Don't Test, Decide

William M. Briggs
E-mail: matt@wmbriggs.com*

Abstract There is no reason to use traditional hypothesis testing. If one's goal is to assess past performance of a model, then a simple measure or performance with no uncertainty will do. If one's goal is to ascertain cause, then because probability models can't identify cause, testing does not help. If one's goal is to decide in the face of uncertainty, then testing does not help. The last goal is to quantify uncertainty in predictions; no testing is needed and is instead unhelpful. Examples in model selection are given. Use predictive, not parametric, analysis.

Keywords: Causation, P-values, Hypothesis Testing, Model Selection, Model Validation, Predictive Probability

1 Testing Is Dead

A vast amount of statistical practice is devoted to "testing", which is the art of deciding, regardless of consequence, that a parameter or parameters of a probability model does or does not take a certain value. Most testing is done with p-values. If the p-value is less than the magic number, the test has been passed; otherwise not. Everybody knows the magic number.

The foremost, and insurmountable, difficulty is that every use of a p-value contains a fallacy or is a mistake. This is proved several times over in [1, 2, 3, 4]. Some have accepted that p-values should be abandoned, but, still desiring to test, they have moved to Bayes factors. See [5] about Bayes factors.

Bayes factors are at best only a modest improvement on model decision making. Unfortunately, their use retains several fundamental misconceptions shared by p-values about what parameters are and do. Many for instance believe that if a test has been passed, a cause has been proved. For example, suppose we have a regression that characterizes uncertainty in "amount of improvement" for some observable

*Corresponding author

(say, in medicine). One parameter in the model represents presence of a new drug. If this parameter is positive and the model's parameter passes its test, frequentist or Bayes, almost everybody believes the new drug "works" *in a causal sense.* Meaning everybody thinks, regardless of any provisos they might issue, that the new drug really does *cause* an improvement above the old one.

This inference is certainly natural, just as it is certainly invalid. It is invalid even in those times in which the parameter really is assoociated with a causal effect. A conclusion in some logical arguments can be true even though the argument itself is invalid. Probability models can't prove cause, though they can assume it. See [6].

In any case, testing is not is what is wanted by the majority of users of statistical models. What they're after is quantifying uncertainty of *this* given *that*. In our example, patients wants to know what are the chances they will improve if they take the new drug. And that's not all. They will also likely want to know the cost, possibility of side effects, and so forth. Patients factor in all these things, along with the model's probability of improvement, into a decision whether to use the new drug or to go with the old, a decision which is personal to them—and *not* to the statistician creating the model. Patients likely won't do this "analysis" in a formal way, but what they *are* doing is deciding and not testing.

Testing is definite. Too definite. Testing under frequentism says a thing is certain—a parameter "is" equal to 0, say—with no measure of uncertainty attached. There are, in frequentist inference, measures of uncertainty of tests, but these are statements about tests different than the one in hand, about what happens to other tests at the limit. The limit is a time of no concern to anybody.

Here we argue that testing should be abandoned and replaced everywhere with deciding.

2 Model Uncertainty

The simplest case is one model. We have background evidence M which is known or accepted such that the uncertainty in the observable y is characterized by the probability model deduced from M. This is

$$\Pr(y|x\mathrm{DM}), \tag{1}$$

where the optional x are other measures probative of y, the optional D is n prior observations (y,x), and M the background information. In M are *all* relevant premises of the model, including premises about priors on the parameters of the model, if any. Finally, y is shorthand for $y \in s$ where s is some relevant set.

Suppose $\Pr(y|x\mathrm{DM}) = p$, then it is *true* the probability of y conditional on this model is p. That is, the probability has been deduced conditional on M. Whether M is true, or useful, is another matter entirely.

Next imagine two models are under contention, M_1 and M_2. Why there might be two, or more, is explored later. That there are two requires having some form of

background knowledge B, from which these two probability models are deduced. B may be as simple as "I am considering M_1 and M_2."

