A Philosopher's Guide to Empirical Success

Malcolm R. Forster†‡

The simple question, what is empirical success? turns out to have a surprisingly complicated answer. We need to distinguish between meritorious fit and 'fudged fit', which is akin to the distinction between prediction and accommodation. The final proposal is that empirical success emerges in a theory dependent way from the agreement of independent measurements of theoretically postulated quantities. Implications for realism and Bayesianism are discussed.

1. Introduction. It would be a miracle for our best scientific theories to be empirically successful if none of their postulated entities existed or if the theories were not approximately or partially true. This is commonly known as the miracle argument, or the cosmic coincidence argument for scientific realism—the view that science provides us with information about the reality behind the observable phenomena. An equally well known response claims that the *truth* of our best scientific theories is not necessary to explain their empirical success; it is sufficient that our theories be true in all their observational claims. Antirealists such as van Fraassen (1980) claim that it is sufficient that everything a theory says about the observed phenomena, past, present, and future is true (that is, the theory is empirically adequate), for this weaker claim also implies that the theory is empirically successful. The debate has naturally focused on the notion of explanation; realists typically claim that empirical adequacy is not

†To contact the author, please write to: Department of Philosophy, University of Wisconsin–Madison, 5185 Helen C. White Hall, 600 North Park Street, Madison, WI 53706; e-mail: mforster@wisc.edu.

‡This paper was written when I was a visiting fellow at the Center for Philosophy of Science at the University of Pittsburgh; I thank everyone for their support.

Philosophy of Science, 74 (December 2007) pp. 588–600. 0031-8248/2007/7405-0004$10.00
Copyright 2007 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/525606 Published online by Cambridge University Press

sufficient for *explanation* in a full blooded sense, and tried to spell that out.[1]

Realists and antirealists have not said much about how empirical success should be defined. Both sides appear to agree that the degree to which a theory 'saves the phenomena' is something like the degree to which a theory *fits* the observed phenomena. In the ideal case, perfect fit requires the truth of the observed consequences of a theory. In the less ideal case, some account of observational error is made, in which case empirical success might be defined in terms of a 'least squares' measure of fit, or by some probabilistic measure of fit using likelihood or the log likelihood.[2]

But what is empirical success, exactly? The problem is surprisingly complicated. For instance, empirical success cannot be goodness of fit with the data, in any unqualified sense, because good fit can be 'fudged', for instance, by introducing many adjustable parameters. 'Fudged' fit is not good, or at least, not something that needs to explained in a realist way. At the same time, it is standard practice in science to use some adjustable parameters; so fit is 'fudged' to some extent in all cases, and therefore we need to distinguish between meritorious fit and fudged fit when they occur together. This will turn out to be related to another distinction—between prediction and accommodation.

In Section 2, the problem is motivated by a simple question—Why are Kepler's laws empirically more successful than Copernicus's theory of planetary motion? Section 3 introduces a positive proposal, in terms of cross-validated fit, while Section 4 explains why this improved answer is still incomplete. It is argued, in terms of a very simple example, that empirical success is intimately tied to the agreement of independent measurements of quantities introduced by a theory. In this view, empirical success is theory laden. Consequences for realism and Bayesians are examined in the final sections.

2. Why Are Kepler's Laws Better than Copernicus's Theory? Kepler's first law of planetary motion says that planets move on elliptical paths with the sun at one focus, while the second law (the area law) says that the line from the sun to the planet sweeps out equal areas in equal times. Note that the particular ellipse or the rate of motions are not specified by these laws; Kepler's laws introduce a number of *adjustable* parameters,

1. See van Fraassen 1980, Chapter 2, for an introduction to the realist debate, and the antirealist position mentioned here is, of course, van Fraassen's Constructive Empiricism.

