

Modeling scientific evidence: the challenge of specifying likelihoods

Patrick Forber
Tufts University

March 30, 2010

In science evidence is objective. Whether data count as evidence for or against a hypothesis should not depend on my unconstrained subjective degrees of belief. Philosophical models of confirmation should aim to capture, in at least some respects, the objective character of scientific evidence. Indeed, this aim drives the resistance to subjective Bayesianism, leading to the development of objective Bayesian accounts, or to the abandonment of Bayesianism all together. Modeling an objective evidence relation in a probabilistic framework faces two challenges: the probabilities must have the right epistemic foundation, and they must be specifiable given the hypotheses and data under consideration. Here I will explore how Sober's (2008; 2009) approach to confirmation handles these challenges of *foundation* and *specification*. Briefly, Sober addresses the first by eschewing subjective credences in favor of the law of likelihood, and tackles the second by illustrating how quantitative predictions of competing hypotheses specify likelihood values. Yet neither challenge is completely overcome. In particular, I will argue that the specification problem proves especially difficult, and undermines the general argument that the law of likelihood provides a more objective representation of scientific evidence.

1 The foundation challenge

Confirmation theory provides a set of competing strategies for modeling the relation between scientific theory and empirical evidence. Most of these models are borrowed from statistics and applied to epistemology. Such epistemological models are normative; they tell us how we should reason about evidence. One aim of confirmation theory, an aim I will take as primary, is to capture the objective nature of scientific evidence, for evidence provides the neutral arbiter, the normative grounds for theory choice. Given this aim, confirmation-theoretic models, and the components of such models, should have the right sort of epistemic foundation. For probabilistic models, the component probabilities need to be *objective*. Identifying the epistemic source of these probabilities, and assessing whether this source makes the probabilities objective in the right way, is the foundation challenge.

The best tool for unpacking this objectivity demand is *probability coordination*. Lewis (1980) proposed the principal principle—that we should set our (subjective) credences to the (objective) chances we observe—as the way to coordinate probabilities in an epistemological model using features of the world. There are some technical problems associated with how exactly we should implement the principal principle (Hall, 1994), but the core idea is right: probabilities in our epistemological models should reflect features of the world. What sorts of features certain probabilities track, and how they track those features, determine how objective they are. If they quantify our uncertainty or opinion, they are more subjective. If they reflect empirical information about the constitution of the world or the nature of accepted theory then they are more objective. In light of probability coordination, meeting the foundation challenge amounts to identifying the processes, objective chance setups, or consensus theories capable of coordinating probabilities across a group of practicing scientists.

Now consider two central strategies for modeling the confirmation relation with probability: Bayesianism and likelihoodism.¹ The Bayesian strategy postulates three kinds of probabilities: priors, posteriors, and likelihoods. These probabilities are usually interpreted as subjective credences. Credences are updated according to Bayes' rule (equation 1), and O confirms H if and only if $Pr(H|O) > Pr(H)$.

$$Pr(H|O) = \frac{Pr(O|H)Pr(H)}{Pr(O)} \quad (1)$$

That is, O confirms H just when my credence in H increases upon learning O . Credences count as subjective because they are coordinated by appealing to the opinions and beliefs of particular individuals. A typical Bayesian way to rescue the objectivity of evidence is to appeal to convergence theorems that show that, given enough evidence, diverse initial credences will eventually converge.²

The likelihood strategy takes a more minimalist approach, avoiding credences and building confirmation-theoretic models with likelihoods. Likelihoods are conditional probabilities that quantify the chance of observing some data given a proposed hypothesis is true. Sober tends to follow the likelihood approach, and defines the central principle as follows.

Law of likelihood: The observations O favor hypothesis H_1 over hypothesis H_2 if and only if $Pr(O|H_1) > Pr(O|H_2)$. And the degree to which O favors H_1 over H_2 is given by the likelihood ratio $Pr(O|H_1)/Pr(O|H_2)$ (Sober, 2008, 32).

Sober favors the likelihood over the Bayesian strategy because there

¹See Fitelson (2007) for a formal comparison of these two approaches and discussion of some problem cases.