In B is the implicit or explicit evidence from which we deduce the prior probability of the models. Explicit evidence might exist. It would have the form $Pr(\mathrm{M}_1|\mathrm{B}) = q$, from which $Pr(\mathrm{M}_2|\mathrm{B}) = 1-q$. If no explicit evidence exists, and only the implicit evidence that these two (and no other) models are possible, then via the statistical syllogism we deduce $q = 1/2$.

Then, as is well known,

$$\Pr(y|x\mathrm{DB}) = \Pr(y|x\mathrm{DBM}_1)\Pr(\mathrm{M}_1|x\mathrm{DB}) + \Pr(y|x\mathrm{DBM}_2)\Pr(\mathrm{M}_2|x\mathrm{DB}). \quad (2)$$

This is expanded in the obvious way to $p > 2$ models. This humble formation is the solution to all old testing problems.

It is important to understand the relationship between any of the models. The first assumption is that any M_i is probative to $y \in s$ for at least some s. Leave aside difficult questions of infinite sets and so-called "improper" (read *not*) probabilities. Any actual y in use in finite (which is to say, all) settings, will actually take at most a finite number of values (even if y is *potentially* infinite). No measurement apparatus exists to probe any y with infinite precision, and no actual decision uses any but finite y.

Therefore, an implicit model which always exists for any actual y is that $\Pr(y = y_i|x\mathrm{D}) = 1/m$, where $y \in \{y_1, y_2, \ldots, y_m\}$, with $m < \infty$. These values represent the smallest possible actual measurements, or smallest decisionable measurements, of y. These will always actually exist in any real-life decision setting; even in those cases where y potentially belongs in some infinite set. This model says, knowing only that y can take values in a known set, the probability it takes any of these values, absent *any* other information, is uniform. This again uses the statistical syllogism for proof.

Thus, in order for an M_i to be probabtive of y, it must be that

$$\Pr(y \in s|x\mathrm{D}) \neq \Pr(y \in s|x\mathrm{DBM}_i) \quad (3)$$

for at least some $s \in \{y_1, y_2, \ldots, y_m\}$, where s is any collection of y_k. If there is instead equality, then M_i adds nothing to our knowledge of the uncertainty of y and is not therefore a proper model *of* y.

The same holds true in comparing any two models. As long as, given B, no M_i can be deduced from any other M_j, and each M_i is probative of y, then each model adds something to the knowledge of the uncertainty of y. Whether this addition is useful or good or cost efficient are, of course, separate questions. All we require is that each model is separately probative in the sense of (3) and that

$$\Pr(y \in s|x\mathrm{DBM}_i) \neq \Pr(y \in s|x\mathrm{DBM}_j) \quad (4)$$

for at least some s. If there is equality at all s, then $\mathrm{M}_i \equiv \mathrm{M}_j$.

It is crucial to understand that if the probabilities for two models only differ at $s = \{y_k\}$ for a singular k, then each model is still adding different information—conditional on D, B, and x, of course.

Another clarification. Suppose B indicates M_1 is a normal model with a conjugate prior on the parameters, and that M_2 is also a normal model with a Jeffrey's prior on the parameters. These are *two separate models* in our terminology because they will give different probabilities to various *s*. Models are *different* when they are not logically equivalent and when they are probative to *y*—conditional on D and B (and *x*, when present).

It is true one of these models might be better with regard to the predictions, and the *decisions conditional on those predictions*. But if that is not known in advance (in B), then there is no way to know that one should prefer one model over the other.

This finally brings us to the difference between testing and deciding and their relationship to uncertainty.

3 Testing Versus Deciding

Suppose we have the situation just mentioned, two normal models with different priors for the observable *y*. We'll assume these models are probative of *y*; they are obviously logically different, and practically different for small *n*. At large *n* the difference in priors vanishes.

A frequentist would not consider these models, because in frequentist theory all parameters are fixed and onotlogically exist (presumably in some Platonic realm), but a Bayesian might work with these models, and might think to "test" between them. What possible reasons are there to test in this case?

