Author & Citation Info | Friends PDF Preview | InPho Search | PhilPapers Bibliography

Inductive Logic

First published Mon Sep 6, 2004; substantive revision Wed Feb 27, 2008; substantive revision Thurs Feb 10, 2011

An inductive logic is a system of evidential support that extends deductive logic to less-than-certain inferences. For valid deductive arguments the premises *logically entail* the conclusion, where the entailment means that the truth of the premises provides a *guarantee* of the truth of the conclusion. Similarly, in a good inductive argument the premises should provide some *degree of support* for the conclusion, where such support means that the truth of the premises indicates with some *degree of strength* that the conclusion is true. Presumably, if the logic of good inductive arguments is to be of any real value, the measure of support it articulates should meet the following condition:

> **Criterion of Adequacy (CoA)**:
> As evidence accumulates, the *degree* to which the collection of true evidence statements comes to *support* a hypothesis, as measured by the logic, should tend to indicate that false hypotheses are probably false and that true hypotheses are probably true.

This article will focus on the kind of the approach to inductive logic most widely studied by philosophers and logicians in recent years. These logics employ conditional probability functions to represent measures of the degree to which evidence statements support hypotheses. This kind of approach usually draws on Bayes' theorem, which is a theorem of probability theory, to articulate how the *implications of hypotheses about evidence claims* influences the degree to which hypotheses are supported by those evidence claims. We will examine the extent to which this kind of logic may pass muster as an adequate logic of evidential support, especially in regard to the testing of scientific hypotheses. In particular, we will see how such a logic may be shown to satisfy the Criterion of Adequacy.

Sections 1 through 3 present all of the main ideas behind the probabilistic logic of evidential support. For most readers these three sections will suffice to provide an adequate understanding of the subject. Those readers who want to know more about how the logic applies when the *implications of hypotheses about evidence claims* (called *likelihoods*) are vague or imprecise may, after reading sections 1-3, skip down to section 6.

Sections 4 and 5 are for the more advanced reader who wants a detailed understanding of some telling results about how this logic may bring about convergence to the truth. These results show that the Criterion of Adequacy is indeed satisfied — that as evidence accumulates, false hypotheses will very probably come to have evidential support values (as measured by their *posterior probabilities*) that approach 0; and as this happens, a true hypothesis will very probably acquire evidential support values (as measured by their *posterior probabilities*) that approach 1.

---

1. Inductive Arguments

Let us begin by considering examples of the kinds of arguments an inductive logic should explicate. Consider the following two arguments:

> **Example 1.**. Every raven in a random sample of 3200 ravens is black. This strongly supports the hypothesis that all ravens are black.

> **Example 2**. 62 percent of voters in a random sample of 400 registered voters (polled on February 20, 2004) said that they favor John Kerry over George W. Bush for President in the 2004 Presidential election. This supports with a probability of at least .95 the hypothesis that between 57 percent and 67 percent of all registered voters favor Kerry over Bush for President (at or around the time the poll was taken).

An argument of this kind is often called an *induction by enumeration* of cases. We may represent the logical form of such arguments semi-formally as follows:

> Premise: In random sample S consisting of n members of population B, the proportion of members that have attribute A is r.
>
> Therefore, with degree of support p,
>
> Conclusion: The proportion of all members of B that have attribute A is between $r-q$ and $r+q$ (i.e., is within *margin of error* q of r).

Let's lay out this argument more formally. The Premise breaks down into three separate premises:[1]

	Semi-formalization	Formalization
Premise 1	The frequency (or proportion) of members with attribute A among the members of S is r.	$F[A,S] = r$
Premise 2	S is a random sample of B with respect to whether or not its members have A	$\mathrm{Rnd}[S,B,A]$
Premise 3	Sample S has exactly n members	$\mathrm{Size}[S] = n$
Therefore	with degree of support p	========{p
Conclusion	The proportion of all members of B that have attribute A is between $r-q$ and $r+q$ (i.e. is within *margin of error* q of r)	$F[A,B] = r \pm q$

Any inductive logic that encompasses such arguments should address two challenges. (1) It should tell us which enumerative inductive arguments should count as *good* inductive arguments rather than as inductive fallacies. In particular, it should tell us how to determine the appropriate *degree p* to which such premises *inductively support* the conclusion, for a given margin of error *q*. (2) It should demonstrably satisfy the CoA. That is, it should be provable (as a metatheorem) that *if* a conclusion expressing the approximate proportion for an attribute in a population is true, *then* it is very likely that sufficiently numerous random samples of the population will provide true premises for *good* inductive arguments that confer *degrees of support p* approaching 1 for that true conclusion — where, on pain of triviality, these *sufficiently numerous* samples are only a tiny fraction of a large population. Later we will see how a probabilistic inductive logic may meet these two challenges.