2. 'Likelihood' is a technical term, which refers to the probability of the observations given the hypothesis (not to be confused with the probability of the hypothesis given the observations, which is a distinctly Bayesian concept).

https://doi.org/10.1086/525606 Published online by Cambridge University Press

such as the size of the ellipse, its eccentricity, orientation, and the rate of motion along its path. A *predictive* hypothesis (or hypothesis, if no confusion will result) assigns a precise Keplerian trajectory to each planet. It is 'predictive' in the sense that it makes exact predictions about the position of any planet at any given time. Kepler's laws define a family of such hypotheses, which I shall call a *model*. Kepler's model, in other words, is a family of predictive hypotheses. Kepler's third law, also known as the harmonic law, says that ratios R^3/T^2, measured independently for each planet, are equal, where R is the mean radius of the planet's motion (the size of the ellipse) and T is the time it takes for the planet to complete one revolution around the sun. The third law introduces no new parameters; it merely constrains the values of the parameters introduced in the first two laws.

Now consider a set of observations that state the relative positions of the planets and the sun at particular times. How should we define the empirical success of Kepler's model relative to this set of observations? Most of the hypotheses in the model will fit the data very badly. So how do we define the fit of a *family* of hypotheses? A charitable definition is that *model fit* is the *best* fit achieved by any hypothesis in the model. But can empirical success be defined as model fit? The answer is no! The argument for this conclusion doesn't depend on how fit is defined; so let's assume that it is the sum of the squared residues, where the residue is the spatial distance between the observed position of a planet and the position specified by a trajectory at the appropriate time.

A Copernican model uses a circle on circle construction (see the caption of Figure 1 for details). Each model introduces the radius of each circle, its period of revolution, and the initial position of each circle as adjustable parameters. The core postulates of Copernicus's *theory* imply nothing about the number of circles that should assigned to each planet, so there are many models compatible with the theory. Now consider a specific model. If its empirical success were defined by how well the best fitting hypothesis in the model fits the data, then would be untrue that Kepler's model is empirically more successful than every Copernican model. Consider an arbitrary Copernican model, C, and compare it to another Copernican model, $C+$, which adds one or more circles to C. Then C is *nested* in $C+$ in the precise sense that all the predictive hypotheses in C are also in $C+$ (**Proof**: Consider the special case in which the added circles have zero radius). The nested property is sufficient to prove that the more complex model can only improve the model fit, for the best fitting hypothesis in C is also in $C+$. Any degree of fit that C can achieve, the more complex model can also achieve. It cannot do worse, and $C+$ will in general do even better. The argument rests solely on the nesting relationship between models—not on how fit is defined.

https://doi.org/10.1086/525606 Published online by Cambridge University Press

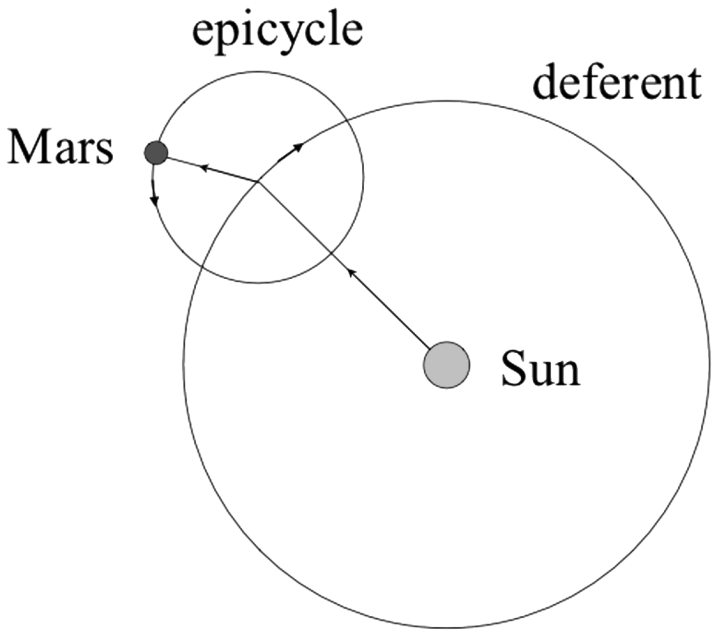

Figure 1. A two-circle Copernican model for the planet Mars. The motion of Mars relative to the sun is modeled as the sum of two vector motions; one represented by the arrow from the sun to the circumference of the main circle, called the deferent, and one from that point to Mars. Each vector has a fixed length and rotates with uniform motion. (The sun could be placed a short distance from the center of the deferent circle, although this would be mathematically equivalent to adding another epicycle.)