First, what is being tested? It could be which model fits D, the past data, better. But because it is always possible to find a model which fits past data perfectly, this cannot be a general goal. In any case, if this is the goal—perhaps there was a competition—then all we have to do is look to see which model fit better. And then we are done. There is no testing in any statistical sense, other than to say which model fit best. There is *no* uncertainty here: one is better *tout court*.

The second and only other possibility is to pick the model which is most likely to fit future data better.

Fit still needs to be explained. There are many measures of model fit, but only one that counts. This is that which is aligned with a decision the model user is going to make. A model that fits well in the context of one decision might fit poorly in the context of another. Some kind of proper score is therefore needed which mimics the consequences of the decision. This is a function of the probabilistic predictions and the eventual observable. Proper scores are discussed in [1]. It is the user of the model, and not the statistician, who should choose which definition of "fit" fits.

There is a sense that one of these models might do better at fitting, which is to say predicting, future observables. This is the decision problem. One, or one subset of models, perhaps because of cost or other considerations, must be chosen from the set of possible models.

There is also the sense that if one does not know, or know with sufficient assurance, which model is best at predictions, or that decisions among models do not

have to be made, that the full uncertainty across models should be incorporated into decisions.

The two possibilities are handled next.

3.1 Put Your Best Fit Forward

Now it is easy to calculate the so-called posterior of every model; i.e.

$$\Pr(\mathrm{M}_i|\mathrm{DB}) = \frac{\Pr(y|x\mathrm{BM}_i)\Pr(\mathrm{M}_i|x\mathrm{B})}{\sum_i \Pr(y|x\mathrm{BM}_i)\Pr(\mathrm{M}_i|x\mathrm{B})}. \tag{5}$$

Recall D = (y,x), the previous observations.

This can be of interest in the following sense. The background information B supplies the models, M_i. There are only three possibilities for B. (1) B specifies a strictly causative M, i.e. it has identified all the causes of y in conjunction with x. There is then no reason to have any rival models. (2) From B we *deduce* a probability model, as in the case of a die throw, for instance. Again, there is no need of a rival model, for the exact probability of y (possibly given x) has been deduced. (3) B specifies only *ad hoc* models, usually chosen by custom and experimentation. There is no sense then that any M_i specifies is the true model, though one may be best at either fitting D or in predicting future y.

The model posteriors thus represent how well *ad hoc* models fit past data. But since our interest is prediction, we want to know how good each model will make predictions, if we are going to pick just one model (or possibly some subset of models). Thus "best" has to be defined by the decisions to be made with it.

Predicting which model will be best runs like this: (1) Fit each model M_i and form (1), i.e. the predictive form of y. (2) Use $\Pr(y|x\mathrm{DM}_i) = p_i$ to probabilistically predict y. (3) Compute $S(\mathrm{D}(y,x), p_i)$, a proper score reusing the previous data D; S_i may be a vector of scores of length n, or a single number. (4) The probability of seeing scores "like" S_i in new predictions (of size n) is the posterior probability of M_i. This "like" has to be strenghtened considerably in future work.

The probabilities and the costs, or rather consequences, of each S_i are then used in a standard decision setting. For example, it could be in the case of two models $\Pr(\mathrm{M}_1|\mathrm{DB}) > \Pr(\mathrm{M}_2|\mathrm{DB})$ but that

$$\Pr(\mathrm{M}_1|\mathrm{DB})S(\mathrm{D}(y,x), p_1) < \Pr(\mathrm{M}_2|\mathrm{DB})S(\mathrm{D}(y,x), p_2)$$

M_2 is picked if higher S are better (for this decision). Note carefully that the costs of the model, and all other elements related to the decision to be made, are incorporated in S.

3.2 Be Smooth

There is not much to be said about using all uncertainty across all models, except to reiterate that all known or accepted uncertainties should be used in making predictions and in ascertaining model performance. That is, use this:

$$\Pr(y|x\text{DB}) = \sum_i \Pr(y|x\text{DBM}_i) \Pr(\text{M}_i|x\text{DB}). \tag{6}$$

In other words, don't test: use all available information. The differences in the models are of no consequence, or they do not matter to any decisions that will be made. Do not pick the "best" model—use all of them! The weights between models are specified in B, as before. The posterior predictive distribution smooths over all models.