Enumerative induction is rather limited in scope. This form of induction is only applicable to the support of claims involving simple universal conditionals (i.e., claims of form 'All *B*s are *A*s') and claims about the proportion of an attribute in a population (i.e., 'The frequency of *A*s among the *B*s is *r*'). And it applies only when the evidence for such claims consists of instances of *B*s observed to be either *A*s or non-*A*s. However, many important empirical hypotheses are not reducible to this simple form, and the evidence for hypotheses is often not composed of simple instances. Consider, for example, the Newtonian Theory of Mechanics:

> All objects remain at rest or in uniform motion unless acted upon by some external force. An object's acceleration (i.e., the rate at which its motion changes from rest or uniform motion) is in the same direction as the force exerted on it; and the rate at which the object accelerates due to a force is equal to the magnitude of the force divided by the object's mass. If an object exerts a force on another object, the second object exerts an equal amount of force on the first object, but in the opposite direction to the force exerted by the first object.

The evidence for (and against) this theory is not gotten by examining a randomly selected subset of objects and the forces acting upon them. Rather, the theory is tested by calculating observable phenomena entailed by it in a wide variety of specific situations — ranging from simple collisions between small bodies to the trajectories of planets and comets — and then seeing whether those phenomena really occur. This approach to testing hypotheses and theories is ubiquitous, and should be captured by an adequate inductive logic.

Many less theoretical instances of inductive reasoning also fail to be captured by enumerative induction. Consider the kinds of inferences members of a jury are supposed to make based on the evidence presented at a murder trial. The inference to probable guilt or innocence is usually based on a patchwork of various sorts of evidence. It almost never involves consideration of a randomly selected sequences of past situations when people like the accused committed similar murders. Or, consider how a doctor diagnoses her patient on the basis of his symptoms. Although the frequency of occurrence of various diseases when similar symptoms were present may play a role, this is clearly not the whole story. Diagnosticians commonly employ a form of *hypothetical reasoning* — e.g., if the patient has a brain tumor, would that account for all of his symptoms?; or are these symptoms more likely the result of a minor stroke?; or is there another possible cause? The point is that a full account of inductive logic should not be limited to enumerative induction, but should also explicate the logic of *hypothetical reasoning* through which hypotheses and theories are tested on the basis of their predictions about specific observations. In Section 3 we will see how a kind of probabilistic inductive logic called "Bayesian Confirmation Theory" captures such reasoning.

2. Inductive Logic and Inductive Probabilities

Probability, and the equivalent notion *odds*, are the oldest and best understood ways of representing partial belief and uncertain inference. Probability has been studied by mathematicians for over 350 years, but the concept is certainly much older. In recent times a number of other related representations of uncertainty have emerged. Many of these have found useful application in computer based artificial intelligence systems that perform inductive inferences in expert domains such as medical diagnosis. This article will explicate the representation of inductive inferences in terms of *probability*. A brief comparative description of some of the most prominent alternative representations may be found in the following supplementary document:

Some Prominent Approaches to the Represention of Uncertain Inferences.

2.1 The Historical Origins of Probabilistic Logic

The mathematical study of probability originated with Blaise Pascal and Pierre de Fermat in the mid-17th century. From that time through the early 19th century, as the mathematical theory continued to develop, the theory was primarily applied to the assessment of risk in games of chance and to drawing simple statistical inferences about characteristics of large populations — e.g., to compute appropriate life insurance premiums based on mortality rates. In the early 19th century Pierre de Laplace made further theoretical advances and showed how to apply probabilistic reasoning to a much wider range of scientific and practical problems. Since that time probability has become an indispensable tool in the sciences, business, and many other areas of modern life.

Throughout its development various researchers appear to have thought of probability as a kind of logic. But the first extended treatment of probability as an explicit part of logic was George Boole's *The Laws of Thought* (1854). John Venn followed two decades later with an alternative empirical frequentist account of probability in *The Logic of Chance* (1876). Not long after that the whole discipline of logic was transformed by new developments in deductive logic.

In the late 19th and early 20th century Frege, followed by Russell and Whitehead, showed how deductive logic could be represented in the kind of rigorous formal system we now call *quantificational logic* or *predicate logic*. For the first time logicians had a fully formal deductive logic powerful enough to represent all valid deductive arguments in mathematics and the sciences — a logic in which the validity of deductive arguments depends only on the logical structure of the sentences involved. This development spurred some logicians to attempt to apply a similar approach to inductive reasoning. The idea was to extend the deductive entailment relation to a notion of *probabilistic entailment* for cases where premises provide less than conclusive support for conclusions. These *partial entailments* are expressed in terms of *conditional probabilities*, probabilities of the form $P[C \mid B] = r$ (read "the probability of C given B is r"), where P is a probability function, C is a conclusion sentence, B is a conjunction of premise sentences, and r is the probabilistic degree of support that B provides for C. Attempts to develop such a logic have varied widely in regard to precisely how the deductive model is emulated.

Some inductive logicians have tried to follow the deductive paradigm very closely by attempting to specify inductive support probabilities in terms of the syntactic structures of premise and conclusion sentences. In deductive logic the syntactic structure of the sentences involved completely determines whether premises logically entail a conclusion. So these logicians attempted to specify inductive support probabilities solely in terms of the syntactic structure of premise and conclusion sentences. In such a system each sentence confers a syntactically specified degree of support on each of the other sentences of the language. The inductive probabilities in such a system are *logical* in the sense that they depend on syntactic structure alone. This kind of conception was articulated to some extent by John Maynard Keynes in his *Treatise on Probability* (1921). Rudolf Carnap pursued this idea with greater rigor in his *Logical Foundations of Probability* (1950) and in several subsequent works (e.g., Carnap 1952). (For details of Carnap's approach see the section on logical probability in the entry on interpretations of the probability calculus, in this *Encyclopedia*.)