Moreover, there is a theorem in mathematics, called the Fourier theorem, that implies that one can, in principle, approximate any planetary trajectory to an arbitrary degree of precision if it uses a sufficient number of circles. This proves that there exists a Copernican model that can approximate any finite set of points sampled from the true planetary trajectories with an arbitrary degree of fit. On the other hand, Kepler's model fits the phenomena only approximately (as we know from Newton's theory). Therefore, there exists a Copernican model that exceeds the best fit achieved by Kepler's laws.

So, we can't define empirical success in terms of model fit if we want to maintain the view that Kepler's model is empirically more successful than every Copernican model. Intuitively, empirical success must somehow take account of the fact that complex Copernican models 'fudge'

https://doi.org/10.1086/525606 Published online by Cambridge University Press

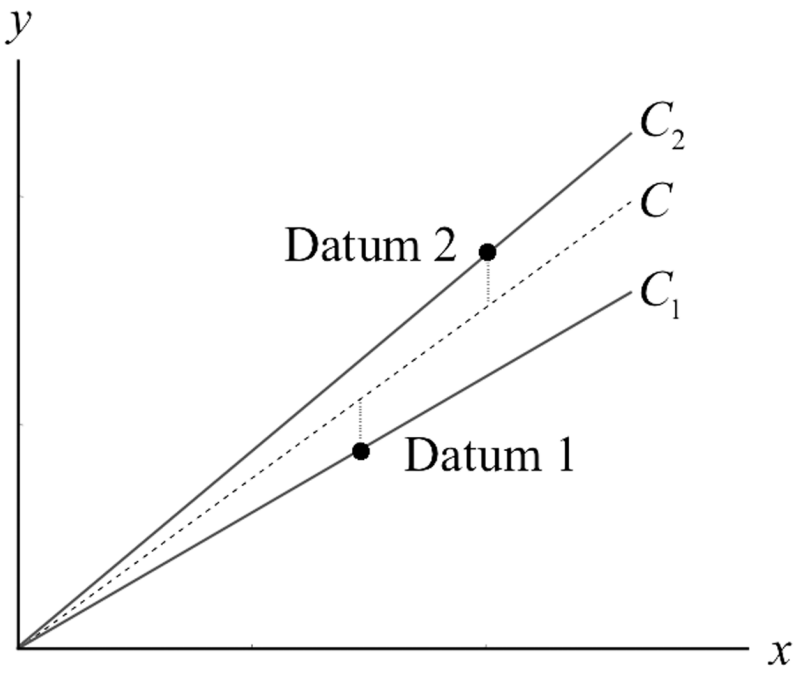

Figure 2. Empirical success in a simple curve fitting example based on a least squares criterion.

their fit by using a large number of circles. It is not so easy to capture this idea precisely.

3. Cross-Validated Fit as a Measure of Empirical Success. Let me begin with a description of the least squares measure of fit to see how it might be modified. Consider a generic curve fitting example in which the model is $y = \beta x$, where x and y are observable quantities, and β is an adjustable parameter. The predictive hypotheses in the model are represented by straight line "curves" passing through the point $(0, 0)$. Now look at the two data points in Figure 2.

