

Counterexamples to a Likelihood Theory of Evidence

Malcolm R. Forster¹
Department of Philosophy
University of Wisconsin-Madison
July 4, 2006

(Forthcoming in *Minds and Machines*)

ABSTRACT: The Likelihood Theory of Evidence (LTE) says, roughly, that only likelihoods matter to the evidential comparison of hypotheses (or models). There exist counterexamples in which one can tell which of two hypotheses is true from the full data, but not from the likelihoods alone. These examples demonstrate the power of other forms of reasoning, such as the consilience of inductions (Whewell, 1858). Bayesian and Likelihoodist philosophies of science are more limited in scope.

¹ My thanks go to all those who responded well to the first version of this paper presented at the University of Pittsburgh Center for Philosophy of Science on January 31, 2006, and especially to Clark Glymour. A revised version was presented at Carnegie-Mellon University on April 6, 2006. I also wish to thank Jason Grossman, John Norton, Teddy Seidenfeld, Elliott Sober, Peter Vranas, and three anonymous referees for valuable feedback.

This paper is part of the ongoing development of a half-baked idea about cross-situational invariance in causal modeling introduced in Forster (1984). I appreciated the encouragement at that time from Jeff Bub, Bill Demopoulos, Michael Friedman, Bill Harper, Cliff Hooker, John Nicholas, and Jim Woodward. Cliff Hooker discussed the idea in his (1987), and Jim Woodward suggested a connection with statistics, which took me 20 years to figure out.

1. Introduction

Consider two simple hypotheses, h_1 and h_2 , with likelihoods denoted by $P(E | h_1)$ and $P(E | h_2)$ respectively, where E is the total observed data—the actual evidence. By definition, a simple statistical or probabilistic hypothesis has a precisely specified likelihood (as opposed to composite hypotheses, or models, which do not—see below). The use of the term “simple” in this context is standard in the classical statistics literature—it does not mean that the hypothesis is simple in any intuitive sense, and it does not imply that the evidence is simple or that the relationship with the hypothesis or that its evidence is simple.

A Likelihood Theory of Evidence (LTE) is presupposed in the standard Bayesian method of comparing hypotheses, according to which two simple hypotheses are compared by their posterior probabilities, $P(h_1 | E)$ and $P(h_2 | E)$. Bayes theorem tells us that

$$\frac{P(h_1 | E)}{P(h_2 | E)} = \frac{P(h_1)}{P(h_2)} \times \frac{P(E | h_1)}{P(E | h_2)}.$$

Therefore, the *evidence*, E , affects the comparison of hypotheses only via their likelihoods. Once the likelihoods are given, the detailed information contained in the data is no longer relevant. That is roughly the thesis stated by Barnard (1947, p. 659):

The connection between a simple statistical hypothesis H and observed results R is entirely given by the likelihood, or probability function $L(R|H)$. If we make a comparison between two hypotheses, H and H' , on the basis of observed results R , this can be done only by comparing the chances of, getting R , if H were true, with those of getting R , if H' were true.

If the likelihood of a hypothesis is viewed as a measure of fit with the data, then LTE says that the impact of evidence on hypothesis comparison depends only on how well the

hypotheses fit the total observed data. It is a surprising thesis, because it implies that the evidence relation between a simple hypothesis and the observed data, no matter how rich, can be captured by a single number—the likelihood of the hypothesis relative to the data.

The LTE extends naturally to the problem of comparing composite hypotheses, which are also called *models* in the statistics literature.² In a trivial case, a model M might consist of a family of two simple hypotheses $\{h_1, h_2\}$, while a rival model, M' , is the family $\{h_3, h_4\}$. For a Bayesian,

$$\frac{P(M | E)}{P(M' | E)} = \frac{P(M)}{P(M')} \times \frac{P(E | M)}{P(E | M')},$$

where the likelihoods $P(E | M)$ and $P(E | M')$, are calculated as averages over the likelihoods of the simple hypotheses in the respective families. Specifically,

$$P(E | M) = P(E | h_1)P(h_1 | M) + P(E | h_2)P(h_2 | M).$$

So, if models are compared by their posterior probabilities, $P(M | E)$ and $P(M' | E)$, then the bearing of the evidence, E , is still exhausted by the likelihoods of the simple hypotheses in each model. Note that the likelihood of a model is not well defined, except by specifying the prior probabilities, $P(h_1 | M)$ and $P(h_2 | M)$, which are usually not given by the model itself. Non-Bayesian statisticians, who avoid the use of prior probabilities, may use likelihoods differently while still subscribing to the LTE (see below). As a final remark about terminology, note that the set of likelihoods defines a mapping from the simple hypotheses in the model to likelihoods. This mapping is

² Terminology varies. In the computer science literature especially, a simple hypothesis is called a model and what I am calling a model is referred to as a model class.

standardly referred to as the *likelihood function* of the model. Likelihoods are always defined relative to a single set of observed data—the total actual evidence.³

It is now easy to formulate the LTE in a way that applies equally well to the comparison of simple hypotheses or models (composite hypotheses):

The Likelihood Theory of Evidence (LTE): The observed data are relevant to the comparison of simple hypotheses (or models) only through the likelihoods of the simple hypotheses (or the likelihood functions of the models).

LTE says nothing about how likelihoods are *used* in the comparison of hypotheses or models. Bayesians compare models by comparing average likelihoods. Non-Bayesians may compare *maximum* likelihoods adjusted by a penalty for complexity, as in Akaike's AIC statistics.⁴ Again, the data enters the comparison only via the likelihoods, so AIC conforms to LTE.⁵ The majority of model selection methods in the statistics literature, such as BIC (Schwarz 1978), Bayes factors (see Wasserman 2000) or posterior Bayes factors (Aitkin 1991), also conform to LTE. Standard model selection criteria are being lumped together for the purposes of this paper because they differ *only* in the *way* they use likelihoods.⁶

Contemporary statistics is divided into three camps; classical Neyman-Pearson statistics (see Mayo 1996 for a recent defense), Bayesianism (*e.g.*, Jefferys 1961, Savage

³ A peculiar thing about the quote from Barnard (above) is that he refers to the likelihood of a simple hypothesis as a probability *function*. It is not a function except in the very trivial sense of mapping a single hypothesis to a single number.

⁴ Akaike 1973, Sakamoto *et al.* 1986, Forster and Sober 1994, Burnham and Anderson 2002.

⁵ In contrast, the Law of Likelihood (LL) is very specific about how likelihoods are used in the comparison of simple hypotheses. Forster and Sober (2004) argue that AIC is a counterexample to LL. Unfortunately, Forster and Sober (2004) mistakenly describe LL as the likelihood principle, which was pointed out by Boik (2004) in the same volume. For the record, Forster and Sober (2004) did not intend to say anything about the likelihood principle—the present paper is the first publication in which I have discussed LP.

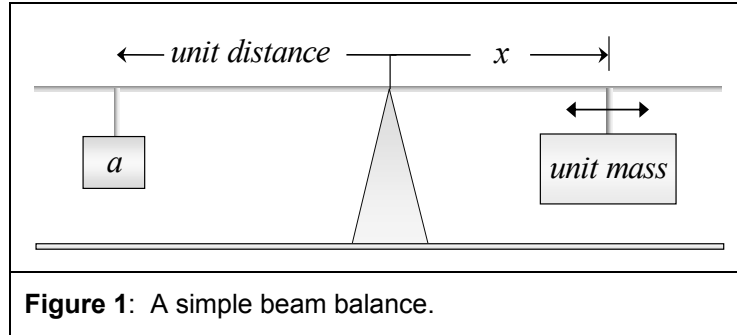
⁶ See Forster (2000) for a description of the best known model selection criteria, and for an argument that the Akaike framework is the conceptually clearest framework for understanding the problem of model selection because it clearly distinguishes criteria from goals.

1976, Berger 1985, Berger and Wolpert 1988), and third, but not last, Likelihoodism (*e.g.*, Hacking 1965, Edwards 1987, and Royall 1997). Likelihoodism is, roughly speaking, “Bayesianism without priors”, where I am classifying the Akaike “predictive” paradigm as a kind of Likelihoodism. Bayesianism and Likelihoodism, as they are understood here, are founded on the Likelihood Principle, which may be viewed as the thesis that LTE applies to the problem of comparing simple hypotheses under the assumption that the background model is true. If what can count as a “background model” is left vague, then the counterexamples to LTE are also counterexamples to the Likelihood Principle.

The Likelihood Principle has been vigorously upheld (*e.g.*, Birnbaum 1962, Royall 1991) in reference to its most important consequence, called Actualism by Sober (1993)—the reasonable doctrine that the evidential support of hypotheses and models should be judged only with respect to data that is actually observed. As Royall (1991) emphasizes in terms of dramatic examples, classical statistical practice has sometimes violated Actualism, and sometimes in the face of very serious ethical issues. But the likelihood principle has other consequences besides Actualism, and these might be false. Or, put another way, a theory of evidence may deny the Likelihood Principle, without denying Actualism. Actualism is strictly adhered to in all the examples discussed in this paper.

Section 2 describes what a fit function is, and introduces the idea of a fit-function principle. Likelihood is described as a measure of fit in Section 3, and relationship between the Likelihood Principle and LTE is discussed there. The two sections following

that present counterexamples to LTE, first in terms of an example with continuous variables (a simple curve fitting problem) and then in terms of binary (yes-no) variables.



2. Fit Functions

Consider a beam balance device (Fig. 1) on which an object a of unknown mass, θ , is hung at a unit distance from the fulcrum. Then the position of the unit mass on the right is adjusted until the beam balances. The experiment can be repeated by taking a off the beam and beginning again. Each repetition is called a trial of the experiment. One can even change the “unit distance” between trials, provided that x is always recorded as a proportion of that distance. In order to experimentally measure the values of postulated quantities, like θ , they must be related to observed quantities, in this case, the distance, x , at which the unit mass is hung to balance the beam.