In the inductive logics of Keynes and Carnap, Bayes' theorem, which is a theorem of probability theory, plays a central role in expressing how evidence comes to bear on hypotheses. (We'll examine Bayes' theorem later.) So, such approaches might well be called *Bayesian logicist* inductive logics. Other well-known Bayesian logicist attempts to develop a probabilistic inductive logic include (Jeffreys, 1939), (Jaynes, 1968), and (Rosenkrantz, 1981).

It is now generally held that the core idea of Bayesian logicism is fatally flawed — that syntactic logical structure cannot be the sole determiner of the degree to which premises inductively support conclusions. A crucial facet of the problem faced by Bayesian logicism involves how the logic is supposed to apply to scientific contexts where the conclusion sentence is some hypothesis or theory, and the premises are evidence claims. The difficulty is that in *any* probabilistic logic that satisfies the usual axioms for probabilities, the inductive support for a hypothesis must depend in part on its *prior probability*. This *prior probability* represents how plausible the hypothesis is supposed to be based on considerations other than the observational and experimental evidence (e.g. perhaps due to relevant plausibility arguments). A Bayesian logicist must tell us how to assign values to these pre-evidential *prior probabilities* of hypotheses, for each of the hypotheses or theories under consideration. Furthermore, this kind of Bayesian logicist must determine these *prior probability* values in a way that relies only on the syntactic logical structure of these hypotheses, perhaps based on some measure of their syntactic simplicities. There are severe technical problems with getting this idea to work. Moreover, various kinds of examples seem to show that such an approach must assign intuitively quite unreasonable prior probabilities to hypotheses in specific cases (see the footnote cited near the end of section 3.2 for details). Furthermore, for this idea to apply to the evidential support of real scientific theories, scientists would have to formalize theories in a way that makes their relevant syntactic structures apparent, and then evaluate theories solely on that syntactic basis (together with their syntactic relationships to evidence statements). Are we to evaluate alternative theories of gravitation (and alternative quantum theories) this way? This seems an extremely doubtful approach to the evaluation of real scientific theories and hypotheses. Thus, it seems that logical structure alone cannot suffice for the inductive evaluation of scientific hypotheses. (This issue will be treated in more detail in Section 3, after we first see how probabilistic logics employ Bayes' theorem to represent the evidential support for hypotheses as a function of *prior probabilities* together with their *evidential likelihoods*.)

At about the time the Bayesian logicist idea was developing, an alternative conception of probabilistic inductive reasoning was also emerging. This approach is now generally referred to as the Bayesian *subjectivist* or *personalist* approach to inductive reasoning (see, e.g., Ramsey, 1926; De Finetti, 1937; Savage 1954; Edwards, Lindman, Savage, 1963; Jeffrey, 1983, 1992; Howson, Urbach, 1993; Joyce 1999). It treats inductive probability as part of a larger normative theory of belief and action known as *Bayesian decision theory*. The principle idea is that the strength of an agent's desires for various possible outcomes should combine with her belief-strengths regarding claims about the world to produce optimally rational decisions. Bayesian subjectivists provide a logic that captures this idea, and they attempt to justify this logic by showing that in principle it leads to optimal decisions about which of various risky alternatives should be pursued. On the Bayesian subjectivist or personalist account of inductive probability, inductive probability functions represent the subjective (or personal) belief -strengths of ideally rational agents, the kind of belief strengths that figure into rational decision making. (See the section on subjective probability in the entry on interpretations of the probability calculus, in this *Encyclopedia*.)

Elements of the logicist conception of inductive logic live on today as part of the general approach called *Bayesian inductive logic*. However, among philosophers and statisticians the term ‘Bayesian’ is now most closely associated with the subjectivist or personalist account of belief and decision. And the term ‘Bayesian inductive logic’ has come to carry the connotation of a logic that involves purely subjective probabilities.

This current usage is misleading since for inductive logics the Bayesian/non-Bayesian distinction should really hang on whether the logic gives Bayes' theorem a prominent role, or whether the logic largely eschews the use of Bayes' theorem in inductive inferences (as do the *classical approaches* to statistical inference developed by R. A. Fisher (1922) and by Neyman and Pearson (1967)). Indeed, any inductive logic that employs the same probability functions to represent both the *probabilities of evidence claims due to hypotheses* and the *probabilities of hypotheses due to those evidence claims* must be a *Bayesian* inductive logic in this broader sense; because Bayes' theorem follows directly from the axioms that each probability function must satisfy, and Bayes' theorem expresses a necessary connection between the *probabilities of evidence claims due to hypotheses* and the *probabilities of hypotheses due to those evidence claims*.