The 'distance' of an arbitrary curve in the model, say C, from the data may be measured by the *sum of squared residues* (SSR), where the residues are defined as the y-distances between the curve and the data points. The residues are the lengths of the vertical lines in Figure 2. If the vertical line is below the curve, then the residue is negative; otherwise it is positive. Squaring the residues ensures that the SSR score is always greater than or equal to zero, and equal to zero if and only if the curve passes through

https://doi.org/10.1086/525606 Published online by Cambridge University Press

all the data points exactly. So, the SSR score is an intuitively good measure of the distance (discrepancy) between a curve and the data.

Now define the curve that *best* fits the data as the curve that has the *least* SSR. Recall that any assignment of numbers to the adjustable parameters determines a unique curve, and vice versa. So, in particular, the best fitting curve automatically assigns numerical values to all the adjustable parameters. These values are the least squares estimates the parameters, and this method of parameter estimation is called the method of least squares.

By fitting a model to the data, we obtain a unique best fitting curve.[3] The values of the parameters determined by this curve are often denoted by a hat. Thus, the best fitting hypothesis in the model is denoted by $y = \hat{\beta}x$. The hypotheses represented by the curves C_1 and C_2 are also in the model, but they have a higher SSR score with respect to the data, even though each fits *one* of the data points perfectly.

More exactly, model fit is calculated in the following way:

Step 1: Find the hypothesis that best fits the data. Denote this hypothesis by C_i.

Step 2: Consider a single datum. Square the residue of this datum determined by C_i.

Step 3: Go back to Step 2 and repeat this procedure for all n data.

Step 4: Sum the SR scores and divide by n.

This number actually measures the badness of fit of the model, so model fit is defined as minus this score. The reason that we take the *average* SSR in Step 4 is that we want to use the goodness of fit score to estimate how well the model will predict a 'typical' data point. The goal is the same as in simple enumerative induction—to judge how well the 'induced' hypothesis *predicts* a 'next instance'; we are assuming that all seen instances are representative of the parent population.

Let us remark that the SSR is determined by the seen data, whereas the *predictive accuracy* of a model is, by definition (see Forster and Sober 1994), a measure of how well the model fitted to an arbitrary data set of a particular size will fit *new* data sampled from the same population (that is, using the same experimental procedure and conditions that produced the seen data). Predictive accuracy is like truth; it is not something that models wear on their sleeves. It is an achievable *goal* of scientific modeling (Forster 2002); but like truth, it can only be *estimated* in terms of the

3. There are exceptions to this, for example when the model contains more adjustable parameters than data.

https://doi.org/10.1086/525606 Published online by Cambridge University Press

seen data. The question is whether the SSR score is the best available means of estimating the predictive accuracy of a model.

If the goal is to estimate the *predictive accuracy* of the model, then it is easy to understand why the SSR estimate is biased. For each datum has been used twice: once in the construction of the 'induced' hypothesis (Step 1), and again to calculate how well the 'constructed' hypothesis predicts a typical data point (Step 2). The problem is not that the seen data are unrepresentative of the parent population. The problem is that the best fitting hypothesis, which is used to make the predictions, has been selected, in part, to minimize the 'predictive' error in the *seen* data. But the goal is to estimate the error in predicting *unseen* data. That is why, for example, the SSR score is badly biased when a model has many adjustable parameters; SSR is measuring the ability of the model to *accommodate* data.[4] The problem has nothing to do with the psychological bias of the scientists; it is a logical problem. And it has a logical solution.

One way of removing this bias is to define empirical success in terms of its leave-one-out cross-validation score (CV score), which turns out to be surprisingly similar to the SSR score.

> **Step 1**: Choose a data point i, called the *test datum*, and find the hypothesis that best fits the remaining $n - 1$ data points. Denote this hypothesis by C_i.
>
> **Step 2**: Square the residue of the test datum with respect to C_i.
>
> **Step 3**: Go to Step 1, and repeat this procedure for all n data (in all experiments).
>
> **Step 4**: Sum the scores and divide by n.

