Counterexamples to a Likelihood Theory of Evidence

Final Draft, July 12, 2006. Forthcoming in *Minds and Machines*.

MALCOLM R. FORSTER[1]
Department of Philosophy, University of Wisconsin-Madison, 5185 Helen C. White Hall, 600 North Park Street, Madison WI 53706 USA. Email: mforster@wisc.edu

Abstract: The Likelihood Theory of Evidence (LTE) says, roughly, that all the information relevant to the bearing of data on hypotheses (or models) is contained in the likelihoods. 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 suggest that some forms of scientific reasoning, such as the consilience of inductions (Whewell, 1858), cannot be represented within Bayesian and Likelihoodist philosophies of science.

Key words: The likelihood principle, the law of likelihood, evidence, Bayesianism, Likelihoodism, curve fitting, regression, asymmetry of cause and effect.

1. Introduction

Consider two simple hypotheses, h_1 and h_2, with likelihoods denoted by $P(E \mid h_1)$ and $P(E \mid h_2)$ 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). If a hypothesis contains a free, or adjustable, parameter, such that different values would change its likelihood, then it is not a simple hypothesis. The term "simple" is standardly used in the classical statistics literature in this way—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 by the standard Bayesian method of comparing hypotheses, according to which two simple hypotheses are compared by their posterior probabilities, $P(h_1 \mid E)$ and $P(h_2 \mid E)$. Bayes theorem tells us that

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

Thus, 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 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-Bayesians statisticians, who want to 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 standardly referred to as the *likelihood function* of the model. Whenever the word ‘likelihood’ occurs in the sequel, it refers to the probability of the total observed data given the hypothesis under consideration.[3]

It is now possible 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 via the likelihoods of the simple hypotheses being compared (or the likelihood functions of the models under comparison). In other words, all the information about the total data that bears on the comparison of a hypothesis with others under consideration, reduces to a single number, namely its likelihood.

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 as well.[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* likelihoods are used.[6]

Even though LTE is vague about how likelihoods are used, it is very precise about what shouldn't be used…namely, everything else! You can't use likelihood defined relative to only part of the data, and you can't consider likelihoods of component parts of the hypothesis with respect to any part of the data. You can't include judgments about the simplicity and unification of hypotheses, or models, or their prior plausibility.

Other principles that fall under LTE, such as the Law of Likelihood (LL), are more specific about how likelihoods are used. LL says, roughly, that that evidence *E* supports h_1 or than it supports h_2, or that *E* favors h_1 over h_2 if and only if the likelihood of h_1 greater than the likelihood of h_2 (i.e., $P(E\,|\,h_1) > P(E\,|\,h_2)$). The terms 'support' and 'favors' are not defined. Challenges to LL, or to LTE, depend on some kind of intuitive grasp of their meanings, at least within the context of particular examples. It is because of this looseness in LL and LTE, that they gradually acquire the status of definitions in the minds of their adherents. Challenging entrenched ways of thinking, in any field, is never easy.

Contemporary statistics is divided into three camps; classical Neyman-Pearson statistics (see Mayo 1996 for a recent defense), Bayesianism (*e.g.*, Jefferys 1961, Savage 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 a 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 after that present counterexamples to LTE, first in terms of an example with binary (yes-no) variables, and then in terms of continuous variables (a simple curve fitting problem).

2. Fit Functions

Consider a simple 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 observed value. 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 its value 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.

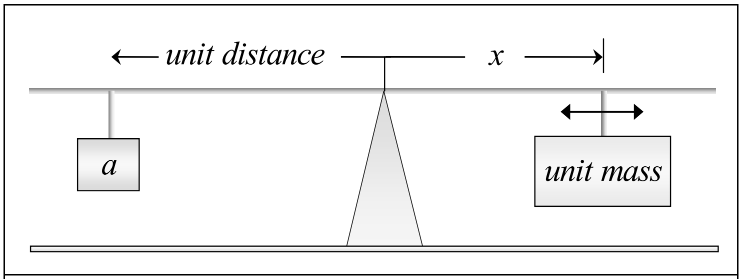

Figure 1: A simple beam balance.