²Whether these theorems deliver is controversial. See Earman (1992); Hawthorne (forthcoming) for discussion.

often is no clear, objective way to constrain credences in the hurly burly of scientific practice. When discussing the contrast, he claims:

What likelihoodists mean by probability is not simply that an agent has some degree of belief [credence]. For one thing, the concept of probability needs to be interpreted more normatively. $Pr(O|H)$ is the degree of belief you *ought* to have in O given H is true. But likelihoodists also like to think of these conditional probabilities as reflecting objective matters of fact (Sober, 2008, 40).

So long as credences are coherent, an agent may assign them however she likes. There is no complete probability model, such as the one governing standard decks of well-shuffled cards, to use to coordinate credences in science. This subjectivity of the credences infects the Bayesian confirmation relation. In contrast, likelihoods depend on what the world is like and what hypotheses predict about the world, not on the whims or opinions of individuals. Thus, Sober relies on them to underwrite an objective account of the confirmation relation.

Sober's move away from Bayesianism to the law of likelihood shows his sensitivity to the foundation problem. In essence, he proposes a criterion of adequacy for selecting the appropriate confirmation-theoretic model:

When prior probabilities can be defended empirically, and the values assigned to a hypothesis' likelihood and to the likelihood of its negation are also empirically defensible, you should be a Bayesian (Sober, 2008, 32).

Thus, a confirmation-theoretic model must meet the foundation challenge in order to provide an adequate representation of scientific evidence.

Likelihoods depend on what a hypothesis predicts and what sorts of chance processes are postulated for producing the target observation. *Prima facie*, these probabilities are empirical defensible in a way that credences

are not. Scientists working within the same theoretical framework should agree upon what a particular hypothesis predicts and on the sorts of processes that are operating in the study system. There are, however, some complications for this solution to the foundation problem. First, in the standard Kolmogorov axiomatization the definition of conditional probability entails that likelihoods can be defined in terms of unconditional probabilities, probabilities that look like credences (equation 2)

$$Pr(O|H) = \frac{Pr(O\&H)}{Pr(H)} \quad (2)$$

It is unclear how apparently objective likelihoods can be defined or determined by subjective credences. Either the objective likelihoods, coordinated given the predictions of the hypotheses, constrain credences in some way and so make them less subjective, or likelihoods and credences must all deal in the same subjective currency and so the “objective” coordination of likelihoods (rather than credences) is illusory.

Sober (2008, 38–41) has a reply to this puzzle: the Kolmogorov definition only makes sense when the unconditional probabilities are “well defined.” This amounts the claim that the definition applies only when the unconditional probabilities also meet the foundational criterion for adequacy. Sober’s take is both attractive and intuitive. Likelihoods certainly look more objective than credences. Insofar as this is true, we should opt for likelihoodism over Bayesianism (given we accept the foundational criterion of adequacy). Yet whether they actually are more objective depends on what, exactly, the probabilities track. And determining this depends on how they are specified. Unpacking this challenge shows that the decision between likelihood and Bayesian modeling strategies is not as simple as proposed

2 The specification challenge

The second challenge, specification, concerns how to determine the component probabilities in a confirmation-theoretic model. The challenge is this: if we use probabilistic confirmation-theoretic models to represent scientific evidence then we must be able to specify precisely the component quantities in the model. This challenge is the local instantiation of a general criterion for constraining interpretations of probability, namely Salmon's (1966) criterion of *ascertainability*. What confirmation theory needs is a method that can, in principle, ascertain the values of the probabilities that appear in a confirmation-theoretic model.

I will focus on the specification of one component in particular: likelihoods. Within the likelihood strategy these probabilities are supposed to have the objective foundation that credences usually lack. In addition, the Bayesian strategy requires likelihoods to update with Bayes' rule. So likelihoods have a special importance in confirmation theory. For likelihoods to be ascertainable, we need a method that specifies a *likelihood function*. Examining some formal details about likelihoods helps make this clear.