In accordance with standard statistical notation, let X denote the distance variable while x refers to its value on a particular trial. The outcome of the first trial might be $X = 18$. The outcome of the next trial might be $X = 19$. It is implausible that the outcomes of a continuous quantity turn out to have integer values (or it be could that the device has a kind of ratchet system that disallows in-between values). X is variable because it can vary from one trial to the next. θ is not variable in this sense because its value does not change between trials, even though its *estimated* value may change as the data accumulate. To mark this distinction, θ is referred to as an *adjustable parameter*.

The standard Newtonian equation relating θ and X turns out to be very simple: $X = \theta$, where θ is an adjustable parameter constrained to have non-negative values ($\theta \geq 0$). A *model* is a set of equations with at least one adjustable parameter. The model in this case is an infinite set of equations, each one assigning different numerical values to θ . A simple *hypothesis* in the model has the form $\theta = 25$, for instance, and the model is the family of all simple hypotheses.

Now do the experiment! We might find that the recorded data in four trials is a sequence of measured X values (18,19,21,22), so the model yields four equations:

$$\theta = 18, \theta = 19, \theta = 21, \theta = 22.$$

Sadly, the data is logically inconsistent with the model; that is, the data falsifies every hypothesis in the model. Should we all go home? If we lower our sights from truth to predictive accuracy, then perhaps not. Some hypotheses in the model definitely do a better job at predicting the data than others. $\theta = 20$ does a better job than $\theta = 537$. Maximizing predictive accuracy is worthwhile, and who knows, some deeper truth-related virtues will also emerge out of the morass.

Definitions of degrees of fit are found everywhere in statistics. For example, the Sum of Squares (SOS) Fit Function in this example is:

$$F(\theta) = (\theta - x_1)^2 + (\theta - x_2)^2 + \dots + (\theta - x_N)^2,$$

where the data is (x_1, x_2, \dots, x_N) . It assigns a degree of fit to every simple hypothesis in the model. Or we could introduce the 0-1 Fit Function that assigns 1 to a hypothesis if it fits perfectly, and 0 otherwise. The SOS function measures badness-of-fit because higher values indicate worse fit, whereas the 0-1 function measures goodness-of-fit. But this is an irrelevant difference because we can always multiply the SOS function by -1 .

The SOS function addresses the problem of prediction when data are “noisy” or when the model is misspecified (i.e., false) . For example, the toy data tell us that the hypothesis $\theta = 20.0$ best fits the observations according to the SOS definition of fit. The minimization the SOS fit function provides a method for estimating the value of theoretical parameters known as the *method of least squares*. Once the best fitting hypothesis is picked out of the model, it can be used to predict unseen data, and the predictive accuracy of the model can be judged by how well it does.⁷

A fit function can be used to address a variety of inferential problems. This part of the story becomes complicated because the different schools of statistical thought (classical Neyman-Pearson statisticians, Bayesians, and Likelihoodists) provide their own methods of statistical inference. I want to abstract away from these controversial particulars and address the more basic question: To what extent is the notion of fit relevant to scientific reasoning? The question is not about whether one definition of fit is better than another. The question is whether any kind of fit do the work usually asked of it in a theory of evidence.

3. The Likelihood Principle

Likelihood is usefully understood as a measure of fit.

Definition: The *likelihood* of a hypothesis (relative to data x) is equal to the probability of x given the hypothesis (not to be confused with the Bayesian notion of the probability of a hypothesis given the data).

⁷ The term ‘predictive accuracy’ was coined by Forster and Sober (1994), where it is given a precise definition in terms of SOS and likelihood fit functions.

Clearly, the likelihood is defined only for hypotheses that are probabilistic. As an illustration, the beam balance model can be turned into a family of probabilistic hypotheses by associating each hypothesis with an error distribution:

$$X = \theta + U ,$$

where U has a normal, or Gaussian, distribution with mean zero and unit variance (according to the model). If we replace the adjustable parameter by a particular number, then we obtain a simple hypothesis in the model, which defines a precise probability density for x (in this example, it implies that the distribution is Gaussian with mean θ and variance 1; note that the model also assumes that different trials of the experiment are probabilistically independent).

Given that the measured value of X is a point value, the likelihood of a datum is zero, strictly speaking, because a beam balance hypothesis assigns only a probability *density* to a point value. This technical problem is finessed by defining likelihood as proportional to the probability that the datum is in the interval from x to $x+k$, where k is sufficiently small. This probability is equal to the probability *density* at x times k . If likelihoods of different hypotheses are compared to the same data, then the value of k , although arbitrary, will be the same for both hypotheses. So in the context of hypothesis comparison, where the likelihoods are always relative to the same set of data, it is not arbitrary to claim that two hypotheses have the same likelihood or that two models have the same likelihood functions.

Berger (1985, p. 28.) states the Likelihood Principle in the following way: “In making inferences or decisions about θ after x is observed, all relevant experimental information is contained in the likelihood function for the observed x .” The first point is

that making decisions about θ is the same as making decisions about simple hypotheses in the model because there is a one-to-one correspondence between simple hypotheses and point values of θ .

Berger continues: “Furthermore, two likelihood functions contain the same information about θ if they are proportional to each other (as functions of θ).” This claim can be understood in terms of the beam balance example, or a slight modification of it. Suppose that in addition to the beam balance data, $x = (18, 19, 21, 22)$, we also recorded the outcome of a coin toss, which lands heads. We might record the expanded data as $((18, 19, 21, 22), H)$. Further suppose that there are two beam balance models, both agreeing on the stochastic equation $X = \theta + U$, and agreeing that the coin toss is probabilistically independent of other event, but disagreeing about the probability of the outcome H . Then each of the models will assign different probabilities to the total data, but their respective likelihood functions will differ only by a constant. Both models should make the same inferences about θ because they contain the same information about θ . In more technical jargon, x is a *sufficient statistic* for θ , and the two models have exactly the same likelihood function with respect to x .

Berger and Wolpert (1988, pp. 19-21) add the following caveat to their version of the Likelihood Principle: “...it only applies for a fully specified model... If there is uncertainty in the model, and if one desires to gain information about which model is correct, that uncertainty must be incorporated into the definition of θ .” In the previous example, suppose that we are unsure about the probability of the event H , so we are uncertain about which of the two models is true. Berger and Wolpert might be saying

something like this:⁸ Let h_1 be the hypothesis that says that the probability of H is $\frac{1}{2}$, while h_2 says that the probability of H is almost 0, say .0000001. Let M_1 be the beam balance model conjoined with h_1 , while M_2 is the beam balance model conjoined with h_2 . Simple hypotheses in the models, such as $h_1 \ \& \ (\theta = 17)$ and $h_2 \ \& \ (\theta = 21)$, can be coded in the parameters by writing $\theta_1 = 17$ and $\theta_2 = 21$, respectively. It is now clear that the likelihood functions for M_1 and M_2 are different despite the fact that the likelihood functions for θ are equivalent *for the purpose of estimating values of θ* . This helps block a simple-minded argument against the Likelihood Principle, which goes something like this: We can tell from the total evidence that M_2 is false because all simple hypotheses in M_2 assign a probability of almost 0 to the outcome H . But the two models are likelihood equivalent because their likelihood functions differ only by a constant. Agreed! This argument is wrong.

This much seems clear: Most Bayesians, if not all, think that in order to gain information about which of two models is correct, it is at least *necessary* for there be *some* difference in the likelihood functions of the models. For if the likelihoods functions of two models were exactly the same, the only way for the posterior probabilities to be different would be for the priors to be different, but a difference in priors does not count as *evidential* discrimination. This is the assumption that I have referred to as the Likelihood Theory of Evidence (LTE).

Whether the different published versions of the Likelihood Principle presuppose the universal truth of LTE is not always clear. Nevertheless, philosophers and

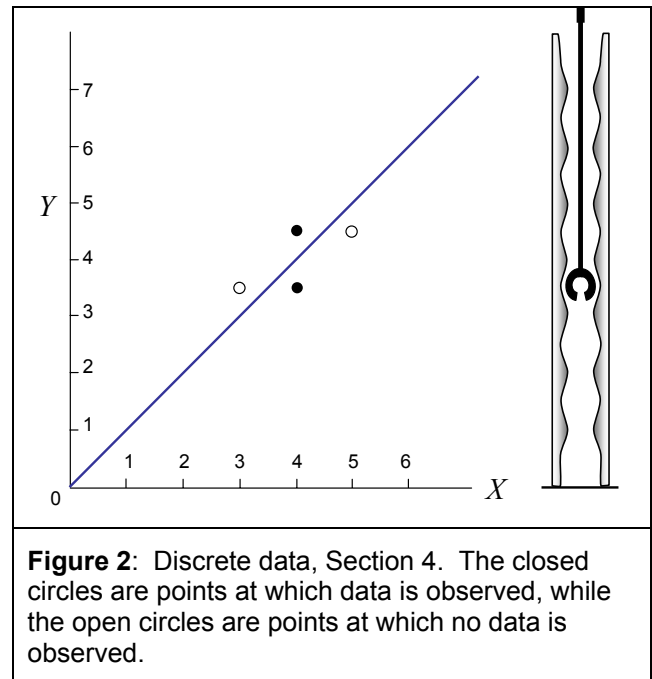
⁸ I owe this suggestion to Jason Grossman.

statisticians who claim that simple hypotheses or models should always be compared by their posterior probabilities are committed to LTE.