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? Not yet, because here are two other options. We could weaken the hypothesis by adding an error term, or we could lower our sights from truth to predictive accuracy. In the next section, we consider the first option; here we consider the second option. In many ways, the two options are entirely related.

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 (in this un-explicated sense) 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:

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

where the data is $(x_1, x_2, \ldots, x_N)$. 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 different 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]

3. The Likelihood Principle

Likelihood is usefully understood as providing some kind of fit function.

> **Definition**: The *likelihood* of a hypothesis (relative to observed 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).

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 events, but disagreeing about the probability of the outcome H. Then each of the models will assign different probabilities to the total data, but there respective likelihood functions will differ only by a constant. Both models should therefore 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 functions with respect to x.

Berger and Wolpert (1988, pp. 19-21) add the following caveat to their version of the Likelihood Principle: That "…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 ½, 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. The simple hypotheses, $h_1 \,\&\, (\theta = 17)$ and $h_2 \,\&\, (\theta = 21)$, for example, 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).

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

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 −½ , or ½, each with probability ½. Strictly speaking, it's a contradiction to say that $Y = 3.5$ and then $Y = 4.5$. 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 attributed to it by the hypothesis under consideration. Since *U* is a random variable, so is $Y - X$ (since $U = Y - X$). But does it follow that *X* and *Y* are random variables? There are three cases to consider.

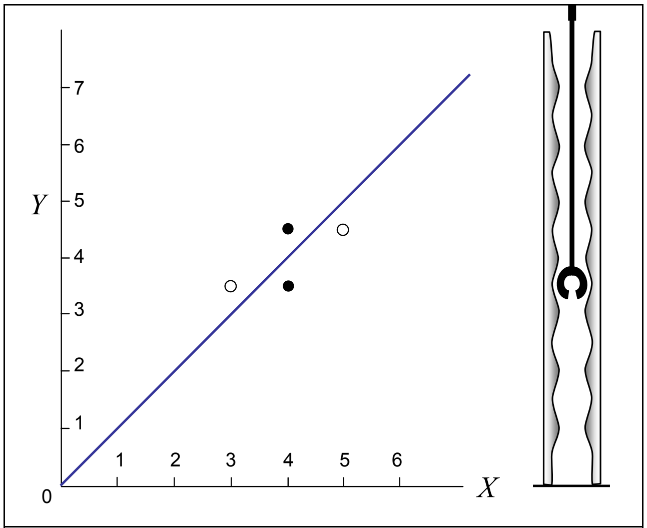

Figure 2: Discrete data, Section 4. The solid dots are points at which data is observed, while the open circles are points at which no data is observed.

Case (1): *X* is variable quantity 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 $Y(X)$. So, a conditional probability like $P(Y = y \mid X = x)$ would 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 value. It is better to write $P_{X=x}(Y = y)$.

Case (2): We could write the equation as $X = Y - U$ 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 correct interpretation of hypothesis A, and it is worth saying why. 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 it implies that if a variable is not random, then it is the exogenous variable (X in this example). This convention implies that case (2) does not apply to hypothesis A—this how the hypothesis represents an asymmetric between X and Y, as is appropriate in causal modeling.

Case (3): X is a random variable. that is, the probabilities $P(X = x)$ are specified 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, 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 probabilities $P_{X=x}(Y = y)$ in Case (1), and define a joint distribution:

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

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 derived by first interpreting the hypothesis as in Case (1), and then *assuming* that

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

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 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

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

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. In this case, B assigns no probabilities to the Y variables. It is easy to see that the conditional likelihood of B is also equal to ¼, and

$L(A) = L(B)$. So, there is no way of distinguishing between *A* and *B* the hypotheses in terms of conditional likelihoods.