A likelihood function determines the probability of making a particular observation as a function of the parameter value(s) for an assumed background statistical model and that observation. To put it another way, the background statistical model includes a family of precise probability distributions indexed by one or more parameters. The likelihood function specifies the probability of making the observation we in fact made supposing one distribution in the family is true.³ For likelihood functions to work in this way, the data set (the observation) is modeled as an instance of random variable (X) with an unknown distribution (θ). Suppose we have two hypotheses: H_1 states that X has distribution θ_1 and H_2 states that X has distribution θ_2 . Formally, the likelihood value (the probability

³See, for example, Royall (1997) for a standard formulation of likelihoods in statistics.

of O given H) for the observation that $X = x$ is specified by the likelihood function $L(\theta; x)$ (as in 3).

$$Pr(X = x|H_1) = L(\theta_1; x), Pr(X = x|H_2) = L(\theta_2; x) \quad (3)$$

Likelihoods come from likelihood functions, so for likelihoods to be objective then our scientific theories must specify adequate likelihood functions.

The specification challenge has both a broad and narrow version. The broad version concerns what sort of information should be included in the statement of the observation (O) and the set of rival hypotheses (H_1, H_2, \dots, H_n). This raises familiar philosophical issues that Sober handles well. The narrow version concerns how particular hypotheses, about (say) natural selection in Galapagos finches, specify actual likelihood values for certain observations, such as a change in mean beak size being correlated with a change in the availability of seed resources. The serious and underappreciated difficulties for the likelihood strategy lurk here. I will discuss both versions of the problem, focusing mainly on the narrow version, and make a connection to formal problems in statistics.

2.1 Broad specification

Meeting the broad specification challenge requires determining what sort of information is necessary to specify likelihoods for any hypothesis-observation pair. How should we fill in the variables H and O to get an accurate and objective value for $Pr(O|H)$?

Sober focuses primarily on the broad version. With respect to O , he defends the principle of total evidence, a claim that all available information about the data set or observation must be included in O . Failure to do so can generate artifactual likelihood values that yield unintuitive results in our confirmation-theoretic model. Data we take as clear evidence

for a hypothesis ends up disconfirming it given an underspecified set of observations (Sober, 2008, 41–46).

With respect to H , Sober makes two relevant points. First, familiar from Duhem-Quine holism, any hypothesis requires auxiliary hypotheses to make predictions, and therefore the auxiliaries are assumed in the likelihood specification. Not just any auxiliary hypothesis will do. Suitable auxiliaries must be true, independently attested, and independent of the truth of hypotheses under test (Sober, 2008, 152). Second, the broad specification of likelihoods requires a description of the observation process. Sober (2009, 82) calls this the “anthropic principle”, a requirement that we include “as complete a description as the agent possesses of the process by which he or she came to observe E [the evidence or observation] is true.” This principle helps the law of likelihood cope with observation selection effects, providing a further correction to our confirmation-theoretic model.

These principles are sufficient for meeting the broad specification challenge. Perhaps their only defect may be that the informational demands often go unsatisfied in practice. Strevens (2009, 96) argues that if we have not established the truth of the relevant auxiliary hypotheses, then we must take into account rival auxiliaries. In these cases, according to the theorem of total probability, specifying the likelihood for a hypothesis-observation pair requires weighting the component likelihoods for the pair plus each rival auxiliary by the credences for the competing auxiliaries. Subjective credences thus reappear in the calculation of the overall likelihood for the hypothesis-observation pair. The frequency of these cases in science is an open question, but scientists certainly must cope with uncertainty about auxiliaries sometimes.

The broad specification challenge concerns general epistemological principles and how these principles can and should guide reasoning about evidence. Good philosophical resources have been brought to bear on this problem. However, these general principles are not sufficient to specify

precise likelihood values for many scientific hypotheses, for the narrow specification challenge presents new difficulties.

2.2 Narrow specification

How do we specify the precise likelihoods for scientific hypotheses? One attractive response to this challenge is that this is not the job for philosophy. Science, when done well, provides the solution. We simply must determine the hypothesis, the proper auxiliaries, and the process of observation for a particular concrete case and compute the value for the likelihood of O given H . Another attractive response is that it does not much matter, so long as scientists agree on qualitative assessments of confirmation (e.g., $Pr(O|H_1) > Pr(O|H_2)$). If there is qualitative agreement then convergence theorems show that science will get there in the end (Hawthorne, forthcoming). Both of these responses hide complications. Getting even qualitative agreement requires identifying a likelihood function, or family of functions, that meet the foundational criterion of adequacy—the (qualitative or quantitative) specification of likelihoods needs to be empirically defensible.