4. Preliminary Examples

When a mass is hung on a spring, it oscillates for a period of time and comes to rest. After the system reaches an equilibrium state, the spring is stretch by a certain amount;

let's denote this variable by Y . To simplify the example, suppose that Y takes on a discrete value $\frac{0}{2}, \frac{1}{2}, \frac{2}{2}, \dots, \frac{14}{2}, \frac{15}{2}$, because in-between positions are not stable. Maybe this is because the motion of the device is constrained by a weightless "ball" attached at the bottom moving inside a lubricated cylinder with a serrated surface (see Fig. 2, right). The mass on the spring consists of a number of identical pellets (*e.g.*, coins). This number is also an observed quantity—denoted by $X = 1, 2, 3, \dots$



Conduct 2 trials of the experiment, and record the observations of (X, Y) : Suppose they are $(4, 3.5)$ and $(4, 4.5)$. The data are represented by the solid dots in Fig. 2. Now consider the hypothesis $A: Y = X + U$, where U is a random error term that takes on values $-\frac{1}{2}$, or $\frac{1}{2}$, each with probability $\frac{1}{2}$. Strictly speaking, we should introduce a different set of variables for each trial of the experiment: $Y_i = X_i + U_i$, for $i = 1, 2$, where the random variables U_i are mutually independent and identically distributed (i.i.d.).

This detail will become important later; in the meantime we shall use $Y = X + U$ as a way of referring to an arbitrary trial of the experiment.

To understand what follows, it is important to understand the meaning of a stochastic equation like $Y = X + U$. The fundamental assertion is that U is a random variable, which means that possible events $U = u$ are assigned a probability value by the hypothesis. Always remember that being a *random* variable is not a god-given property of a variable—it is a status accorded by the hypothesis under consideration. Since U is a random variable, so is $Y - X$ (by the equation). But it does it follow that X and Y are random variables? There are three cases to consider.

Case (1): X is just an ordinary variable with no probability distribution associated with it. As a variable, it can a particular value, say x . Since x is a just a number, it follows that $x + U$ is a random variable. So $X + U$ maps possible values of X to random variables. In a sense, we might think of Y as a random *function* rather than a random variable, written $Y(X)$. Nevertheless, a conditional probability like $P(Y = y | X = x)$ would appear to be unambiguous because a unique random variable, $Y(x)$, is picked out. But this may be misleading, for this conditional probability cannot be obtained from the Kolmogorov definition of conditional probabilities, because $P(X = x)$ has no meaning. It is better to write $P_{X=x}(Y = y)$, provided that it is understood correctly.

Case (2): We could write the equation $X = Y - U$ and treat Y as an ordinary variable, in which case, X is a random function, and the hypothesis provides “conditional probabilities” $P_{Y=y}(X = x)$. This not the intended interpretation of hypothesis A , and it is worth saying more. When writing $Y = X + U$, it is assumed that X is the independent (or exogenous) variable, while Y is the dependent (or endogenous variable). This has real

consequences in the context of stochastic equations, for I take it to imply that if ordinary variables are involved in the equation, then it is the exogenous variable (X in this example). This convention implies that case (2) does not apply to hypothesis A —it gives the hypothesis a content that is asymmetric between X and Y , as is appropriate in causal modeling.

Case (3): X is a random variable. That is, probabilities $P(X = x)$ are given by the hypothesis. This is not sufficient to make Y a random variable—one needs a joint probability distribution $P(X = x, U = u)$ as well. Once that is specified, then $P(X = x, Y = y)$ is well defined, and the distribution of U is derivable from its defining equation $U = Y - X$. In causal modeling, it is standardly assumed that U is probabilistically independent of the exogenous variable X : Given this assumption, then it is sufficient to specify the probabilities $P(X = x)$ to obtain the joint distribution $P(X = x, Y = y)$. Another way of doing this would be to add $P(X = x)$ to the conditional probabilities $P_{X=x}(Y = y)$ in Case (1):

$$P(X = x, Y = y) \triangleq P(X = x)P_{X=x}(Y = y) \quad (*)$$

It is interesting to ask whether these two methods are equivalent. The answer is yes, by the following argument. First note that

$$P_{X=x}(Y = y) = P_{X=x}(U = y - x) = P(U = y - x).$$

By (*), $P(X = x, Y = y) = P(X = x)P(U = y - x)$.

But $P(X = x, Y = y) = P(X = x, U = y - x)$.

Therefore, $P(X = x, U = y - x) = P(X = x)P(U = y - x)$, for all y . This proves, for all u ,

$$P(X = x, U = u) = P(X = x)P(U = u),$$

which is what we wanted to show. This is conceptually revealing—the mysterious independence between exogenous variables and the error term is explained by first interpreting the hypothesis as in Case (1), and then *assuming* that

$$P(Y = y | X = x) = P_{X=x}(Y = y).$$

Returning to our example, Y is a function of U , and U is a random variable. But what is the status of X ? In Case (1), the number of pellets making up the mass is not usually thought of as having a probability. The problem is that if X has no probability distribution associated with it, then hypothesis A has no likelihood relative to the *total* evidence, and so the likelihood theory of evidence (LTE) does not apply.

What happens if we use the conditional likelihoods, which are defined? In our example, the conditional likelihood of hypothesis A is equal to

$$L(A) = P_{X_1=4}(Y_1 = 3.5)P_{X_2=4}(Y_2 = 4.5) = \frac{1}{4}.$$

Now compare this with an alternative hypothesis B (for Backwards) with equations $X_1 = Y_1 + U_1$ and $X_2 = Y_2 + U_2$, where U_1 and U_2 are error terms that are identically distributed to those postulated by A . That are also treated in accordance with Case (1); this time B assigns no probabilities to the Y variables. It is easy to see that the conditional likelihood of B is also equal to $\frac{1}{4}$. $L(A) = L(B)$. So, there is no way of distinguishing between A and B the hypotheses in terms of conditional likelihoods.

This is significant because the two hypotheses *can* be distinguished on the basis of the data. First note that no matter how times we duplicate the observed data, the conditional likelihoods will remain the same. Concretely, suppose that the data points (4,3.5) and (4,4.5) are observed 10 times each, as would be expected if A were true. But

it tells us immediately that B is false. Why? Let me explain the point in a way that generalizes easily to other examples. Hypothesis B entails a *constraint*:

Constraint:
$$P_{Y=3.5}(X = 4) = P_{Y=4.5}(X = 5).$$

(Both probabilities are equal to $P(U = \frac{1}{2})$.) But the data show that $P_{Y=3.5}(X = 4)$ is close to 1 while $P_{Y=4.5}(X = 5)$ is close to 0. In other words, two independent measurements of $P(U = \frac{1}{2})$ that not only disagree with the hypothesized value ($\frac{1}{2}$), but also disagree with each other.

The example is already a counterexample in the following sense: We are told that either A or B is true, and we can tell from the data that A is true and B is false. But there is nothing in the *likelihoods* that distinguishes between them.

Perhaps a subscriber to LTE might deny that LTE applies to hypotheses that are incomplete in this sense.⁹ They might insist that the example violates the spirit of the principle of total evidence, even though here are no data “hidden from view”, or withheld in any way.

In any case, it is not difficult to modify the example so that the full likelihoods are well defined. We must first recognize that each trial of the experiment is modeled in terms of its own set of variables, so the equation for trial i is $X_i = Y_i + U_i$, where these variables do not appear in other equation. The only constraint that B postulates between different trials is that error terms, U_i , are independent and identically distributed (i.i.d.). If we add probability distributions for the exogenous variables Y_i , then there is no rule that they must be identically distributed. They might be constrained in other ways, or

⁹ The problem is the same one discussed in Forster 1988b.

they might be entirely unconnected. So, consider the augmented hypothesis, call it B' , that says that $P(Y_i = y_i) = 1$, for all i , where y_i happens to be the observed value of the variable in trial i . Likewise, consider the hypothesis A' that adds the assumption that $P(X_i = 4) = 1$, for all i . These are real hypotheses that are 100% consistent with probability theory. Now we are told that either A' or B' is true. Can we tell which one from the data? Yes, in the same way as before— B' is false because it logically entails B , and B is false. But does this now up in the likelihoods? No! Because the relative likelihoods of the two hypotheses have not changed.

One reaction might be to object that the hypotheses are ad hoc, or that they must have been constructed with a knowledge of the data. This won't save LTE because it denies the relevance of extra-empirical, historical, or psychological considerations (so do I, by the way). Besides, it is *very* clear from this example that there is no need to appeal to extra-logical considerations! We can tell which hypothesis is false from the data alone! It's not a counterexample to all logical theories of evidence. It's a counterexample to likelihood theories of evidence.

Why are likelihood theories of evidence so popular? Success can always be backwards-engineered by restricting one's attention to the right class of hypotheses. This is common practice in the field of Bayes nets (see for example Pearl 2000), and our running example provides a nice illustration of how it works. A sufficient condition for success is to first augment A and B with probability distributions that are identically distributed. Of course, it is still possible to add distributions so the B beats A , *but now it's possible to blame the poor fit of the marginal probabilities*. To demonstrate the effect, let's add the best i.i.d. marginal distributions possible—namely, that ones that fit

the marginal data the best. Then poorness of fit cannot be blamed. In our example, we need to add to A , $P(X_i = 4) = 1$, for all i , resulting in hypothesis A'' . Clearly this makes not difference to the likelihood: $L(A'') = L(A)$. To B , we add

$P(Y_i = 3.5) = \frac{1}{2} = P(Y_i = 4.5)$. Now $L(B'') = (\frac{1}{2})^{20} L(A'')$. So, B'' is less likely than A'' , which is the right answer. The mystery is: Why are we adding things to A and B instead of comparing them against the data directly when we know that it works? In order to get the right likelihoods? In order to *make* the LTE work?