The key difference is the test datum i is no longer used twice because is not used to "construct" the hypothesis C_i in Step 1. It is therefore an unbiased estimate of the ability of the model to *predict* new data. The use of CV scores places simple and complex models on an even playing field; there is no need to factor in nonempirical virtues such as simplicity or unification. Moreover, the CV score provides a measure of empirical success that is acceptable to realists and antirealists alike.

3.1. Remark 1. In Section 2, we proved that if model A is nested in model B, then A can never fit the seen data better than B *no matter how fit is defined*. But A can have better *cross-validated* fit than B, so why isn't there a contradiction here? The reason is that model fit was previously

4. This does not undermine the least squares method of parameter estimation. There is no bone to pick with statisticians here.

https://doi.org/10.1086/525606 Published online by Cambridge University Press

defined by the fit of its best fitting member, while cross-validated fit is not defined in that way. It does not measure the ability of the model to fit the *total* data, but rather its ability to predict some of the data from the rest of the data. In a sense, it is not a measure of fit at all. Fit with the *total* data is a measure of accommodation. Good models achieve good fit (as do some bad models), so good models are good accommodators. Accommodation is not something bad. It is *mere* accommodation that is bad.

3.2. Remark 2. Not only does the CV score more perspicuously estimate the predictive ability of a model, but it also gives finer grained information about these abilities. To show this, let C be the curve that best fits the total data, and C_i the curve that best fits the data with datum i left out. If SR_i is the squared residue of datum i relative to C, and PE_i is the squared predictive error of datum 1 relative to C_i, then, by definition, $\text{CV} = 1/n \sum PE_i$ and $\text{SSR} = 1/n \sum SR_i$. Trivially, $\text{CV} = \text{SSR} + 1/n \sum (PE_i - SR_i)$. So CV is equal to the SRR plus a term that corrects the model fit for 'fudging'. What is not so trivial is that $(PE_i - SR_i)$ is greater than or equal to zero *for each datum*.[5] What this means is that the degree of fudging is estimated for each datum, so that the comparison between CV and SR scores is heuristically valuable. Cross-validated fit can point to the particular data that are not predicted well by the model and pose specific questions about the reliability of those data or how the model might be modified to improve those particular predictions.

3.3. Remark 3. The leave-one-out CV score approximates the AIC score (Stone 1977) when the conditions of Akaike's theorem hold (see Akaike 1973; Forster and Sober 1994; Hitchcock and Sober 2004). So, the CV score can do everything that AIC can do, and more because the CV score does not depend on the assumptions of Akaike's theorem. Suppose, for example, that the conditions of Akaike's theorem do not hold in the planetary astronomy example (Kieseppä 1997) and that AIC is a biased estimate of predictive accuracy. The CV score still avoids the double use problem, and is an intuitively plausible estimate of predictive accuracy.

5. **Proof**: Let F be the SSR of the remaining data relative to C, while F_1 is the SSR of the remaining data relative to C_1. Both F and F_1 are the sum of $n - 1$ squared residues. By definition, C fits the *total* data at least as well as C_1. Moreover, the SSR for C *relative to the total data* is just $SR_1 + F$ while the SSR of C_1 relative to the total data is $PE_1 + F_1$. Therefore, $PE_1 + F_1 \geq SR_1 + F$. On the other hand, C_1 fits the $n - 1$ data at least as well as C, again by definition of "best fitting." This implies that $F \geq F_1$. Putting the two inequalities together: $PE_1 + F_1 \geq SR_1 + F \geq SR_1 + F_1$ implies that $PE_1 \geq SR_1$, which is what we set out to prove.

https://doi.org/10.1086/525606 Published online by Cambridge University Press

While leave-one-out CV provides a better definition of empirical success than AIC, leave-one-out CV is not the whole story, as the following examples are designed to show.