There are two sources for complications. First, many cases in science lack sufficient empirical information to specify a precise likelihood for a hypothesis. Second, the relevant physical processes are often too complex to be modeled accurately with a likelihood function. I will discuss both sources, using Sober's clear examples to structure the argument.

Medical diagnosis provides an example of how likelihood specification should work. Suppose we have a simple diagnostic test that returns a positive or negative result for some disease, and we want to test a particular patient for this disease. We have two hypotheses, our patient either has the disease (D) or she does not ($\neg D$). Observing a positive result (say) confirms D over $\neg D$ since $Pr(+|D) \gg Pr(+|\neg D)$. The likelihood model succeeds in capturing what is objective about the evidence because the

component probabilities (the four likelihoods) are coordinated using frequency data (Figure 1). We have independent access to whether patients have the disease and this is the basis for medical research on the reliability of the test, estimating the rates of false positives and negatives. The research yields sufficient information to specify the likelihoods in an empirically defensible way. This is a best case scenario for the likelihood strategy.

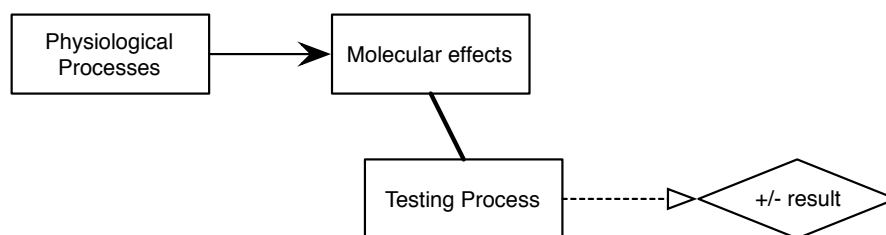


Figure 1: A schematic representation of the medical diagnostic case. Given ample frequency data on patients with or without the disease and their test results we can coordinate the four relevant likelihoods $Pr(+/- | D)$ and $Pr(+/- | \neg D)$.

Sober (2008, 192–201) illustrates how the specification should be done in evolutionary biology with a schematic test of selection plus drift (SPD) versus pure drift (PD). To generate likelihoods for SPD and PD Sober assumes a highly idealized model of the evolutionary process, a simple quantitative trait with a known fitness function, and that scientists have an uncontroversial and complete optimality model for the trait in question. In principle, this is not a problem, for it makes possible a simple and abstract analysis of how tests of this kind should work. In practice, we lack the empirical information to coordinate likelihoods for such a test. Consider, say, Galapagos finches and whether a change in beak size is the result of SPD or PD. The SPD hypothesis postulates that beak size evolves by natural selection in response to changing availability of seed types in the environment, the main source of food for our species of finch. The

PD hypothesis postulates that beak size simply drifts. The empirical test compares changes in beak size to changes in seed availability over time and generates a measure of correlation between trait and environmental variables. To model this instance of scientific evidence using likelihoods we need to specify objective likelihoods for the observed correlation given SPD and PD (Figure 2). Yet evolutionary biologists simply lack the sort of frequency data necessary to objectively coordinate the likelihoods. In contrast to the medical diagnosis case, there are no test populations where we have independent access to the evolutionary process. We cannot estimate the reliability of the test in order to empirically defend the likelihoods for Galapagos finch case.

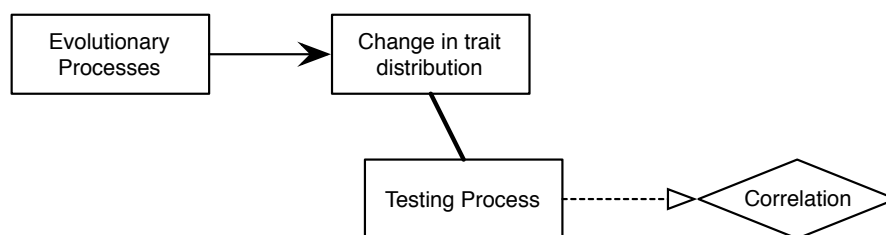


Figure 2: A schematic representation of the evolutionary case. We need empirical information to coordinate the likelihoods $Pr(\text{correlation}|SPD)$ and $Pr(\text{correlation}|PD)$.