While *A* and *B* cannot be distinguished by their conditional likelihoods, the two hypotheses *can* be distinguished on the basis of the data. First note that no matter how many times we duplicate the observed data, the conditional likelihoods will still be equal. 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 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, the two independent measurements of $P(U=\frac{1}{2})$ not only disagree with the hypothesized value (½), but also disagree with each other.

The example is already a counterexample to LTE 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.

A subscriber to LTE can deny that LTE applies to hypotheses that are incomplete in this sense.[9] They might insist that the example violates the principle of total evidence because the likelihoods are not relative to the full data, even though there 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 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 show up in the likelihoods? No! Because the likelihoods of the two hypotheses still equal.

A more direct way of presenting the counterexample would be to construct hypotheses from *A* and *B* by adding an auxiliary hypothesis about the values of the exogenous variable in each case. Let **X** = **x** be statement about the sequence of observed values of the variables, $X_1, X_2, \ldots, X_{20}$, and define hypothesis *A** as the conjunction of *A* and **X** = **x**. Similarly *B** is the conjunction of *B* and **Y** = **y**. Then it is easy to the that *A** and *B** have well defined likelihoods relative to the full data, and that the likelihoods are equal, that is, $L(A^*) = L(B^*)$. These are not conditional likelihoods. LTE implies that if you can tell which one is true from the data, then you can tell from the likelihoods. But, in this example, you can tell which is true from the data, by the same method as before. *A* is true, and **X** = **x** is true, therefore *A** is true. Similarly, *B* is true, and **Y** = **y** is true, therefore *B** is false. Therefore, LTE is false.

One might object that the hypotheses have been constructed with full knowledge of the data. This won't save LTE because it denies the relevance of non-empirical, historical, or psychological considerations. And rightly so, in this example, for we can tell which hypothesis is false from the data alone!

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 added distributions*. 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 ¼. 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 $(\frac{1}{4})^{20}$.

Let's back up a little. Bayesian philosophers of science say that *E confirms* hypothesis *H* if and only if $P(H \mid E) > P(H)$. They also say that E' would confirm *H* better than *E* confirms *H* if and only if $P(H \mid E') - P(H) > P(H \mid E) - P(H)$.[11] So, if the Bayesian theory is to match our intuitions in this example, then $P(B'' \mid E'') > P(B'' \mid E)$. But, since $P(E'' \mid B'') = P(E \mid B'')$, 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 many 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 any particular sequence of 50 heads and 50 tails. In isolation, this is not a counterexample to LTE, because it concerns a single hypotheses and two data sets. But there is a connection. In the coin tossing example, a string of all heads is strong evidence that that the coin is not fair because there is a relevant statistic, the observed relative frequency of heads, which has a probability distribution sharply peaked around the value ½ (according to the null hypothesis that the coin is fair). But the observed value is very far from that value, so a classical hypothesis test will reject the hypothesis. If the alternative hypotheses are ones that make the same background assumptions, but change the parameter value (the coin bias), then this classical test does not conflict with LTE because the statistic is *sufficient*. Then, by definition, the problem involves the comparison of hypotheses in a restricted set (a model) such that the likelihood function (relative to full data) is equal to a constant times the likelihood function relative to the sufficient statistic. That is what it means for a statistic to be sufficient. The problem is that in general scientific reasoning, we are interested in comparing hypotheses in different models. There the standard notion of statistical sufficiency breaks down, and the relevant statistics may not be sufficient in the sense required by the LTE.

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 are also 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