4. The Special Case of Perfect Fit. The previous section considered the general case in which fit might not be perfect. In this section we pay closer attention to the ideal case in which the fit with the seen data is perfect. More concretely, suppose that there are three data points (1, 1), (2, 2), and (3, 3), and consider the empirical success of the rival models:

$$\text{ONE}: \quad y = \beta x,$$

$$\text{TWO}: \quad y = a + bx,$$

where x and y are observable quantities, while β, a, and b are adjustable parameters. Both models have zero SSR scores, which means that they accommodate the data perfectly. But note that both have perfect leave-one-out CV scores as well, because they are successful in predicting any datum from the other two. Yet, intuitively, we want to say that ONE is empirically more successful than TWO. In this example, we need to consider something like leave-two-out cross-validated fit in order to distinguish ONE from TWO. ONE has the virtue of only requiring a single data point to uniquely determine a predictive hypothesis, which then fits the other two data points perfectly. TWO does not achieve this kind of empirical success because it requires two data points to determine a unique predictive hypothesis.

It is interesting to see that the empirical success of ONE over TWO is exhibited very naturally in terms of a logical concept of prediction (in which a prediction is any observable consequence deduced from the model). In particular, TWO makes the following prediction:

$$\text{if } y_1 = x_1 \text{ and } y_2 = x_2, \text{ then } y_3 = x_3,$$

plus two similar conditionals obtained by permuting the indices. On the other hand, ONE also predicts that,

$$\text{if } y_1 = x_1, \text{ then } y_2 = x_2 \text{ and } y_3 = x_3,$$

and two others obtained by permuting the indices. Therefore ONE is *predictively* stronger and therefore empirically more successful, and this fact should be included in any complete measure of empirical success. The leave-one-out CV score is therefore an incomplete measure of empirical success.

An equivalent statement is that the total data provide three independent and agreeing measurements of the parameter β (Harper 2002, 2007; Myrvold and Harper 2002), or equivalently, the parameters of ONE are more

https://doi.org/10.1086/525606 Published online by Cambridge University Press

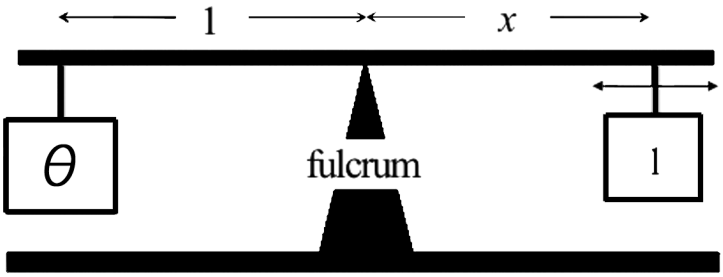

Figure 3. The beam balance experiment.

strongly *overdetermined* by the *total* data (Forster 1988; Norton 2000a, 2000b).

Here is another example. Suppose a beam is supported at the center on a fulcrum (Figure 3). Two objects will balance each other when hung on the beam on opposite sides of the fulcrum if and only the distance from the fulcrum times the downward force on each object is equal. Let one object be a kilogram mass, while the mass of the other object, denoted by θ, is unknown except for the assumption that is it greater than zero. If x denotes the distance of the one kilogram mass from the fulcrum, and d is the distance that the other mass is hung on the other side, then Newton's theory of motion implies that if the objects balance, then $x = \theta d$ (given unstated auxiliary assumptions). We can simplify this further by supposing that the object with the unknown mass θ is always hung exactly one unit distance from the fulcrum, while the kilogram mass is moved back and forth until the beam balances. Then the equation simplifies to $x = \theta$, where x is an observable quantity, and θ is an adjustable parameter.

