecast. **15**, 559–570 (2000)
Hitchcock, C.: Probabilistic causation. In: The Stanford Encyclopedia of Philosophy (Winter 2016 Edition) (2016). https://plato.stanford.edu/archives/win2016/entries/causation--probabilistic
Jaynes, E.T.: Probability Theory: The Logic of Science. Cambridge University Press, Cambridge (2003)
Keynes, J.M.: A Treatise on Probability. Dover Phoenix Editions, Mineola (2004)
Mulder, J., Wagenmakers, E.J.: Editor's introduction to the special issue: Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments. J. Math. Psychol. **72**, 1–5 (2016)
Nguyen, H.T.: On evidence measures of support for reasoning with integrated uncertainty: a lesson from the ban of p-values in statistical inference. In: Huynh, V.N., Inuiguchi, M., Le, B., Le, B., Denoeux, T. (eds.) Integrated Uncertainty in Knowledge Modelling and Decision Making, pp. 3–15. Springer, Cham (2016)
Nosek, B.A., Alter, G., Banks, G.C., et al.: Estimating the reproducibility of psychological science. Science **349**, 1422–1425 (2015)
Park, D.J., Berger, B.K.: Brand placement in movies: the effect of film genre on viewer recognition. J. Promot. Manag. **22**, 428–444 (2010)

Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge (2000)
Quine, W.V.: Two Dogmas of Empiricism. Harper and Row, Harper Torchbooks, Evanston (1953)
Trafimow, D.: The theory of reasoned action: a case study of falsification in psychology. Theory Psychol. **19**, 501–518 (2009)
Trafimow, D.: Implications of an initial empirical victory for the truth of the theory and additional empirical victories. Philos. Psychol. **30**(4), 411–433 (2017)
Trafimow, D., Marks, M.: Editorial. Basic Appl. Soc. Psychol. **37**(1), 1–2 (2015)
Wilks, D.S.: Statistical Methods in the Atmospheric Sciences, 2nd edn. Academic Press, New York (2006)