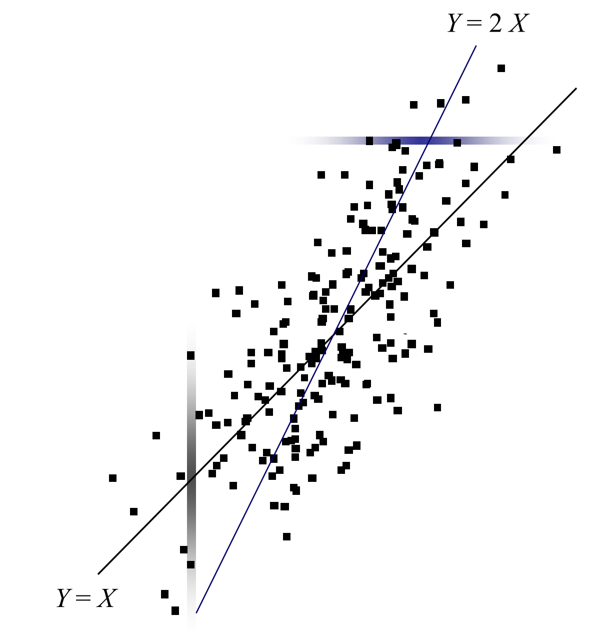

Figure 3: There are two ways of generating the same data: The Forward method and the Backward method (see text). It is impossible to tell which method was used from the data alone.

(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: Consider two hypotheses about how the data are generated. 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 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 from the data alone method which was used.

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 specifies a probability distribution for an x value, which we write as $p_F(x)$, and then specifies the probability density y given x; in symbols, $p_F(y \mid x)$. This determines the joint distribution $p_F(x, y) = p_F(x)p_F(y \mid x)$. Similarly, the Backward hypothesis determines a joint distribution for x and y of the form $p_B(x, y) = p_B(y)p_B(x \mid y)$. 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. In this case, they also cannot be distinguished by the full data, which is why Example 1 is not a counterexample to LTE.

Example 2: The hypotheses considered in Example 1 are the standard in textbooks, but they are not the only ones possible. Consider the comparison of two simple hypotheses that are not likelihood equivalent with respect to the data in Fig. 3. This will not be a counterexample to LTE either, but it does raise some important worries, which will be exploited in subsequent examples. Let us specify the two hypotheses more concretely by assuming that $\mu = 0$, so that the data are centered around the point (0,0). We are told that one of two hypotheses is true:

F_2: $Y = X + U$, and X is standard Gaussian.

B_2: $X = Y + V$, and Y is Gaussian with mean 0 and variance 2,

where again 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 values 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 ½). 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 independent or 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_2 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 $p_F(y \mid x) = p_B(x \mid y)$. 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, 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,

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

Since B_2 has a lower likelihood than F_2, it must be because $p_B(y)$ is smaller than $p_F(x)$, on average.

This is odd because the specification of marginal distributions, $p_B(y)$ and $p_F(x)$ is not what we think of as the essential content of a 'causal' hypothesis. The falsity of B_2 is already apparent from the *pattern* that forms when *x* values generated from *y* values, *even we look at a narrow range of y values*. 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* = 2*X* with half the variance, just as we 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* = 2*X* from the line *Y* = *X* (the 'explainable' variation) and the other due to the smaller random variation about the line *Y* = 2*X*. 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*.

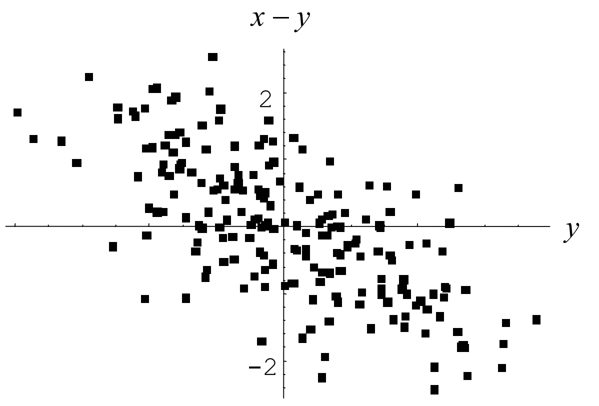

Figure 4: The *x* residuals (*x* – *y*) plotted as a function of *y*. The residual has an average variation of 1, but the variation varies in a systematic way for different values of *y*.

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. That is, *X* has a uniform distribution between the listed value minus $\delta/2$ and the listed value plus $\delta/2$. If we define $\varepsilon = 1/\delta$, then for all observed *x*, $P_F(x) = \varepsilon dx$. 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 other variable is generated is the same as before, in each case. Now, we are told that one of these hypotheses 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. So, this is a counterexample to LTE.