Now consider a set of experiments with three objects, labeled a, b, and c, hung on the beam balance by themselves, and in pairs. There are six experiments in total, one with a, one with b, one with c, one with a*b, one with b*c, and one with a*c, where a*b refers to the composite object consisting of a placed with b, and so on. In each experiment, suppose we make a single measurement and the results are, respectively, $x_1 = 3$, $x_2 = 4$, $x_3 = 5$, $x_4 = 7$, $x_5 = 9$, and $x_6 = 8$. Treat the masses of all six objects as unknown. Then the model, which we call the primitive model (PRIM) produces six unknown quantities in six equations: $x_1 = m(a)$, $x_2 = m(b)$, $x_3 = m(c)$, $x_4 = m(a*b)$, $x_5 = m(b*c)$, and $x_6 = m(a*c)$. PRIM is not able to make any predictions of any part of the data from any other part of the data, and therefore has no empirical success. On the other hand, the usual Newtonian model (NEWT) is supplemented by

https://doi.org/10.1086/525606 Published online by Cambridge University Press

the Law of Composition of Masses (LCM), which states that mass of composite objects are equal to the sum of the masses of the component parts. It implies, for example, that $m(a*b) = m(a) + m(b)$. NEWT has six equations in three unknowns, so each parameter has two independent measurements, which agree. NEWT's parameters are overdetermined by the data.

In terms of the distinction between prediction and accommodation, the empirical success of NEWT is exhibited by the truth of the predictions $x_4 = x_1 + x_2$, $x_5 = x_2 + x_3$, and $x_6 = x_1 + x_3$. PRIM makes no such predictions, even though it accommodates the data perfectly well.

5. Lessons for Realism. The agreement of independent measurements is now directly linked to the overdetermination of parameters. The miracle argument exploits this fact as follows. Surely, it would be a miracle if such measurements agreed if there was no quantity being measured or no single quantity being measured. Why should independent measurements agree unless they are measurements of something? Hence, the realist's explanation of the empirical success of NEWT, for example, naturally supposes that there exists an additive property objects have, namely mass. Of course, it is logically possible that the agreement is a cosmic coincidence, or a brute fact; so antirealism is still a logically consistent position.

Put another way, if empirical success were 'fit with data' then a realist would be claiming to explain fudged fit as well as meritorious fit. But fudged fit should not be explained in a realist way. The explanation of meritorious fit, on the other hand, appeals to specific parts of the theory in a way that is directly relevant to realist intuitions.

6. Problems for Bayesianism? If fit is defined in terms of the probability of the data given then hypothesis (the *likelihood* of the hypothesis relative to the data), then it is a measure of mere accommodation. Does this present problems for probabilistic approaches to confirmation such as Bayesianism? Consider the example of Section 4 again. Which model, NEWT or PRIM, has the highest probability *given the data*?

To examine this question, consider a very specific chance setup. Suppose that someone randomly chooses with probability 1/2 whether to generate the data using NEWT or PRIM.

If NEWT is chosen, then the masses of a, b, and c are assigned values from 1 to 81, such that all possible assignments are equally probable. That is, the probability of any assignment is $1/81^3 = 1/9^6$. These mass values determine the corresponding values of x_1, x_2, x_3 without error, as well as the values of x_4, x_5, x_6 since the composite masses are determined by LCM. The likelihood of NEWT is therefore $1/9^6$.

If PRIM is chosen as the data generating process, then the masses of

https://doi.org/10.1086/525606 Published online by Cambridge University Press

a, *b*, and *c*, and the composite masses *a***b*, *b***c*, and *a***c* are independently assigned a value from 1 to 9, such that all possible data sets compatible with this hypothesis have probability $1/9^6$.

Note that the two models are logically independent; neither entails the other. The data in Section 4 could have been generated by either model. So, the likelihood relative to both hypotheses is exactly the same, namely $1/9^6$. Since the prior probabilities are equal, and the likelihoods are the same, it follows from Bayes theorem that the posterior probabilities are the same. This result is unexpected because it appears that the data provides evidence for the truth of LCM and therefore some evidence in favor of NEWT over PRIM. But there is no contradiction.