In this article the *probabilistic inductive logic* we will examine is a *Bayesian* inductive logic in the broader sense. This logic will not presuppose the *subjectivist Bayesian theory* of belief and decision, and will avoid the objectionable features of Bayesian logicism. Later we will see that there are good reasons to distinguish *inductive probabilities* from Bayesian *degree-of-belief probabilities* and from *purely logical probabilities*. So, the probabilistic logic articulated in this article will be presented in a way that depends on neither of these conceptions of what the probability functions are. However, this version of the logic will be general enough that it may be fitted to a Bayesian subjectivist or Bayesian logicist program, if one desires to do that.

2.2 Probabilistic Logic: Axioms and Characteristics

All logics derive from the meanings of terms in sentences. What we now recognize as *formal deductive logic* rests on the meanings (i.e., the truth-functional properties) of the standard logical terms. These terms, and the symbols we will employ to represent them, are as follows: ‘not’, ‘~’; ‘and’, ‘·’; ‘or’, ‘∨’; truth-functional ‘if-then’, ‘⊃’; ‘if and only if’, ‘≡’; the quantifiers ‘all’, ‘∀’, and ‘some’, ‘∃’; and the identity relation, ‘=’. The meanings of all other terms (i.e., names, and predicate and relational expressions) are permitted to “float free”. That is, the logic depends neither on their meanings nor on the truth-values of sentences containing them. It merely supposes that these other terms are meaningful, and that sentences containing them have truth-values. Deductive logic then tells us that the logical structures of some sentences — i.e., the syntactic arrangements of their logical terms — preclude them from being jointly true of any possible state of affairs. That is the notion of *logical inconsistency*. The notion of *logical entailment* is interdefinable with it. A collection of premise sentences *logically entails* a conclusion sentence just when the negation of the conclusion is *logically inconsistent* with those premises.

An inductive logic must, it seems, deviate from this paradigm in several significant ways. For one thing, logical entailment is an absolute, all-or-nothing relationship between sentences, whereas inductive support comes in degrees of strength. For another, although the notion of *inductive support* is analogous to the deductive notion of *logical entailment*, and is arguably an extension of it, there seems to be no inductive logic extension of the notion of *logical inconsistency* — at least none that is inter-definable with *inductive support* in the way that *logical inconsistency* is inter-definable with *logical entailment*. That is, *B logically entails A* just when $(B\cdot{\sim}A)$ is *logically inconsistent*. However, it turns out that when the unconditional probability of $(B\cdot{\sim}A)$ is very nearly 0 (i.e., when $(B\cdot{\sim}A)$ is “nearly inconsistent”), the degree to which *B inductively supports A*, $P[A \mid B]$, may range anywhere between 0 and 1.

Another notable difference is that when *B logically entails A*, adding a premise *C* cannot undermine the entailment — i.e., $(C\cdot B)$ must entail *A* as well. This property of *logical entailment* is called *monotonicity*. But *inductive support* is *nonmonotonic*. In general, depending on what *A*, *B*, and *C* mean, adding a premise *C* to *B* may substantially raise the degree of support for *A*, or may substantially lower it, or may leave it completely unchanged — i.e., $P[A \mid C\cdot B]$ may have a value much larger than $P[A \mid B]$, or a much smaller value, or it may have the same, or nearly the same value as $P[A \mid B]$.

In a formal treatment of probabilistic inductive logic, inductive support is represented by conditional probability functions defined on sentences of a formal language *L*. These conditional probability functions are constrained by certain rules or axioms that are sensitive to the meanings of the logical terms (i.e., ‘not’, ‘and’, ‘or’, etc., the quantifiers ‘all’ and ‘some’, and the identity relation). The axioms apply without regard for what the other terms of the language may mean. In essence the axioms specify a family of *possible support functions*, $\{P_\beta, P_\gamma, \ldots, P_\delta, \ldots\}$ for a given language *L*. Although each support function satisfies these same axioms, the further issue of which among them provides an appropriate measure of *inductive support* is not settled by the axioms alone. That may depend on additional factors, such as the meanings of the non-logical terms in the language.

A good way to specify the rules or axioms of the logic of inductive support functions is as follows. Let *L* be a language for predicate logic with identity, and let ‘⊨’ be the standard *logical entailment* relation — i.e. the expression ‘$B \vDash A$’ says “*B logically entails A*” and the expression ‘$\vDash A$’ says “*A* is a tautology”.

> A support function is a function P_α from pairs of sentences of *L* to real numbers between 0 and 1 that satisfies the following rules or axioms:
>
> 1. $P_\alpha[D \mid E] < 1$ for at least one pair of sentences *D* and *E*.
>
> For all sentence *A*, *B*, and *C*,
>
> 2. If $B \vDash A$, then $P_\alpha[A \mid B] = 1$;
> 3. If $\vDash (B \equiv C)$, then $P_\alpha[A \mid B] = P_\alpha[A \mid C]$;
> 4. If $C \vDash {\sim}(B\cdot A)$, then $P_\alpha[(A \vee B) \mid C] = P_\alpha[A \mid C] + P_\alpha[B \mid C]$ or $P_\alpha[D \mid C] = 1$ for every *D*;
> 5. $P_\alpha[(A\cdot B) \mid C] = P_\alpha[A \mid (B\cdot C)] \times P_\alpha[B \mid C]$.