In particular, there is no reason in the world to engage in the usual practice of model selection, if model selection is used to discover causes or "links". Links are causes but said in other words. Modelers speak of links when the modeler knows that probability models cannot discover causes, but still wants to say his measure is a cause; hence, "link". There are innumerable headlines like this: "Eating bananas is *linked to* irascibility." People who want to be irascible then begin eating bananas, and vice versa, believing in the causal power of the "link".

Model selection often happens like this. A modeler is using a regression model (which are ubiquitous), and begins with a selection of measures x_i, for $i = 1 \ldots p$. Now what is often missed about modeling, or rather not emphasized, is that at this point the modeler has already made an infinite number of "tests" by not including every possible x there is. And he did this not using any information from *inside* the model. He did *no* testing, but made decisions. This is a direct acknowldgement that statistical *testing is not needed*.

There must have been some reason for the set of x chosen, some suspicion (which belongs formally in B) that these x are related in the causal chain of y in some way. There is therefore, based on that reason alone, no reason to toss any of these x away. Unless it is for reasons of costs and the like. But that situtation (decision) has already been covered in the previous sub-section.

It may also be that there are too many x, that $p > n$, for instance, and the modeler does not want to use methods where this can be handled. It may be, too, that some of the x were included for dubious reasons. It could be, that is to say, the modeler wants to try $\text{M}_1 = \text{M}(x_1, x_2, \ldots, x_p)$ and $\text{M}_2 = \text{M}(x_1, x_2, \ldots, x_{-j}, \ldots, x_p)$, i.e. a model with and without x_j.

Absent cost and the like there just is no reason to test, i.e. to decide between M_1 and M_2. Let B decide the weighting, the initial q_i, use all uncertainty and don't test. Unless q_i is high for a dubiously entered x_j, the model with it will not get great weight in the posterior prediction calculation. Use both models and compute (6).

Not that I recommend it, but this is the solution to the all-subsets regression "problem". For interest, an example is given below.

4 Giving It A Try

4.1 Deciding

This is a somewhat contrived example, using appendicitis data from [7]. There are $n = 443$ patients on which the presence of appendicitis is measured, along with age, sex, white blood count, and the results from each of ten medical examinations, which are either positive or negative. In reality, these are simple costless tests, such as examining for nausea or right lower quandrant pain. For our purposes we will assume that we can only "afford" one of these tests and have to pick one—perhaps because of monetary or time costs.

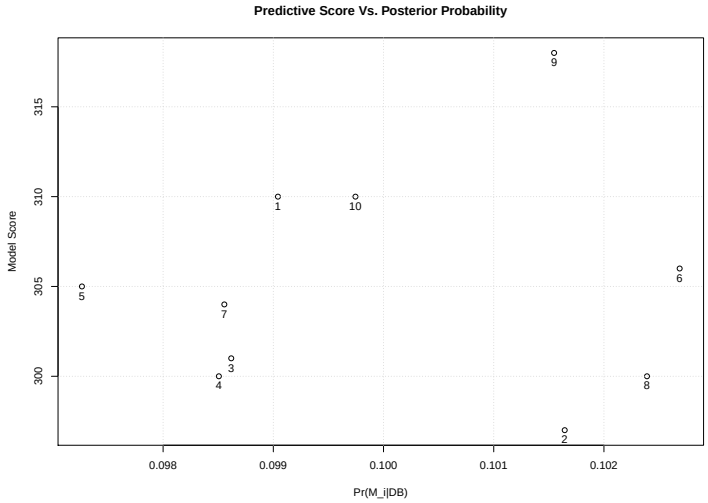

Fig. 1 This is the model posterior, computed by (5), plotted against the score of each model. See the text for model score calculation. The number of the test is also given.