In the examples just considered, two hypotheses are compared against a single data set. Philosophers of science also consider questions about the comparative impact of two hypothetical data sets on a single hypothesis. Is likelihood the right measure of comparison in this case? Label the previous data set E , and consider B'' . B'' is just an everyday hypothesis that assigns each of four data points (in Fig. 2) a probability of $\frac{1}{4}$. We have already seen that B'' is clearly refuted by E , since all the data are concentrated on two of the four points (the solid dots in Fig. 2).¹⁰ Compare this to an equally large set of data E'' that is evenly spread amongst all four points (still with 20 data points in total). My intuition says that E'' confirms B'' better than E confirms B'' . After all, E is inconsistent with B'' , whereas E'' conforms to the B'' as well as any data set imaginable (of that size). Right?! Not according the theory of confirmation advocated by Bayesian philosophers of science! For the probability of E'' given B'' is the same as the probability of E given B'' ; both are equal to $(\frac{1}{4})^{20}$.

¹⁰ While the refutation is not refutation in the strict logical sense, the number of data in the example can be increased to whatever number you like, so it becomes arbitrarily close to that ideal.

Let's back up a little. Bayesian philosophers of science say that E confirms hypothesis H if and only if $P(H | E) > P(H)$. They also say that E' would confirm H better than E confirms H if and only if $P(H | E') - P(H) > P(H | E) - P(H)$.¹¹ So, if the Bayesian theory is to match our intuitions in this example, then $P(B'' | E'') > P(B'' | E)$. But, since $P(E'' | B'') = P(E | B'')$, that can only happen if $P(E) > P(E'')$. It is strange to me that objective facts about confirmation should ever depend on how surprising or how improbable the evidence is, but let's leave that to one side. It is certainly possible to place the example in a historical context in which $P(E) \leq P(E'')$, or one in which $P(E)$ is *much* less than $P(E'')$. In the latter case, Bayesians are forced to say that B'' is much better confirmed by E than by E'' . But that conclusion is absurd in this example!

This issue is well known to Bayesian statisticians, and to some philosophers of science: For example, the hypothesis that a coin is fair assigns the same probability to a string of 100 heads as it does to a random 50-50 sequence of heads and tails. The usual lesson drawn from this is that confirmation is comparative—alternative hypotheses about the coin's bias will assign different probabilities to the two data sets, so that the likelihoods ratio between two hypotheses is still meaningful. This is why statisticians formulate the Likelihood Principle in the context of comparing simple hypotheses in a model. And that is why the previous examples were comparative in nature. What I've tried to show is that even likelihoods ratios are sometimes unreliable indicators of evidential relevance.

¹¹ Fitelson (1999) shows that choice of the difference measure does matter in some applications. But that issue does not arise here.

5. The Asymmetry of Regression

The same challenge to LTE extends to the linear regression problem ('regression' is the statistician's name for curve fitting). In these examples, you may be told that one of the two hypotheses or models is true, and you are invited to say which one is true on the basis of the data. They are examples in which anyone can tell from the full data (with moral certainty) which is true, but nobody can tell from a knowledge solely of the likelihoods. It is not because Bayesians, or anyone else, are using likelihoods in the wrong way. It's because the relevant information is not there!

Suppose that data are generated by the 'structural' or 'causal' equation $Y = X + U$, where X and Y are observed variables, and U is a normal, or Gaussian, random variable with mean zero and unit variance, where U is probabilistically independent of X .¹² To use this to generate pairs of values (x, y) , we must also provide a generating distribution for X , represented by the equation $X = \mu + W$, where μ is the mean value of X , and W is another standard Gaussian random variable, probabilistically independent of U , also with zero mean and unit variance. Two hundred data points are shown in Fig. 3. The vertical bar centered on the line $Y = X$ represents the probability density of y given a particular value of x .

Example 1: There are two indistinguishable ways of generating these data. The Forward method randomly chooses an x value, and then determines the y value by adding a Gaussian error above or below the Forward line ($Y = X$). This is the method described in the previous paragraph. The Backward method randomly chooses a y value according

¹² Causal modeling of this kind has received a great deal of attention in recent years. See Pearl (2000) for a comprehensive survey of recent results, as well as Woodward (2003) for an introduction that is more accessible to philosophers.

to the equation $Y = \mu + \sqrt{2} Z$, where Z is a Gaussian variable with zero mean and unit variance. Then the x value is determined by adding a Gaussian error (with half the variance) above or below the Backward line $Y = 2X$ (slope = 2). This probability density represented by the horizontal bar centered on the line $Y = 2X$ (see Fig. 3). In this case, the ‘structural’ equation is $X = \frac{1}{2}\mu + \frac{1}{2}Y + \frac{1}{\sqrt{2}}V$, where V is standard Gaussian (mean zero and unit variance) such that V is probabilistically independent of Y . It is impossible to tell which method was used from the data alone.

Example 1 is not a counterexample to the likelihood theory of evidence (LTE). As it applies to this example, LTE says that if two simple hypotheses cannot be distinguished on the basis of their likelihoods (let’s say that they are *likelihood equivalent*) then they cannot be distinguished on the basis of the full data. Why are the two hypotheses likelihood equivalent? The

Forward hypothesis first specifies a probability distribution for an x value, which we write as $p_F(x)$, and then specifies the probability of a y value given that x value; in symbols, $p_F(y|x)$. This determines a joint probability distribution $p_F(x,y) = p_F(x)p_F(y|x)$. The Backward hypothesis determines

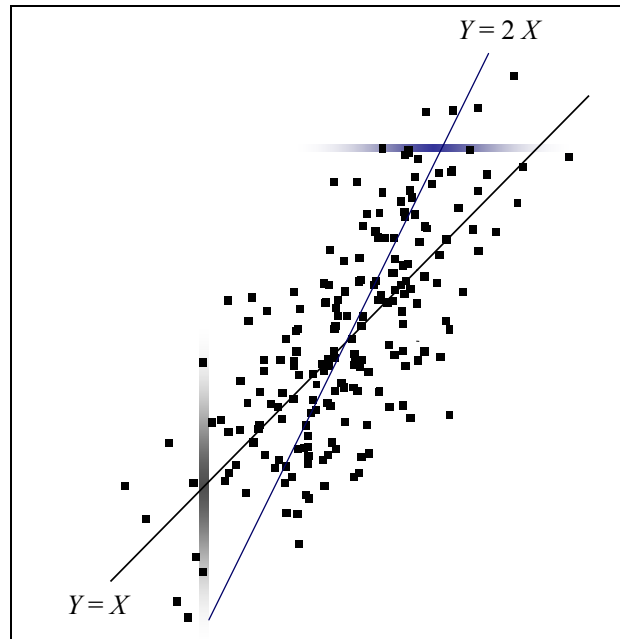


Figure 3: There are two ways of generating the same data: The Forward method and the Backward method (see text). It is impossible to tell which method was used from the data alone.

a joint distribution for x and y of the form $p_B(x, y) = p_B(y)p_B(x | y)$. Under the stated conditions, it is possible to prove that for all x and for all y , $p_F(x, y) = p_B(x, y)$. So, they cannot be distinguished in terms of likelihoods. But they also cannot be distinguished by the full data, which is why it is not a counterexample to LTE.

Example 2: Now consider the comparison of two simple hypotheses that are not likelihood equivalent with respect to the same data (Fig. 3). This is not a counterexample to LTE either, but it does raise some important worries, which will be exploited in the examples that follow. Let us specify the two hypotheses more concretely by assuming that $\mu = 0$, where we also assume that the data are centered around the point $(0,0)$. We are told that one of two hypotheses are true:

$$F_2 : \quad Y = X + U \text{ and } X = W .$$

$$B_2 : \quad X = Y + V \text{ and } Y = \sqrt{2}Z ,$$

where again W , U , and V are mutually independent Gaussian random variables with zero mean and unit variance, such that U is independent of X and V is independent of Y . F_2 is the same hypothesis as in Example 1. The difference is in the Backwards hypothesis.

The marginal y values are generated in the same way as before, but now B_2 says that the x value are generated from the line $Y = X$ (rather than $Y = 2X$) using a Gaussian error of mean zero and unit variance (as opposed to a variance of $1/2$). It is intuitively clear that B_2 will fit the data worse (have lower likelihood) than the previous hypothesis, B_1 .

What's remarkable in the present case is that, when compared to F_2 , the lower likelihood of B_2 arises not from its generation of x values, but from the fact that there is larger variation in y values than in the x values. This is strange because we intuitively regard

the generation of the “exogenous” variable to be an inessential part of the causal hypothesis.

To demonstrate this phenomenon, consider an arbitrary data point (x, y) . From the fact that F_1 and B_2 generate the y and x values, respectively, from the line $Y = X$, and the fact that an arbitrary point is equidistant from this line in *both* the vertical and the horizontal directions, it follows that $p_F(y | x) = p_B(x | y)$. For the benefit of the technocrats amongst us, both are equal to $(2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}(x-y)^2}$.

For each hypothesis, the likelihood is obtained by multiplying the probabilities of the data points together (assuming that the data points are mutually independent), where each probability has the form:

$$P_F(x < X < x + dx, y < Y < y + dy) = k p_F(x, y),$$

$$P_B(x < X < x + dx, y < Y < y + dy) = k p_B(x, y),$$

where $k = dx dy$. Furthermore,

$$p_F(x, y) = p_F(x)p_F(y | x), \text{ and } p_B(x, y) = p_B(y)p_B(x | y).$$

Therefore, B_2 has a lower likelihood than F_2 because the marginal probability density of y is lower. This is indeed the case because the variance of Y is twice the variance of X , which flattens out the distribution for y .