This axiomatization takes conditional probability as basic, as seems appropriate for *evidential support functions*. These functions agree with the usual unconditional probability functions when the latter are defined — just let $P_\alpha[A] = P_\alpha[A \mid (D \vee {\sim}D)]$. However, these axioms permit conditional probabilities $P_\alpha[A \mid C]$ to remain defined even when condition statement C has probability 0 (i.e., even when $P_\alpha[C \mid (D \vee {\sim}D)] = 0$).

Notice that conditional probability functions apply only to pairs of sentences, a conclusion sentence and a premise sentence. So in probabilistic inductive logic we represent finite collections of premises by conjoining them into a single sentence. Rather than say, ‘A is supported to degree r by the set of premises $\{B_1, B_2, B_3,\ldots,B_n\}$’, we say ‘$A$ is supported to degree r by the premise $(\ldots((B_1 \cdot B_2) \cdot B_3) \cdot \ldots \cdot B_n)$’, and write this as ‘$P[A \mid (\ldots((B_1 \cdot B_2) \cdot B_3) \cdot \ldots \cdot B_n)] = r$’.

Let us briefly consider each axiom, 1-5, to see how plausible it is as a constraint on a quantitative measure of inductive support, and how it extends the notion of deductive entailment. First, notice that adopting an inductive support scale between 0 and 1 is merely a convenience. This scale is usual for probabilities; but any other scale might do as well.

Rule (1) is a non-triviality requirement. It says that at least one sentence must be supported by another to degree less than 1. We might instead have required that $P_\alpha[(A \cdot {\sim}A) \mid (A \vee {\sim}A)] < 1$; but this turns out to be derivable from Rule (1) together with the other rules.

Each *degree-of-support* function P_α on L measures *support strength* with numerical values between 0 and 1, with maximal support at 1. When B *logically entail* A, the support of A based on B is maximal. This is what Rule (2) asserts. It comports with the idea that an inductive support function is a generalization of the deductive entailment relation.

Rule (3) is equally obvious. It says that whenever B is logically equivalent to C, as premises each must provide precisely the same amount of support to every conclusion.

Rule (4) says that inductive support “adds up” in a plausible way. When C logically entails the incompatibility of A and B, the support C provides each separately must sum to the support it provides for their disjunction. The only exception is in cases where C acts like a contradiction and supports all sentences to degree 1.

To understand what Rule (5) says, think of a support function P_α as describing a measure on possible worlds or possible states of affairs. ‘$P_\alpha[C \mid D] = r$’ says that the proportion of worlds in which C is true among those where D is true is r. Rule (5) then

says the following: if A is true in fraction r of worlds where B and C are true together, and if B (together with C) is true in proportion q of all the C-worlds, then A and B (and C) should be true together in fraction r of that proportion q of B (and C) worlds among the C-worlds.[2]

From these five rules all of the usual theorems of probability theory are easily derived. For example, logically equivalent sentences are always supported to the same degree: if $C \vDash (B \equiv A)$, then $P_\alpha[A \mid C] = P_\alpha[B \mid C]$. The following generalizations of the Addition Rule (4) may be proved as well:

$$P_\alpha[(A \vee B) \mid C] = P_\alpha[A \mid C] + P_\alpha[B \mid C] - P_\alpha[(A \cdot B) \mid C].$$

If $\{B_1, \ldots, B_n\}$ is any finite set of sentences such that for each pair B_i and B_j, $C \vDash {\sim}(B_i \cdot B_j)$ (i.e., the members of the set are mutually exclusive, given C), then

$$P_\alpha[((B_1 \vee B_2) \vee \ldots \vee B_n) \mid C] = \sum_{i=1}^{n} P_\alpha[B_i \mid C],$$

unless $P_\alpha[D \mid C] = 1$ for every sentence D.

If $\{B_1, \ldots, B_n, \ldots\}$ is any countably infinite set of sentences such that for each pair B_i and B_j, $C \vDash {\sim}(B_i \cdot B_j)$, then

$$\lim_n P_\alpha[((B_1 \vee B_2) \vee \ldots \vee B_n) \mid C] = \sum_{i=1}^{\infty} P_\alpha[B_i \mid C],$$

unless $P_\alpha[D \mid C] = 1$ for every sentence D.[3]

In the context of inductive logic it makes good sense to supplement the above rules with two additional rules. One is this:

6. If A is an axiom of set theory or any other piece of pure mathematics employed by the sciences, or if A is analytically truth (given the meanings of terms in L associated with support function P_α), then, for all C, $P_\alpha[A \mid C] = 1$.

The idea is that inductive logic is about evidential support for contingent claims. Nothing can count as empirical evidence against non-contingent truths. They should be maximally supported by all claims *C*.