There is an alternative way to objectively coordinate the likelihoods: appeal to empirical generalizations about evolution or background information about finch ecology. The complexity of the evolutionary process makes this problematic. Note the substantive assumptions built into the SPD versus PD test. They include many idealizations—patently *false* assumptions—as auxiliaries or components of the SPD and PD hypotheses. This is necessary because both evolution and organisms are complex. Yet it undermines the objectivity of the likelihoods for SPD and PD. Building idealizations into the test compromises the requirement for true auxiliaries and makes possible methodological disagreements about the ide-

alizations. For example, Sober's test assumes an optimality model. The viability of optimality modeling in evolution is controversial. And there is the further assumption that there is a consensus optimality model for the system to apply the test. Were there agreement about optimality modeling in general or the optimality model for finches in particular, then the likelihoods would be more objective. Such a consensus is often the exception, rather than the rule, in evolutionary biology.

These two problems, lack of empirical information for coordination and overly complex physical processes, are substantive version of formal problems in statistical inference. Drawing the connections between these problems and more formal problems will help support my case that narrow specification is a genuine challenge.

2.3 Formal problems with substantive implications

There are recognized formal problems that face any determination of the appropriate likelihood function for statistical inference. How to frame and parameterize an inference problem is one. How to cope with so-called nuisance parameters is another. Due to these problems, determination of a likelihood function depends, to a large degree, on art and application.

Recall that a likelihood function treats the data as the value of a random variable and so must assume a statistical model. The statistical model identifies a family of probability distributions as a function of the observed data and instantiated values of one or more parameters. Given this background model, the likelihood function ($L(\theta; x)$) specifies the probability of observing the data (x) given the distribution determined by the parameter value(s) (θ).

Framing an inference problem, such as the SPD vs. PD test, thus requires treating the data as the outcome of a random variable with an unknown distribution, determining the statistical model to identify the sets of possible distributions, and parameterizing the problem in an informa-

tive way to index these possible distributions. Different choices about how to frame an inference problem lead to different likelihood functions. Bayarri et al. (1988); Bayarri & DeGroot (1992) argue that subjective considerations must enter into these framing decisions, and that such decisions about what likelihood function to use affect the subsequent statistical inferences made from the target data. They examine different likelihood functions for a particular inference problem and show that the choice of function produces different results. Moreover, there are no general methodological or epistemic arguments for preferring one sort of function over the others; each way of determining a likelihood function is useful in different inference problems.

The problem of nuisance parameters has a similar character. This problem arises when the likelihood function depends on more than one parameter. Additional parameters complicate the inference problem, and often are not of direct interest. Hence they are “nuisance” parameters. There is no generally viable strategy for dealing with nuisance parameters. Instead, there are a set of ad hoc methods available for managing the problem, such reparameterizing the problem to eliminate them or use more complex devices, such conditional likelihoods or likelihood profiles, to control the effects of the extra parameters (Royall, 1997). The decision to use a particular method depends on the details of the inference problem and on what works. How to handle nuisance parameters is another decision with a subjective element that has non-trivial consequences for the resulting likelihood function (Bayarri & DeGroot, 1992).

In the context of scientific practice, these formal problems have substantive implications for the objective status of likelihoods. One goal of confirmation-theoretic models is to capture the objectivity of evidence in real scientific cases. The flexibility in framing and parameterizing an inference problem to determine a likelihood function, and the variety of ways for handling any nuisance parameters, undermine the supposed objectiv-

ity of likelihoods. Representing a complex natural process with a tractable statistical model inherently involves simplification and abstraction. Coordination would still be possible, if sufficient information were available to the likelihoods empirically defensible, as in the medical diagnostic case. For the finch case, if the evolutionary process were simple enough then there would be a consensus statistical model to determine an objective likelihood function. Or, if the right sort of empirical data were available, on (say) how often directional selection processes produce an observed degree of correlation between trait and environmental variables, then it would be possible to coordinate likelihoods in the right way. The complexity and lack of information make these formal issues into real problems for the claim that likelihoods are more objective than credences.