This is odd because the specification of marginal probabilities is not what we think of as the essential content of a ‘causal’ hypothesis. The falsity of B_2 is apparent from the *pattern* that forms when x values generated from y values, *even if we only look at data with the same y value*. The x values do not vary randomly to the left and to the right of the line $Y = X$, as B_2 claims. Instead, they vary randomly to the left and the right

of the line $Y = 2X$ with half the variance, just as would expect if B_1 were true. This is easily seen by plotting residuals $(x - y)$ against y (see Fig. 4). The residual variance is equal to 1 because it is sum of two terms— one due to the deviation of the line $Y = 2X$ from the line $Y = X$ (the ‘explainable’ variation) and the other due to the smaller random variation about the line $Y = 2X$. In contrast, if we were to plot the y residuals against x , then there would be no discernible correlation between the y residuals and x . The fact that the marginal variance of the y values is twice the marginal variance of the x values seems to be irrelevant to the evidential comparison, so it is strange that it should play such a pivotal a role in the likelihood comparison.

If the competition were between B_1 and B_2 , then B_1 would win the likelihood comparison and there would be nothing strange about this because it would successfully explain the anti-correlation in Fig. 4. All we have done in this example is to replace B_1 with its the likelihood

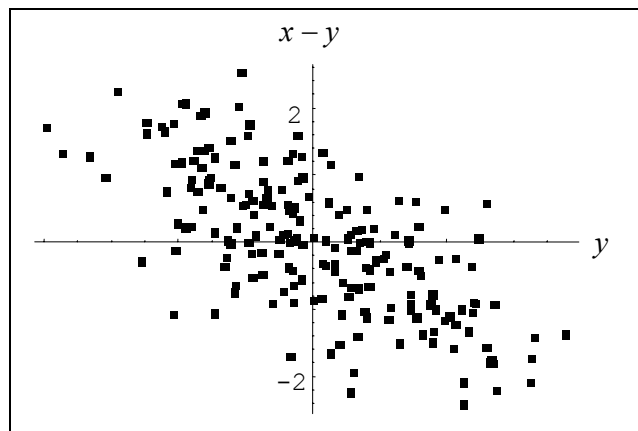


Figure 4: The x residuals $(x - y)$ plotted as a function of y . The residual has an average variation of 1, but the variation varies in a systematic way for different values of y .

equivalent hypothesis F_2 , which also explains the anti-correlation.

Example 3: We will have a counterexample to LTE if the competing hypotheses can be constructed so that the marginal components of their likelihoods are the same. Suppose that we are told an alternative story about how the marginal values are generated. According to this version of Forward hypothesis, the x values are read from a predetermined list of values, and then a slight stochastic element imposed on the final

value by randomly generating a small error from a uniform distribution of (small) width δ around the listed value. If we define $\varepsilon = 1/\delta$, then $P_F(x) = \varepsilon dx$ for all x . Similarly, B asserts that y values are first drawn from a list and then randomized in the same way, so that $P_B(y) = \varepsilon dy$. The story about how the second variable is generated is the same as before, in each case. Similarly, we are told that either F or B is true. Can we tell which one? Yes, by looking at the behavior of the residuals (as explained in Example 2). Can we tell if we are just given the likelihoods? No, because the likelihoods are the same (remember that we are only told the likelihoods, and not details of the calculation, which could give away information about the original data). So, this is a clear-cut counterexample to LTE.

The Backward hypothesis is derived from B_2 by changing the story about how the “exogenous” variable is generated. If we were to replace this hypothesis instead with a variant of B_1 , with the same story about the exogenous variable, then the Backward and Forward hypotheses would be genuinely indistinguishable on the basis of the full data—either one could have generated the data. Yet, in this case, the Backward hypothesis would have the higher likelihood! This does not contradict LTE because I have formulated it in a way that is completely neutral about *how* likelihoods are used. Nevertheless, it is a counterexample to the Law of Likelihood (Hacking 1965, Royall 1997), which claims that evidence E supports A better than B or is stronger evidence for A than for B if and only if the likelihood of A is greater than the likelihood of B .

Example 4: An interesting variation of Example 3 makes only one change. Instead of a list of x values that are distributed in a Gaussian way around a central value ($x = 0$), suppose that list comprising of two clusters— 200 values distributed around $x = -10$ with an apparently Gaussian distribution, and a list of 200 x values centered around $x = +10$ with a similar distribution. The y values are generated in the same way as before. This is

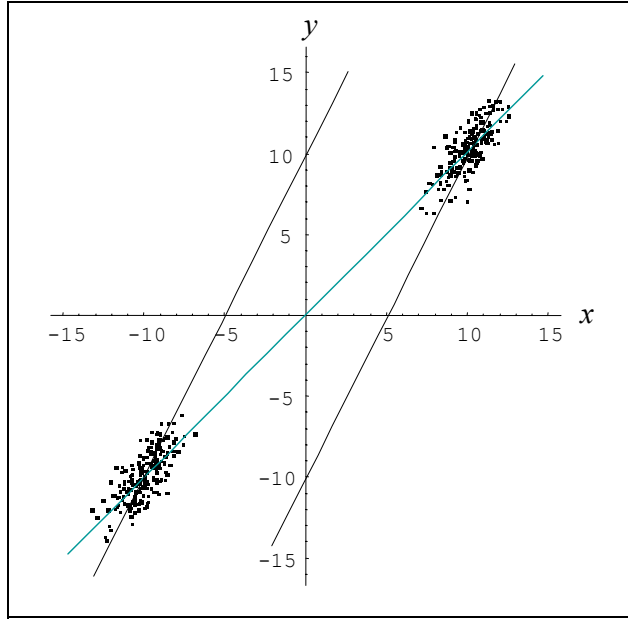


Figure 5: The asymmetry of regression. If a regression analysis is performed on the two clusters of data separately, then the Forward regression lines will coincide. But the Backward regression lines are very distinct.

hypothesis F . The data, which are actually generated according to F , are shown in Fig. 5. B is the false hypothesis that is analogous to that in Example 3, with the only difference being the obvious one, that the list of y values now form two clusters, one centered at $y = -10$ and the other at $y = +10$. If we are given the full data, then we can tell that B is false by looking at the residuals, as before. So this is also a counterexample to the Likelihood Theory of Evidence (LTE).

Example 5: Example 4 is easily turned into an example *model* comparison. Construct the *models* F and B by the ‘causal’ equations $Y = \alpha + \beta X + \sigma U$ and $X = a + bY + sZ$, respectively, where U and Z are standard Gaussian, U is probabilistically independent of X , and Z is independent of Y . F and B are models because the equations contain adjustable parameters. The marginal distributions for each

trial are added in the same way as in Example 4—they don't introduce any adjustable parameters, even though the distributions vary from one trial to the next (they are not identically distributed). F is true because one of the simple hypotheses in F is true: Suppose that the data are generated by $Y = \frac{10}{\sqrt{101}}X + U$. This choice of coefficients ensures that the variances of X and Y in the data are the same when the data are clustered around $X = -10$ and $X = +10$, as shown in Fig. 5. With respect to the single-clustered data (Fig. 3), the likelihood functions of models F and B are not equal. But, with respect to the data in Fig. 5, the maximum likelihoods of each model are now the same, which means that the likelihood of any hypothesis in one model can be matched by the likelihood of a hypothesis in the other model (see the Appendix for the proof). In other words, the two models are equally good at accommodating the total data. But they are predictively very different, as we are about to show.

To complete the argument, we need only explain how the data (in Fig. 4) tell us which model is true. One way would be to show that *every* hypothesis in B is false by plotting the residuals, as explained in Example 2. But there is an easier way...

The idea is to fit each model to the two clusters of data separately and compare the independent estimates of the parameters obtained from the best fitting curves. I will describe this in a way that is reminiscent of the “test of hypotheses” that William Whewell called the *consilience of inductions* (Whewell 1858, 1989). Let X_1 and Y_1 refer to the cluster of data on the lower left of Fig. 5, while X_2 and Y_2 refer to the cluster on the upper right. Then F can be rewritten in terms of two stochastic equations,

$$Y_1 = \alpha_1 + \beta_1 X_1 + \sigma U_1, \text{ and } Y_2 = \alpha_2 + \beta_2 X_2 + \sigma U_2, \text{ plus two constraints } \alpha_1 = \alpha_2 \text{ and}$$

$\beta_1 = \beta_2$.¹³ The two stochastic equations are not rival models; they are parts of the same model (let's call them submodels). This way of writing the model makes no essential changes—it is just a different way of describing the same family of probability distributions. If we fit the submodels to their respective clusters of data, we obtain independent estimates of the parameters from the best fitting lines, which we can then use to test the constraints.

The results will be as follows. Using the data in Fig. 5, the independent measurements of the F parameters will agree closely (by any statistical criterion). But the B model will *fail* the same test. To see this, rewrite B as $X_1 = a_1 + b_1 Y_1 + sV_1$, and $X_2 = a_2 + b_2 Y_2 + sV_2$, plus the constraints $a_1 = a_2$ and $b_1 = b_2$. The constraint $b_1 = b_2$ will be verified, but the constraint $a_1 = a_2$ is not close to being true. As shown in Fig. 5, the estimated values are approximately $a_1 = -10$ and $a_2 = +10$. No statistical analysis can conclude that these are independent measurements of a single quantity. The data shows plainly that B is the false model, and therefore LTE is false.

To put the point another way, B is false because it fails to *predict* features of one cluster of data from the rest of the data. When we fit B to the lower data cluster, we get a backwards regression curve that approximates the line $Y = -10 + 2X$ (the steep line on the left in Fig. 5). Recall from Example 1, and Fig. 3, that this is the line from which B could have generated the lower data without us being able to tell. But we can tell that it did not generate the upper cluster of data—because the line does not pass anywhere near the points in the upper right quadrant. B fails at this kind of cross-situational prediction,

¹³ The word ‘constraint’ is borrowed from Sneed (1971), who introduced it as a way of constraining submodels. Although the sense of ‘model’ assumed here is different from Sneed’s, the idea is the same.

even though it is able to *accommodate* the full data perfectly well. The Likelihood Theory of Evidence fails because likelihood functions merely determine degrees of accommodation, not prediction.