One important respect in which inductive logic should follow the deductive paradigm is in not presupposing the truth-values of contingent sentences. No *inductive support function* P_α should permit a tautological premise to assign degree of support 1 to a contingent claim — i.e., $P_\alpha[C \mid (B \vee {\sim}B)]$ should always be less than 1 when *C* is contingent. For, the whole idea of inductive logic is to provide a measure of the extent to which contingent premise sentences indicate the likely truth-values of contingent conclusion sentences. This idea won't work properly if the truth-values of some contingent sentences are presupposed. Such presuppositions would make inductive logic enthymematic. It may hide significant premises in inductive support relationships.

However, it is common practice for probabilistic logicians to sweep provisionally accepted contingent claims under the rug by assigning them probability 1. This saves the trouble of repeatedly writing a given contingent sentence *B* as a premise, since $P_\gamma[A \mid B{\cdot}C]$ will just equal $P_\gamma[A \mid C]$ whenever $P_\gamma[B \mid C] = 1$. Although this device is useful, such probability functions should be considered mere abbreviations of proper, logically explicit, non-enthymematic, inductive support functions. Thus, properly speaking, an inductive support function $P_\alpha$should not assign probability 1 to a sentence *relative to all possible premises* unless that sentence is either (i) logically true, or (ii) an axiom of set theory or some other piece of pure mathematics employed by the sciences, or (iii) unless according to the interpretation of the language that $P_\alpha$presupposes, the sentence is *analytic*, and so outside the realm of evidential support. Thus, we adopt the following version of the so-called "axiom of regularity".

7. If, for all *C*, $P_\alpha[A \mid C] = 1$, then *A* is a logical truth or an axiom of set theory or some other piece of pure mathematics employed by the sciences, or *A* is analytically true (according to the meanings of the terms of *L* as represented in P_α).

This is more a convention than an axiom. Taken together with (6) it tells us that a support function P_α counts as non-contingently true just those sentences that it assigns probability 1 on every premise.

Some Bayesian logicists (e.g. Carnap) thought that inductive logic might be made to depend solely on the logical form of sentences, just like deductive logic. The idea was, effectively, to supplement axioms 1–7 with additional axioms that depend only on the logical structures of sentences, and to introduce enough such axioms to reduce the number of possible support functions to a single uniquely best confirmation function. It is now widely agreed that this project cannot be carried out in a plausible way. Perhaps there are additional rules that should be added to 1–7. But it is doubtful such rules can suffice to specify a single, uniquely qualified support function based only on logical structure. We will se why in Section 3, but only after first seeing how inductive probabilities capture the relationship between hypotheses and evidence.

2.3 Two Conceptions of Inductive Probability

Axioms 1–7 for conditional probability functions merely place formal constraints on what may properly count as a *degree of support function*. Each function P_α satisfying these rules may be viewed as a possible way of applying the notion of *inductive support* to a language *L* that respects the meanings of the logical terms, much as each possible *truth-value assignment* for a language represents a possible way of assigning truth-values to its sentences in a way that respects the semantic rules expressing the meanings of the logical terms. The issue of which of the *possible* truth-value assignments to a language represents the *actual* truth or falsehood of its sentences depends on more than this — it depends on the meanings of the non-logical terms and on the state of the actual world. Similarly, the degree to which some sentences *actually* support others in a fully meaningful language must rely on something more than merely satisfying the axioms for support functions. It must, at least, rely on what the sentences of the language mean, and perhaps on much more besides. But, what more? Various "interpretations of probability", which offer accounts of how support functions are to be understood, may help by filling out our conception of what *inductive support* is really about. There are two prominent views.

One reading is to take each P_α as a measure on possible worlds, or possible states of affairs. The idea is that, given a fully meaningful language (and, perhaps relative to the inferential inclinations of a particular agent, α) '$P_\alpha[A \mid B] = r$' says that among the worlds in which *B* is true, *A* is true in proportion *r* of them. There will generally not be a single privileged way to define such a measure on possible worlds. Rather, it may be that each of a number of functions P_α, P_β, P_γ, …, etc., satisfying the constraints imposed by axioms 1-7 can represent a viable measure of the *inferential import* of propositions expressed by sentences of the language. This idea needs more fleshing out, of course. The next section will give some indication of how that might go.

Subjectivist Bayesians offer an alternative reading of the support functions. First, they usually take unconditional probability as basic, and they take conditional probabilities as defined in terms of them: the conditional probability '$P_\alpha[A \mid B]$'is defined as a ratio of unconditional probabilities, $P_\alpha[A \cdot B]/P_\alpha[B]$.*Subjectivist Bayesians* take each unconditional probability function P_α to represent the belief-strengths or confidence-strengths of an ideally rational agent, α. On this understanding '$P_\alpha[A] = r$' says, "the strength of α's belief (or confidence) that *A* is truth is *r*." Subjectivist Bayesians usually tie such belief strengths to what the agent would be willing to bet on *A* turning out to be true. Roughly, the idea is this. Suppose that an ideally rational agent α would be willing to accept a wager that would yield (no less than) $*u* if *A* turns out to be true and would lose him $1 if *A* turns out to be false. Then, under reasonable assumptions about how much he desires money, it can be shown that his belief strength that *A* is true should be $P_\alpha[A] = 1/(u+1)$. And it can further be shown that any function P_α that expresses such betting-related belief-strengths on all statements in agent α's language must satisfy axioms for unconditional probabilities analogous to axioms 1–5.[4] Moreover, it can be shown that any function P_β that satisfies these axioms is a possible rational belief function for some ideally rational agent β. These relationships between belief-strengths and the desirability of outcomes (e.g., gaining money or goods on bets) are at the core of *subjectivist Bayesian decision theory*. *Subjectivist Bayesians* usually take *inductive probability* to just *be* this notion of *probabilistic belief-strength*.