3 Specification and epistemic foundations

The complications with specification make the objectivity of likelihoods uncertain at best. Idealizations, necessary to represent an inference problem with a likelihood function, lack the strong epistemic foundation that true auxiliary hypotheses have, as evidenced by the rich philosophical investigation over evaluating idealizations and tradeoffs (Levins, 1966; Matthewson & Weisberg, 2009). The lack of complete information in real scientific cases, and the lack of consensus over methodological strategies for formulating hypotheses, also present difficulties. Without complete information or consensus scientists must make untested assumptions, or agree to disagree on methodological practices. It is possible to specify some likelihood values in this way, but these likelihoods depend more and more on inclinations and beliefs of the scientific practitioners, and fall further from satisfying the foundational criterion of adequacy. Idealizations and lack of information undermine the apparently objective epistemic foundation of likelihoods.

In particular cases, or kinds of cases, the Bayesian strategy or the likelihood strategy may provide the more objective confirmation-theoretic model. Whether the Bayesian or likelihood strategy is the most general or objective framework for understanding evidence across the sciences is an empirical question, decided on how the strategies handle representative sets of cases, not on general epistemological arguments regarding the apparent objectivity of their components.

Thus, we have no *general* reason to expect the likelihood strategy to provide more objective confirmation-theoretic models than the Bayesian strategy. Simply relying on likelihoods and eschewing credences does not give the likelihood strategy an objective edge when representing scientific evidence. Whether a particular Bayesian or likelihood model counts as more or less objective than the other depends entirely on the details of the application. Do the component probabilities in the model meet the foundational criterion for adequacy? Assessing this question about objectivity requires meeting the specification challenge. Do we have sufficient empirical information to coordinate likelihoods and/or credences? Do we have a consensus model for the target system to ground the likelihood function? How we specify likelihoods is not a merely formal problem for statistics, nor an empirical problem for science, but a genuine problem for epistemology.

References

- Bayarri, M. J. & DeGroot, M. H. (1992). Difficulties and ambiguities in the definition of a likelihood function. *Journal of the Italian Statistical Society*, 1, 1–15.
- Bayarri, M. J., DeGroot, M. H., & Kaldane, J. B. (1988). What is the likelihood function? In S. S. Gupta & J. O. Berger (Eds.), *Statistical decision theory and related topics IV* (pp. 3–27). New York: Springer Verlag.

- Earman, J. (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge: MIT Press.
- Fitelson, B. (2007). Likelihoodism, bayesianism, and relational confirmation. *Synthese*, 156, 473–489.
- Hall, N. (1994). Correcting the guide to objective chance. *Mind*, 103(412), 505–517.
- Hawthorne, J. (forthcoming). Confirmation theory. In S. Bandyopadhyay & M. R. Forster (Eds.), *Philosophy of Statistics*, volume 7 of *Handbook of the Philosophy of Science*.
- Levins, R. (1966). The strategy of model building in population biology. *American Scientist*, 54, 421–431.
- Lewis, D. (1980). A subjectivist's guide to objective chance. In R. Jeffrey (Ed.), *Studies in Inductive Logic and Probability*, volume 2 (pp. 263–293). Berkeley: University of California Press.
- Matthewson, J. & Weisberg, M. (2009). The structure of tradeoffs in model building. *Synthese*, 170, 169–190.
- Royall, R. M. (1997). *Statistical Evidence: A likelihood paradigm*. New York: Chapman and Hall/CRC.
- Sober, E. (2008). *Evidence and Evolution: The Logic Behind the Science*. Cambridge: Cambridge University Press.
- Sober, E. (2009). Absence of evidence and evidence of absence: Evidential transitivity in connection with fossils, fishing, fine-tuning, and firing squads. *Philosophical Studies*, 143, 63–90.
- Strevens, M. (2009). Objective evidence and absence: Comment on Sober. *Philosophical Studies*, 143, 91–100.