6. The Reversibility of Binary Variables

Not all philosophers are comfortable with the mathematics of continuous variables, so this section includes an analogous example that uses only binary variables. Those reader who are already satisfied with the examples in the previous section will lose nothing by skipping to Section 6, except perhaps a minor excursion into the fascinating world of statistical mechanics.

All the models discussed in this section supervene on the same underlying microphysics, which I shall describe in detail. A free particle with a fixed energy is placed in a cubical box whose sides are 1 meter in length. Choose an x - y - z coordinate frame so that the x -direction is perpendicular to two faces of the box. An internal wall is placed in the middle, at the position $x = \frac{1}{2}$. If the particle is in the left partition at time t , then we record the event as X , and otherwise as \bar{X} . Collisions with any wall are elastic, which means the particle is reflected off the wall with the same speed.

The device operates in the following way. At time t , the central dividing wall is removed, and then exactly one second after that, the x -faces of the box are removed, and the particle is detected exiting the box to the left (recorded as Y) or to the right (recorded as \bar{Y}).

What are the “transition” probabilities $P(Y | X)$ and $P(Y | \bar{X})$? To answer this question, we introduce some very elementary statistical mechanics. The theory is based

on Newtonian mechanics, rather than quantum mechanics, only because the physics is easier.

Consider a free particle that finds itself enclosed within the box at time t . The particle is free in the sense that it is not subjected to any force except when it collides with the walls of its enclosure. Collisions with the walls are elastic, which means that it is reflected off the wall without losing energy; so, the component of the velocity in the perpendicular direction has a minus sign added to it, but does not change in magnitude. The constant kinetic energy implies that it has a constant speed, which we will take to be $\frac{1}{2}$ meter per second. The only thing relevant to predicting the outcome of the experiment is the x component of the particle's initial velocity (which can vary from $-\frac{1}{2}$ to 0, to $+\frac{1}{2}$) and the x component of its initial position. What probability distribution should we assign to the x component of its velocity? Assuming that there is no privileged direction in which it is moving; that is, the distribution of possible directions of motion is uniform, the initial x velocities are uniformly distributed between $-\frac{1}{2}$ to $\frac{1}{2}$.¹⁴ The probability distribution of initial x positions of the particle within its enclosure is also assumed to be uniform, and probabilistically independent of the velocity. That is all we need.

These assumptions are conveniently expressed in terms of the space of possible initial states of the particle, called the *phase space*. The left half of the rectangle in Fig. 6 represents the phase space of the particle if X is true, while the right part is the phase space if \bar{X} is true. (Those unfamiliar with state spaces can think of it as a Venn diagram of possible worlds.) The probability distribution of initial states is uniform over the

¹⁴ The result follows from what I call the orange peel theorem. Cut a spherical orange into slices of equal thicknesses. Then each slice has the same area of rind on it.

relevant area of phase space (this is an example of what is known as the microcanonical distribution in classical statistical mechanics).

Now assume that X is true. The theory says that the initial state of the particle is uniformly distributed over the

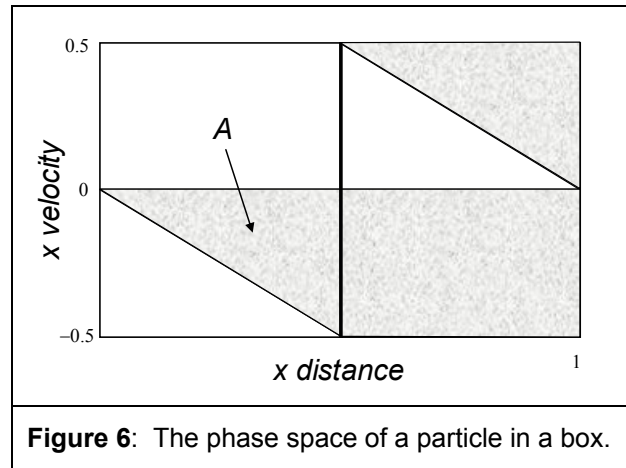


Figure 6: The phase space of a particle in a box.

accessible phase space (the region to the left of the vertical line in Fig. 6). The particle moves with constant speed for one second. Where it ends up depends on how fast it's moving and where it starts. If it's on the top line in phase space, it moves $\frac{1}{2}$ meter to the right. If it's on the center line, it doesn't move at all. If it's on the bottom line, it moves $\frac{1}{2}$ meter to the left. The shaded region in Fig. 6 represents all the possible initial states (irrespective of whether X or \bar{X} is true) that result in the event Y . The states in region A are the only ones in which both X and Y are true.¹⁵ Since region A represents $\frac{1}{4}$ of the states in which X is true, the probability of Y given X is $\frac{1}{4}$. By a similar argument, the probability of Y given \bar{X} is $\frac{3}{4}$. In symbols, $P(Y | X) = \frac{1}{4}$ and $P(Y | \bar{X}) = \frac{3}{4}$.

The conditional, or transition, probabilities are therefore invariant in the sense that they do not depend on the marginal probability $P(X)$. Of course, they are not invariant conditional on future events—for example, $P(Y | X, Y) = 1 \neq \frac{1}{4}$. Nor do they have to be equal to the particular number calculated above. The number $\frac{1}{4}$ was obtained by

¹⁵ To see this directly, note that if the particle has an initial state above the triangle A then it moves to the right, but one second is not long enough for it to bounce off the wall. So, it would exit to the right. If the particle has a state below the triangle, it will bounce off the left wall before the second is up, and also exit to the right. Therefore, out of those states that make X true, the particle exits to the left (that is, Y is true) if and only if it is in state A at time t .

assuming a uniform microcanonical distribution over the accessible state space, whereas the invariance would still follow if we replaced that distribution by another one, provided that it is the same for each particle. So, let's introduce adjustable parameters α and β to represent the values of the conditional probabilities, and allow the model itself to extract information about the distribution from the data. No constraint has been placed on the unconditional (marginal) probabilities of X and \bar{X} . They can be anything.

To formulate the model precisely, consider N trials of the experiment, and introduce the space of possible events in the i th trial as Boolean combinations of X_i and Y_i , which denote the occurrence of events of types X and Y in that trial. The model introduces N adjustable parameters for the probabilities of the X_i , $\pi_i = P(X_i)$, and two adjustable parameters for the transition probabilities: $\alpha = P(Y_i | X_i)$, and $\beta = P(Y_i | \bar{X}_i)$, for $1 \leq i \leq N$. We could make the model more analogous to F in Example 6 of the previous section by not making the π_i adjustable, but instead fixing their values to 1 if X_i occurs and 0 if \bar{X}_i occurs. Instead, I intend to vary the example a little by allowing π_i to take on either of the values 1 or 0. Clearly, if the parameter is adjusted to the “wrong” value, the total likelihood will be 0, so the extension is actually fairly trivial. α and β , on the other hand, range over all values from 0 to 1.

The model specifies the probability of each possible outcome of a trial of the experiment ($1 \leq i \leq N$) as follows:

$$P(X_i Y_i) = \pi_i \alpha, \quad P(X_i \bar{Y}_i) = \pi_i (1 - \alpha), \quad P(\bar{X}_i Y_i) = (1 - \pi_i) \beta, \quad P(\bar{X}_i \bar{Y}_i) = (1 - \pi_i) (1 - \beta).$$

Assuming that the outcomes of different trials are probabilistically independent, for example, $P(X_1\bar{Y}_1X_2Y_2) = P(X_1\bar{Y}_1)P(X_2Y_2)$, each hypothesis in the model has a well defined likelihood with respect to the observed data.

For example, imagine that the observed frequencies over 240 trials of the experiment are $n(XY) = 30$, $n(X\bar{Y}) = 90$, $n(\bar{X}Y) = 90$, and $n(\bar{X}\bar{Y}) = 30$. What is the likelihood function of the model? It consists of two factors—one for the marginal probabilities and one for the conditional probabilities. Consider the marginal probabilities first; let π'_i be π_i or $(1 - \pi_i)$ depending on whether X or \bar{X} occurs in trial i . Then the first factor in the likelihood function is $\pi'_1\pi'_2 \cdots \pi'_{240}$, while the second factor is $\alpha^{30}(1 - \alpha)^{90} \beta^{90}(1 - \beta)^{30}$. Denoting the model by F (for Forward), its likelihood function, L_F is:

Model F :
$$L_F = (\pi'_1\pi'_2 \cdots \pi'_{240})\alpha^{30}(1 - \alpha)^{90} \beta^{90}(1 - \beta)^{30}.$$

Now consider the rival theory and model, denoted by B (for Backward). The Backward theory views the particle is being ejected backwards in time by the detectors with speed randomly distributed between 0 and $\frac{1}{2}$ such that its space position is also uniformly distributed at time $t+1$ within the whole volume. That is, the distribution of states at time $t+1$ is uniformly distributed over the lower half of the phase space in Fig. 6. The particles move backwards for one second, so that at time t the distribution is given by the shaded region in Fig. 6. The question is: What proportion of the shaded region lies in the left partition? As before, the only states that make X and Y true lie in region A at time t , which is $\frac{1}{4}$ of the total shaded region. So, $P(X | Y) = \frac{1}{4}$. By a similar argument, $P(X | \bar{Y}) = \frac{3}{4}$. Since we don't want to rely on the microcanonical distribution, we

introduce parameters $a = P(X_i | Y_i)$ and $b = P(X_i | \bar{Y}_i)$, for $1 \leq i \leq N$. The marginal probabilities of Y and \bar{Y} can be anything, so we introduce $p_i = P(Y_i)$ for $1 \leq i \leq N$. let p'_i denote p_i or $(1 - p_i)$ depending on whether Y or \bar{Y} occurs in trial i . The model now specifies a joint distribution for each trial of the experiment ($1 \leq i \leq N$):