Undoubtedly real agents do believe some claims more strongly than others. And, arguably, the belief strengths of real agents can be measured on a probabilistic scale between 0 and 1, at least approximately. And clearly the inductive support of evidence for hypotheses should influence the strength of an agent's belief in those hypotheses. However, there is good reason for caution about viewing *inductive support functions* as Bayesian belief-strength functions, as we will see a bit later. So, perhaps an agent's support function is not simply *identical to* his belief function, and perhaps the relationship between *inductive support* and *belief-strength* is somewhat more complicated.

In any case, some account of what support functions are supposed to represent is clearly needed. The belief function account and the possible worlds account are two attempts to provide this. Let us put this interpretative issue aside for now. One may be

able to get a better handle on what inductive support functions *really are* after one sees how the inductive logic that draws on them is supposed to work.

3. The Application of Inductive Probabilities to the Evaluation of Scientific Hypotheses

One of the most important applications of a formal inductive logic is to the confirmation or refutation of scientific hypotheses. The logic should explicate the notion of evidential support for all sorts of hypotheses, ranging from simple diagnostic claims (e.g., "the patient is infected with the HIV") to scientific theories about the fundamental nature of the world, like quantum mechanics or the theory of relativity. We'll now look into how support functions (a.k.a. confirmation functions) represent the logic of hypothesis confirmation. This kind of inductive logic is often referred to as *Bayesian Confirmation Theory*.

Consider some exhaustive set of mutually incompatible hypotheses or theories about some subject matter, $\{h_1, h_2, \ldots\}$. The set of alternatives may be very simple, e.g., {"the patient has HIV", "the patient is free of HIV"}. Or, when the physician is trying to determine which among a range of diseases is causing the patient's symptoms, the alternative hypotheses may consist of a long list of possible diseases. For the cosmologist the alternatives may be a list of several alternative gravitational theories, or several versions of the "same theory". Where inductive logic is concerned, even a slightly different version of a given theory will count as a distinct theory if it differs from the original in empirical import. (This should not be confused with the converse claim, which is the positivistic assertion that theories with the same empirical content are really the same theory. Inductive logic doesn't require you to buy that!)

In general there may be finitely or infinitely many such alternatives under consideration. They may all be considered at once, or they may be constructed and compared over a long historical period. One may even think of the set of alternative hypotheses as consisting of all logically possible alternatives expressible in a given language about a given subject matter — e.g., all possible theories of the origin and evolution of the universe expressible in English and mathematics. Although testing every possible alternative may pose practical challenges, it turns out that the logic works much the same way in the logically ideal case as it does in realistic cases.

If the set of alternative hypotheses is finite, it may contain a *catch-all hypothesis* h_K that says that none of the other hypotheses are true (e.g., "none of the other known diseases is present"). When only some number u of explicit alternative hypotheses is under consideration, h_K is just the sentence $(\sim h_1 \cdot \ldots \cdot \sim h_u)$.

Evidence for scientific hypotheses consists of the results of specific experiments or observations. For a given experiment or observation, let '*c*' represent a description of the relevant *conditions* under which it is performed, and let '*e*' represent a description of the result, the *evidential outcome* of conditions *c*.

Scientific hypotheses often require the mediation of background knowledge and auxiliary hypotheses to help them express claims about evidence. Let '*b*' represent all background and auxilliary hypothese not at issue in the assessment of the hypotheses h_i, but that mediate their implications about evidence. In cases where a hypothesis is deductively related to evidence, either $h_i \cdot b \cdot c \vDash e$ or $h_i \cdot b \cdot c \vDash \sim e$.

For example, h_i might be the Newtonian Theory of Gravitation. A test of the theory might involve a condition statement *c* describing the results of some earlier measurements of Jupiter's position, and describing the means by which the next position measurement will be made; the outcome description *e* states the result of this additional position measurement; and the background information (or auxiliary hypotheses) *b* might state some already well confirmed theory about the workings and accuracy of the devices used to make the position measurements. Thus, if $(c \cdot e)$ occurs, this may be considered good evidence for h_i, given *b*, as the *hypothetico-deductive* account of confirmation maintains. On the other hand, if from $h_i \cdot b \cdot c$ we calculate some outcome incompatible with *e*, then $h_i \cdot b \cdot c \vDash \sim e$. In that case from deductive logic alone we get that $b \cdot c \cdot e \vDash \sim h_i$, and h_i is said to be *falsified* by $b \cdot c \cdot e$.