In this example, 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 new story about how the exogenous values are generated, then the Backward and Forward hypotheses would be genuinely indistinguishable on the basis of the full data—either one could have generated the data, and we couldn't know from the data which is true. 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*.

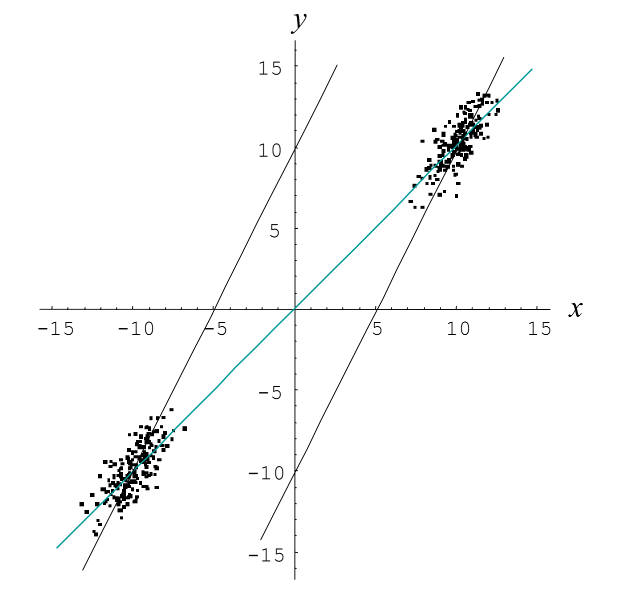

Figure 5: The asymmetry of regression. If a regression analysis is performed on the two clusters of data separately, then the Forward regression lines will coincide. But the Backward regression lines are very distinct.

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

There are exceptions to the rule that all the empirical information relevant to the comparison of hypotheses or models is contained in the likelihoods. Likelihood measures how well a hypothesis 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. Very often, these 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, or 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 and the Bayesian philosophies of science founded on it, will always fail to provide a complete theory of scientific reasoning.

Statisticians have traditionally restricted their application of the Likelihood Theory of Evidence 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 theory can provide.

In recent years, statisticians have turned their attention to the problem of model comparison, or model selection. 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.[14] Criteria such as AIC (Akaike 1973) and BIC (Schwarz 1989) are examples because they are based on the maximum likelihood, which is a feature of the likelihood function. Bayes Factors compare average likelihoods derived directly from the likelihood function.[15]

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 always been a standard part of statistical practice. The problem is that theory lags practice. Future theories of statistical inference should pay more attention to well discussed ideas in philosophy of science, such as William Whewell's concept of scientific induction (which he calls 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 forces.[16] A more general theory of scientific reasoning may also 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). At the present time, these ideas have not been fully explored.

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

$$VarX = \tfrac{1}{2}\tfrac{1}{N/2}\sum x_i^2 + \tfrac{1}{2}\tfrac{1}{N/2}\sum x_j^2 \,,$$

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 where equal to −10, and all the x_j were equal to +10, then $VarX$ would be equal to 100. To that, one must add the effect of the local variances. More exactly,

$$VarX = \tfrac{1}{2}\tfrac{1}{N/2}\sum((x_i+10)-10)^2 + \tfrac{1}{2}\tfrac{1}{N/2}\sum((x_j-10)+10)^2 = 101 \,.$$

From the equation $Y = \frac{10}{\sqrt{101}} X + U$, it follows that $VarY = \frac{100}{101}101 + 1 = 101$. Standard formulae for regression curves now prove that $X = \frac{10}{\sqrt{101}} Y$ is the backwards regression line, 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.

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*, 42:441-500.

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.

Notes:

[1] 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 on different parts of the manuscript.

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 has taken me 20 years to figure out.

[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.

[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.

[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.

[8] I owe this suggestion to Jason Grossman.

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

[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.

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

[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.

[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.

[14] 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.

[15] Wasserman (2000) provides a nice survey.

[16] 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.