To analyze the situation further, divide the data into two parts: $\text{data1} = \{x_1 = 3, x_2 = 4, x_3 = 5\}$ and $\text{data2} = \{x_4 = 7, x_5 = 9, x_6 = 8\}$. Then

$$P_{\text{NEWT}}(\text{Data}) = P_{\text{NEWT}}(\text{data1})P_{\text{NEWT}}(\text{data2}|\text{data1})$$

$$P_{\text{PRIM}}(\text{Data}) = P_{\text{PRIM}}(\text{data1})P_{\text{PRIM}}(\text{data2}|\text{data1}),$$

where Data is the total data. The instructive point is that the predictive probability $P_{\text{NEWT}}(\text{data2}|\text{data1})$ is very high for NEWT, whereas the corresponding probability for PRIM is low. Nevertheless, the likelihoods with respect to the total evidence are the same because the probability of data1 is higher for PRIM than for NEWT. In this example, the result is correct because the probabilities of data1 are given by the models, and are therefore predictive probabilities. But is this true in all examples?

The problem is that models in *physics* do not postulate any stochastic mechanism for generating the values of parameters such as mass. The probabilities of data1 are therefore imposed from the outside by Bayesian statisticians. The problem is that the likelihood (relative to the total evidence) is then an inadequate measure of empirical success (see Forster 2006 for other examples). Likelihoods are, in general, measures of accommodation, and the distinction between prediction and accommodation is washed away. Popper (1959) got it right a long time ago: science aims not at truth, or high probability of truth, but at *informative* truth. Tautologies have high probabilities; they accommodate everything, and make no predictions at all!

REFERENCES

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle", in B. N. Petrov and F. Csaki (eds.), *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, 267–281.

Forster, Malcolm R. (1988), "Unification, Explanation, and the Composition of Causes in Newtonian Mechanics", *Studies in the History and Philosophy of Science* 19: 55–101.

https://doi.org/10.1086/525606 Published online by Cambridge University Press

——— (2002), "Predictive Accuracy as an Achievable Goal of Science," *Philosophy of Science* 69: S124–S134.
——— (2006), "Counterexamples to a Likelihood Theory of Evidence," *Mind and Machines* 16: 319–338.
Forster, Malcolm R., and Elliott Sober (1994), "How to Tell When Simpler, More Unified, or Less *Ad Hoc* Theories Will Provide More Accurate Predictions", *British Journal for the Philosophy of Science* 45: 1–35.
Harper, William L. (2002), "Howard Stein on Isaac Newton: Beyond Hypotheses", in David B. Malament (ed.), *Reading Natural Philosophy: Essays in the History and Philosophy of Science and Mathematics*. La Salle, IL: Open Court, 71–112.
——— (2007), "Newton's Methodology and Mercury's Perihelion Before and After Einstein." *Philosophy of Science*, in this issue.
Hitchcock, Christopher R., and Elliott Sober (2004), "Prediction versus Accommodation and the Risk of Overfitting", *British Journal for the Philosophy of Science* 55: 1–34.
Kieseppä, I. A. (1997), "Akaike Information Criterion, Curve-Fitting, and the Philosophical Problem of Simplicity", *British Journal for the Philosophy of Science* 48: 21–48.
Myrvold, Wayne, and William L. Harper (2002), "Model Selection, Simplicity, and Scientific Inference", *Philosophy of Science* 69: S135–S149.
Norton, John D. (2000a), "The Determination of Theory by Evidence: The Case for Quantum Discontinuity, 1900–1915", *Synthese* 97: 1–31.
——— (2000b), "How We Know about Electrons", in Robert Nola and Howard Sankey (eds.), *After Popper, Kuhn and Feyerabend*. Dordrecht: Kluwer, 67–97.
Popper, Karl (1959), *The Logic of Scientific Discovery*. London: Hutchinson.
Stone, M. (1977), "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion", *Journal of the Royal Statistical Society B* 39: 44–47.
van Fraassen, Bas (1980), *The Scientific Image*. Oxford: Oxford University Press.

https://doi.org/10.1086/525606 Published online by Cambridge University Press