The models are logistic regressions, each with age, sex, and white blood count and one of the ten tests. A uniform distribution over models is used for our B ($q_i = 1/10$). The scoring-decision function is to assume a posterior predictive probability of appendicitis greater than 1/3 indicates the presence of appendicitis, such that an action along those lines is taken, such as an ultrasound or exploratory laparotomy. Predictions which are accurate incur no costs. False positives receive a penality of 1, and false negatives, which are medically more dangerous, receive a penality of 2. All tests are thought to cost the same amount (in time or expensese).

Fig. 1 shows the results. This is the model posterior, computed by (5) by the score of each model. The number of the test is also given. It is clear the model with the highest posterior probability, test 6, does not have the lowest score.

Picking which model to implement requires some form of decision analysis. Here we computed a sum total score, but the distribution of scores could have been used instead. Something along those lines is illustrated in the next section. What's really needed is a *second* modeling of scores. These scores are all in-sample based on the posterior predictive distributions of the observable, and as such they are predictions of what future scores would look like—given an individual model is used.

The lesson is clear: model posterior probabilities do not translate directly into a preference ordering. That depends on the score and decisions to be made. This example, as said, is contrived, but the steps taken would be exactly the same in real situations.

4.2 Full Uncertainty

Data made available from [8] is used to predict levels of prostate specific antigen (PSA) > 10, which are levels commonly taken to indicate prostate cancer or prostate difficulties. Available as probative measures are log cancer volume, log prostate weight, age, log of benign prostatic hyperplasia levels, seminal vesicle invasions, log of capsular penetration, and Gleason score. The data is of size $n = 97$, with 67 of the observations used to fit models, and 30 set aside to assess model performance. The `rstanarm` package version 2.18.2 with default priors in R version 3.5.2 was used for all calculations.

It is not known which combination of these measures best predict, in a logistic regression, PSA > 10. Some attempt at an all-subset regression might be attempted. There are seven potential measures, giving $2^7 = 128$ different possible models. It was decided, in our B, to give each of these 128 models identical initial weight. We will not decide between any of them, and will instead use all of them, as in equation (6) to make posterior probability predictions. We will test the performance of the "grand" model on the 30 hold outs.

As a matter of interest in how measures relate to predictions, the posterior prediction of PSA > 10 as a function of Age is given in Fig. 2.

The red dots are the model (6), i.e. the grand model integrated across all 128 possible submodels. The predictions for those models are also given as small black dots for comparison. The probabilities for high PSA for 40 year olds differs only little from 80 year olds. Whether this difference in important depends on the decisions that will be made by the model. It is not up to the statistician to decide what these decisions will be, or there import.

The Brier score, which is the square of the difference between the observable and its predicted probability, was used to assess model performance. The scores for all models, including the grand model, are given in Fig. 3.

The asterisks are the mean Brier score in the hold out data; thus, these are genuine out-of-sample performances. The range of Brier scores for the 30 data points are given as veritical lines. The red dashed line is the grand model. The blue line is the "best" model according to minimum mean Brier score.

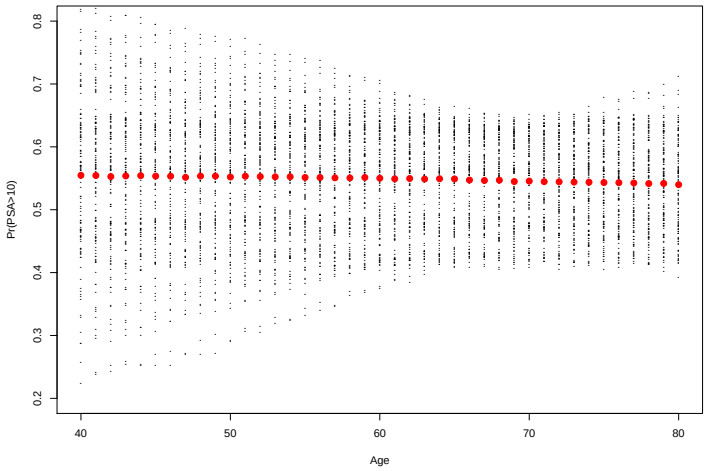

Fig. 2 The red dots are the model (6), i.e. the grand posterior predictive model integrated across all 128 possible submodels. The predictions for those models are also given as small black dots for comparison.