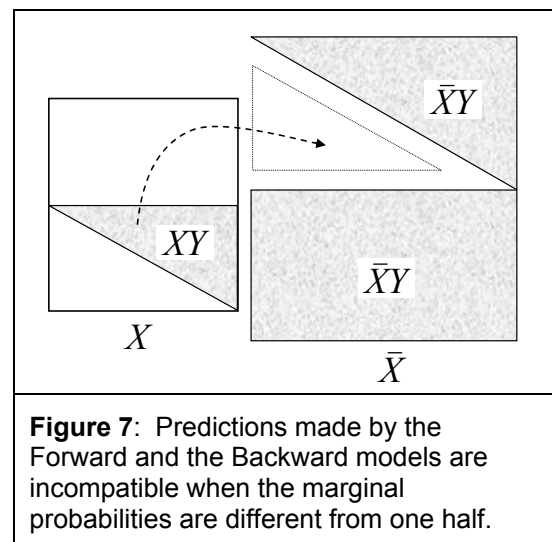
$$P(X_i Y_i) = p_i a, P(X_i \bar{Y}_i) = (1 - p_i) b, P(\bar{X}_i Y_i) = p_i (1 - a), P(\bar{X}_i \bar{Y}_i) = (1 - p_i) (1 - b).$$

Assuming that the outcomes in different trials are probabilistically independent, the likelihood function, denoted by L_B is:

Model B :
$$L_B = (p'_1 p'_2 \cdots p'_{240}) a^{30} b^{90} (1 - a)^{90} (1 - b)^{30}.$$

Under the correspondence $\pi'_i \leftrightarrow p'_i$, $\alpha \leftrightarrow a$, and $\beta \leftrightarrow b$, the likelihood functions are the same ($L_F = L_B$), so the Forward model and the Backward model are likelihood equivalent. In Bayesian terms, the likelihood equivalence implies that no matter what prior distribution is placed over the parameters, the likelihood of F will be matched by the likelihood of B (calculated in both cases as the average of the likelihoods of the hypotheses weighted by the prior). In which case, the posterior probability of F can only be higher than the posterior probability of B by adjusting the prior probability of B to be lower than F . But does it all reduce to merely subjective degrees of belief? That is the question before us.

The Forward model predicts that the forward transition probabilities would be exactly the same even if the marginal



probabilities for X and \bar{X} were different. Suppose that the Forward theorist is told that $P(X) = \frac{1}{3}$ and $P(\bar{X}) = \frac{2}{3}$. The relative sizes of the phase space volumes have been redrawn in Fig. 7 to reflect the change. There is no change in the relative proportions within the X region of phase space, or within the \bar{X} region, which is why the forward transition probabilities are the same. But the backward probabilities have changed. For example, $P(X|Y)$ is calculated by placing all the gray areas together (see Fig. 7), and asking for the area of XY region in proportion to all the gray areas. This is now less than $\frac{1}{4}$. In fact, it is exactly equal to $\frac{1}{7}$. The backward probabilities have changed exactly because the forward probabilities are invariant (Forster 1984, Sober 1994).

On the other hand, the Backward theorist is told that the marginal is $P(Y) = \frac{7}{12}$. But the Backward model predicts that the backward probabilities are unchanged. That is, $P(X|Y) = \frac{1}{4}$, which contradicts what the Forward model says. The models make incompatible predictions.

Why doesn't this refute the Likelihood Principle? It seems to establish that the models are *empirically* inequivalent. But empirical equivalence is different from evidential equivalence, which is what the Likelihood Principle is about. Evidential equivalence is empirical equivalence *relative to the actual data*. The test situation just described refers to *counterfactual* data, so there is no obvious violation of the Likelihood Principle.

But the test data just described need

	Laboratory 1		Laboratory 2	
	Y	\bar{Y}	Y	\bar{Y}
X	10	30	20	60
\bar{X}	60	20	30	10

Table 1: Data collected in 2 laboratories. Should the success of a model in predicting aspects of the second data set from the first count as a part of the evidential support for a model?

not be counterfactual. Suppose that the actual data are composed of two data sets, collected in different laboratories, such that in the first data set the relative frequency of X is $\frac{1}{3}$. The data are shown in Table 1. Given that the numbers are quite small, there will be sampling errors in the estimation of the probabilities. But that is easily fixed by adding whatever number of zeros you want to each entry, while the relative frequencies are the same. So, we may assume that the data provide very accurate estimations of the transition probabilities within each subset of data. The Forward model predicts that the forward probabilities will be the same for both halves of the data, while the Backwards model predicts that they will be different. It is easy to see from Table 1 that the Forward model is right and the Backward model is wrong. Yet, as we have already shown, the two models are likelihood equivalent. It seems to be a clear counterexample to the Likelihood Theory of Evidence. Now the tug-of-war begin.

Objection (1) Surely, there must be ways of subdividing the data in ways that would lead to the opposite conclusion. For example, randomly select 70 of the 120 data in which Y is true, and put them in data set 1. Then randomly select 50 of the 120 \bar{Y} data, and add them to data set 1. Put the remaining data in data set 2. Then the backward transition probabilities will not vary between data sets 1 and 2, and the forward transition probabilities will vary.

Reply: We can always mimic the ways that the Backward theory *assumes* that Nature assort the instances. But that is the point. The competing theories are about how *Nature* assort instances, and not about how *we* are able to sort data (or not). The test relies on a *natural* assortment of the data; each model makes a definite prediction, one is right, and one is wrong.

Objection (2): What counts as a natural collection of instances is not specified by the model itself. If specified at all, it is specified by the background theory, which then constrains the submodels. In particular, it is only the theory that specifies which parameters are invariant, and which are not. If the models stand in isolation from the theory, then the Likelihood Principle applies.

Reply: So, let's reconstruct the models so that they say explicitly how they apply to the subsets of data, and what constraints hold between the submodels. Label the data points so that the first 120, $1 \leq i \leq 120$, are the ones collected in Laboratory 1. Let F_1 be the Forward model that applies to the Laboratory 1 data, and let F_2 be the model that applies to the Laboratory 2 data. The model equations are the same as before except that we use different parameters in each submodel; α_1 and β_1 in F_1 , α_2 and β_2 in F_2 . Now define a composite model F^* to be the conjunction of F_1 and F_2 plus the constraints $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$. Similarly, define B^* to be the conjunction of B_1 and B_2 plus the constraints $a_1 = a_2$ and $b_1 = b_2$. (Sound familiar?) From here the argument is relatively straightforward. F^* is likelihood equivalent to F . This might not be obvious given that F^* and F have different numbers of parameters. But the constraints reduce the *effective* number of parameters in F^* to two. Under the mapping $\alpha \leftrightarrow \alpha_1$, and $\beta \leftrightarrow \beta_1$, with the constraints understood, the likelihood functions of F^* and F are the same. In the same way, B^* and B are likelihood equivalent. But we have already established that F and B are likelihood equivalent, therefore (by the transitivity of equivalence relations) F^* and B^* are likelihood equivalent. If we test the constraints, F^* passes and B^* fails, so F^* and B^* are not evidentially equivalent. Therefore the Likelihood Theory of Evidence is false.

7. Conclusion

The examples described in this paper are designed to show that measures of fit (with the total observed evidence) do not capture the full impact of evidence on hypotheses. Hypotheses do more. They also relate different parts of the data together, or fail to do so (see Example 3 and Example 4, Section 4). Likelihood is a measure of fit, and it is no exception. It is false that all the empirical information relevant to the comparison of hypotheses or models is contained in the likelihoods. In the case of models (families of simple hypotheses), the likelihood function determines how well the model is able to *accommodate* the data, but it leaves out important information about how well it can *predict* one part of the data from another. Often, the predictive achievements are conveniently summarized in terms of the agreement of independent measurements of the theoretical quantities posited by the models.

The empirical overdetermination of parameters, coefficients (Whewell 1958, Forster 1988), or constants (Norton 2000), played a pivotal role in Newton's argument for universal gravitation (Whewell 1958, Forster 1988, Harper 2002), and in Perrin's argument for the existence of atomic constituents of matter (see Norton 2000). That is why the Likelihood Theory of Evidence (see Section 1), and the Bayesian philosophies of science founded on it, will always fail to provide an adequate theory of scientific reasoning.

Statisticians have traditionally restricted their application of the Likelihood Principle to a narrower set of inferential problems—mainly, those involving the estimation of parameter values under the assumption that the model that defines them is true. But how does science establish the correctness of a model in the first place? That

question calls for a deeper understanding of scientific reasoning than any version of the Likelihood Principle can provide.

In recent years, statisticians have turned their attention to the problem of model comparison, or model selection, as it has come to be called. Unfortunately, most of the proposed model selection criteria are based on the comparison of single numbers derived from the likelihood function, and are therefore prone to the limitation described here.¹⁶ Criteria such as AIC (Akaike 1973) and BIC (Schwarz 1989) are examples because they are based on the maximum likelihood, the number of data, and the number of adjustable parameters. Bayes Factors compare average likelihoods derived directly from the likelihood function.¹⁷

Nevertheless, there is no reason why statistical methods cannot be used in evaluating the predictions of models, such as the predicted agreement of independent measurements; and this has, indeed, been a standard part of statistical practice. The problem is that theory lags behind practice. Future theories of statistical inference may connect with well discussed ideas in philosophy of science, such as William Whewell's notion of the colligation of facts and the consilience of inductions (Whewell 1958, 1989). Glymour taps into many of the same ideas in his early writings (*e.g.*, Glymour 1980) and Forster (1988) uses Whewellian ideas in replying to arguments against the existence of

¹⁶ Myrvold and Harper (2002) criticize the Akaike criterion of model selection (Forster and Sober 1994) because it underrates the importance of the agreement of independent measurements in Newton's argument for universal gravitation (see Harper 2002 for an intriguing discussion of Newton's argument). While this paper supports their conclusion, it does so in a more precise and general way. The important advance in this paper is (1) to point out that the limitation applies to all model selection criteria based on the Likelihood Principle and (2) to pinpoint exactly where the limitation lies. Nor is it my conclusion that statistics does not have the resources to address the problem.