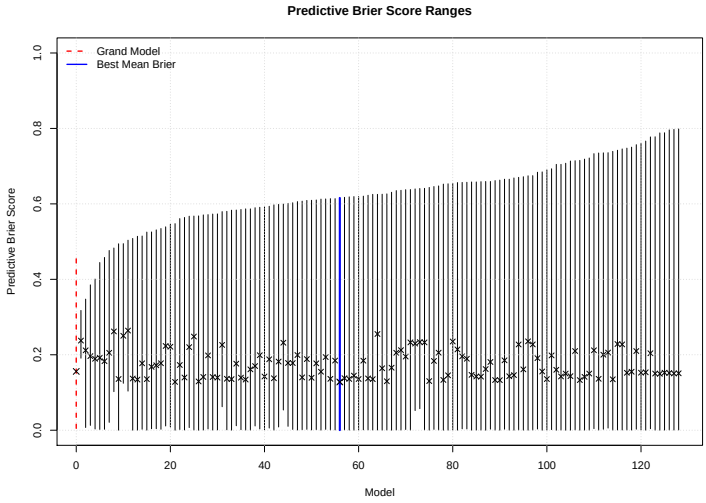

Fig. 3

Some points are interesting. The model with the least range, but higher than average mean score, was the so-called null model; i.e. a logistic regression with only an intercept. All models must be judged relative to this, since it says, in effect, every man has the same fixed chance of a PSA > 10. If a more "sophisticated" model cannot beat this, it should not be used. Which we can see is the case for many models.

The mean Brier score of the grand model is not much higher than the best model. The grand model has the advantage of a shorter range of Brier scores, which is to say, a lower maximum (worse) score. Indeed, only 10 models have smaller ranges, but of those none has a lower mean score. And 5 of those models have higher minimum scores, which is worrisome.

Now whether the mean Brier score is truly best, or whether the model with a low mean score but also with low maximum is best, attributes belonging to the grand model, depends on the decision to which the model will be put.

The grand model uses all measures, at various weights, as it were. The lowest-mean-score model used log cancer volume, log prostate weight, age, and Gleason score. We saw above that age was not especially predictive, though.

5 Last Words

The strategy of searching or computing all subset of linear models is not recommended. It was only shown here to put the model averaging (or rather integrating) procedure through its paces, and to provide a ready-made set of comparison models for a problem. What would be best to demonstrate the true potential is to have actual rival models for an observable, each with different advocates. This would nicely demonstrate the differences between deciding between models, based on predictions, and on using the full uncertainty to make the fairest predictions.

It is clear that we should put these techniques to the test. Don't test: decide to decide.

References

1. W.M. Briggs, *Uncertainty: The Soul of Probability, Modeling & Statistics* (Springer, New York, 2016)
2. W.M. Briggs, in *Beyond Traditional Probabilistic Methods in Economics*, ed. by V. Kreinovich, N. Thach, N. Trung, D. Thanh (Springer, New York, 2019), pp. 22–44
3. W.M. Briggs, H.T. Nguyen, D. Trafimow, in *Structural Changes and Their Econometric Modeling*, ed. by V. Kreinovich, S. Sriboonchitta (Springer, New York, 2019), pp. 3–17
4. W.M. Briggs, Asian Journal of Business and Economics **1**, 37 (2019)
5. W.M. Briggs, H.T. Nguyen, Asian Journal of Business and Economics **1**, (accepted) (2019)
6. W.M. Briggs, arxiv.org/abs/1507.07244 (2015)
7. R. Birkhahn, W.M. Briggs, P. Datillo, S.V. Deusen, T. Gaeta, American Journal of Surgery **191(4)**, 497 (2006)

8. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning* (Springer, New York, 2009)