¹⁷ Wasserman (2000) provides a nice survey.

forces.¹⁸ A more general theory of scientific inference may connect with an old argument for scientific realism described by Earman (1978), and independently by Friedman (1981), both of which are discussed in Forster (1986). But at the present time, none of these ideas is very well developed.

Appendix

Theorem: If the maximum likelihood hypothesis in F is $Y = \frac{10}{\sqrt{101}}X + U$ and the observed variance of X is 101, then the observed variance of Y is also 101. Thus, the maximum likelihood hypothesis in B is $X = \frac{10}{\sqrt{101}}Y + Z$, and they have the same likelihood. Moreover, for any α , β , and σ , there exist values of a , b , and s such that $Y = \alpha + \beta X + \sigma U$ and $X = a + bY + sZ$ have the same likelihood.

Partial Proof: The observed X variance of data distributed in two Gaussian clusters with unit variance centered at $X = -10$ and $X = +10$, where the observed means of X and Y are 0, is equal to

$$\text{Var}X = \frac{1}{2} \frac{1}{N/2} \sum x_i^2 + \frac{1}{2} \frac{1}{N/2} \sum x_j^2,$$

where x_i denotes X values in the lower cluster and x_j denotes X values in the upper cluster.

If all the x_i were equal to -10 , and all the x_j were equal to $+10$, then $\text{Var}X$ would be equal to 100. To that, one must add the effect of the local variances. More exactly,

$$\text{Var}X = \frac{1}{2} \frac{1}{N/2} \sum ((x_i + 10) - 10)^2 + \frac{1}{2} \frac{1}{N/2} \sum ((x_j - 10) + 10)^2 = 101.$$

From the equation $Y = \frac{10}{\sqrt{101}}X + U$, it follows that $\text{Var}Y = \frac{100}{101}101 + 1 = 101$. Standard

formulae for regression curves now prove that $X = \frac{10}{\sqrt{101}}Y$ is the backwards regression line,

¹⁸ Hooker (1987) and Norton (1993, 2000) discuss relevant issues and examples; in fact, there is a wealth of good literature in the philosophy of and history of science that deserves serious attention from outsiders.

where the observed residual variance is also equal to 1. Therefore, the two hypotheses have the same conditional likelihoods, and the same total likelihoods. It follows that the hypotheses $Y = \frac{10}{\sqrt{101}}X + \sigma U$ and $X = \frac{10}{\sqrt{101}}Y + \sigma Z$ have the same likelihoods for any value of σ . It is also clear that for any α , β , and σ , there exist values of a , b , and s such that $Y = \alpha + \beta X + \sigma U$ and $X = a + bY + sZ$ have the same likelihoods.

References

- Aitkin, M. (1991): "Posterior Bayes Factors," *Journal of the Royal Statistical Society B* **53**: 111-142.
- Akaike, H. (1973): "Information Theory and an Extension of the Maximum Likelihood Principle." B. N. Petrov and F. Csaki (eds.), *2nd International Symposium on Information Theory*: 267-81. Budapest: Akademiai Kiado.
- Barnard, G. A. (1947) "Review of Wald's 'Sequential analysis'", *Journal of the American Statistical Association*, **42**: 658-669.
- Berger, James O. (1985): *Statistical Decision Theory and Bayesian Analysis*. Second Edition, Springer-Verlag, New York.
- Berger, James O. and Wolpert, Robert L. (1988) *The Likelihood Principle*. 2nd edition. Hayward, California: Institute of Mathematical Statistics.
- Birnbaum, A. (1962): "On the Foundations of Statistical Inference (with discussion)", *Journal of the American Statistical Association* **57**: 269-326.
- Boik, Robert J. (2004): "Commentary", in Mark Taper and Subhash Lele (eds), *The Nature of Scientific Evidence*, Chicago and London: University of Chicago Press, 167-180.
- Burnham, Kenneth P and Anderson, David R. (2002): *Model Selection and Multi-Model Inference*. New York: Springer Verlag.
- Earman, John (1978). "Fairy Tales vs. an Ongoing Story: Ramsey's Neglected Argument for Scientific Realism." *Philosophical Studies* **33**: 195-202.
- Edwards, A. W. F. (1987): *Likelihood*. Expanded Edition. The John Hopkins University Press: Baltimore and London.
- Fitelson, Branden (1999): "The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity," *Philosophy of Science* **66**: S362-78.
- Forster, Malcolm R. (1984): *Probabilistic Causality and the Foundations of Modern Science*. Ph.D. Thesis, University of Western Ontario.

- Forster, Malcolm R. (1986): "Unification and Scientific Realism Revisited." In Arthur Fine and Peter Machamer (eds.), *PSA 1986*. E. Lansing, Michigan: Philosophy of Science Association. Volume 1: 394-405.
- 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.
- Forster, Malcolm R. (1988b): "Sober's Principle of Common Cause and the Problem of Incomplete Hypotheses." *Philosophy of Science* **55**: 538-59.
- Forster, Malcolm R. (2000): "Key Concepts in Model Selection: Performance and Generalizability," *Journal of Mathematical Psychology* **44**: 205-231.
- Forster, Malcolm R. (forthcoming): "The Miraculous Consilience of Quantum Mechanics," in E. Eells and J. Fetzer (eds.) *Probability in Science*. Open Court.
- 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.
- Forster, Malcolm R. and Elliott Sober (2004): 'Why Likelihood?,' in Mark Taper and Subhash Lele (eds), *The Nature of Scientific Evidence*, Chicago and London: University of Chicago Press, 153-165.
- Forster, Malcolm R. and Elliott Sober (2004): 'Reply to Boik and Kruse,' in Mark Taper and Subhash Lele (eds), *The Nature of Scientific Evidence*, Chicago and London: University of Chicago Press, 181-190.
- Friedman, Michael (1981). "Theoretical Explanation," in *Time, Reduction and Reality*. Edited by R. A. Healey. Cambridge: Cambridge University Press. Pages 1-16.
- Glymour, Clark (1980). "Explanations, Tests, Unity and Necessity." *Noûs* **14**: 31-50.
- Hacking, Ian (1965). *Logic of Statistical Inference*. Cambridge: Cambridge University Press.
- 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*. Chicago and La Salle, Illinois: Open Court. 71-112.
- Hooker, Cliff A. (1987): *A Realistic Theory of Science*. Albany: State University of New York Press.
- Jeffreys, Harold (1961): *Theory of probability*. Third Edition. Oxford, The Clarendon press.
- Mayo, Deborah G. (1996): *Error and the Growth of Experimental Knowledge*. Chicago and London, The University of Chicago Press.
- Mermin, David N. (1990) "Quantum Mysteries Revisited." *American Journal of Physics*, August 1990, pp.731-4.
- Myrvold, Wayne and William L. Harper (2002), "Model Selection, Simplicity, and Scientific Inference", *Philosophy of Science* **69**: S135-S149.

- Norton, John D. (1993): "The Determination of Theory by Evidence: The Case for Quantum Discontinuity, 1900–1915", *Synthese* **97**: 1-31.
- Norton, John D. (2000): "How We Know about Electrons", in Robert Nola and Howard Sankey (eds.) *After Popper, Kuhn and Feyerabend*, Kluwer Academic Press, 67-97.
- Pearl, Judea (2000): *Causality: Models, Reasoning, and Inference*. Cambridge University Press.
- Pearl, Judea (2000): *Causality: Models, Reasoning, and Inference*. Cambridge University Press.
- Royall, Richard M. (1991): "Ethics and Statistics in Randomized Clinical Trials (with discussion)," *Statistical Science* **6**: 52-88.
- Royall, Richard M. (1997): *Statistical Evidence: A likelihood paradigm*. Boca Raton: Chapman & Hall/CRC.
- Savage, L. J. (1976) "On rereading R. A. Fisher (with discussion)", *Annals of Statistics*, **4**:441-500.
- Sakamoto, Y., M. Ishiguro, and G. Kitagawa (1986): *Akaike Information Criterion Statistics*. Dordrecht: Kluwer Academic Publishers.
- Schwarz, Gideon (1978): "Estimating the Dimension of a Model." *Annals of Statistics* **6**: 461-5.
- Sneed, Joseph D. (1971): *The Logical Structure of Mathematical Physics*. Dordrecht: D. Reidel.
- Sober, Elliott (1993): "Epistemology for Empiricists." In H. Wettstein (ed.), *Midwest Studies in Philosophy*. Notre Dame: University of Notre Dame Press; pp. 39-61.
- Sober, Elliott (1994): "Temporally Oriented Laws," in Sober (1994) *From A Biological Point of View - Essays in evolutionary philosophy*, Cambridge University Press, pp. 233 - 251.
- Wasserman, Larry (2000): "Bayesian model selection and model averaging." *Journal of Mathematical Psychology* **44**: 92-107.
- Whewell, William (1858): *Novum Organon Renovatum*, Part II of the 3rd the third edition of *The Philosophy of the Inductive Sciences*, London, Cass, 1967.
- Whewell, William (1989). In Butts, Robert E. (ed.) *Theory of Scientific Method*. Hackett Publishing Company, Indianapolis/Cambridge.
- Woodward, James. (2003): *Making Things Happen: A Theory of Causal Explanation*. Oxford and New York: Oxford University Press.