Duhem (1906) and Quine (1953) are generally credited with alerting inductive logicians to the importance of auxiliary hypotheses. They point out that scientific hypotheses often make little contact with evidence claims on their own. Rather, most scientific hypotheses only make testable predictions relative to background claims or auxiliary hypotheses that tie them to that evidence. Typically auxiliaries are highly confirmed hypotheses from other scientific domains. They often describe the operating characteristics of various devices (e.g. measuring instruments) used to make observations or conduct experiments. They are usually not at issue in the testing of h_i against its competitors, because h_i and its alternatives usually rely on the same

auxiliary hypotheses to tie them to the evidence. But even when an auxiliary hypothesis is already well-confirmed, we cannot simply assume that it is unproblematic, or just *known to be true*. Rather, the evidential support or refutation of a hypothesis *h* is *relative to* whatever auxiliaries and background information (in *b*) is being supposed. In other contexts the auxiliary hypotheses used to test h_i may themselves be among a collection of alternative hypotheses that are themselves subject to evidential support or refutation. (Furthermore, to the extent that competing hypotheses employ different auxiliary hypotheses in accounting for evidence, the evidence only tests each such hypothesis in conjunction with its distinct auxiliaries against alternative hypotheses packaged with their distinct auxiliaries.) Thus, what counts as a *hypothesis to be tested*, h_i, and what counts as auxiliary hypotheses and background information, *b*, and even to some extent what counts as the conditions *c* for an experiment or observation, will always depend on the epistemic context — on what alternative hypotheses are being tested by the same experiments or observations, and on what claims are being presupposed or held fixed for present purposes, and on what claims are considered to be immediate preconditions for the evidential outcome *e*. No statement is intrinsically a *hypotheis*, or intrinsically an *auxiliary* (or a *background condition*), or intrinsically an *evidential condition*. Rather, those are roles statements may play in an epistemic context, and the very same statement may play different roles in different confirmational contexts.

In a probabilistic inductive logic the degree to which evidence $c \cdot e$ supports a hypothesis h_i relative to background *b* is represented by the *posterior probability* of h_i, $P_\alpha[h_i \mid b \cdot c^n \cdot e^n]$. It turns out that the *posterior probability* of a hypothesis depends on just two kinds of factors: (1) its *prior probability*, $P_\alpha[h_i \mid b]$, together with the prior probabilities of its competitors, $P_\alpha[h_j \mid b]$, etc.; and (2) the *likelihood* of evidential outcomes *e* according to h_i, given that *b* and *c* are true, $P[e \mid h_i \cdot b \cdot c]$, together with the likelihoods of outcomes according to its competitors, $P[e \mid h_j \cdot b \cdot c]$, etc. In this section we will first examine each of these two kinds of factors in some detail, and then see precisely how the values of posterior probabilities depend on them.

3.1 Likelihoods

In probabilistic inductive logic *the likelihoods* carry the empirical import of hypotheses. A *likelihood* is a support function probability of form $P[e \mid h_i \cdot b \cdot c]$. It expresses how likely it is that outcome e will occur according to hypothesis h_i.[5] If a hypothesis together with auxiliaries and observation conditions deductively entails an evidence claim, the axioms of probability make the corresponding likelihood objective in the sense that every support function must agree on its values: i.e., $P[e \mid h_i \cdot b \cdot c] = 1$ if $h_i \cdot b \cdot c \vDash e$; $P[e \mid h_i \cdot b \cdot c] = 0$ if $h_i \cdot b \cdot c \vDash {\sim}e$. However, in many cases the hypothesis h_i will not be deductively related to the evidence, but will only imply it probabilistically. There are (at least) two ways this might happen. Either h_i may itself be an explicitly probabilistic or statistical hypothesis, or it may be that an auxiliary statistical hypothesis, as part of background b, connects h_i to the evidence. Let's briefly consider examples of each.

A blood test for HIV has a known false-positive rate and a known true-positive rate. Suppose the false positive rate is .05 — i.e., the test incorrectly shows the blood sample to be positive for HIV in 5% of all cases where HIV is not present. And suppose the true-positive rate is .99 — i.e., the test correctly shows the blood sample to be positive for HIV in 99% of all cases where HIV really is present.When a particular patient's blood is tested, the hypotheses under consideration are ‘the patient is infected with HIV’, h, and ‘the patient is not infected with HIV’, ${\sim}h$. In this context the known test characteristics function as background information, b. The experimental condition c merely states that this patient was subjected to a blood test for HIV, which was processed by the lab in the usual way. Let us suppose that the outcome e states that the result is positive for HIV. The relevant likelihoods, then, are $P[e \mid h \cdot b \cdot c] = .99$ and $P[e \mid {\sim}h \cdot b \cdot c] = .05$.

In this example the values of the likelihoods are entirely due to the statistical characteristics of the accuracy of the test, which is carried by the background information b. The hypothesis h being tested is not itself statistical.

This kind of situation may, of course, arise for much more complex hypotheses. The hypothesis of interest may be some deterministic physical theory, say Newtonian Gravitation Theory. Some of the experiments that test this theory relay on somewhat imprecise measurements that have known statistical error characteristics, which are expressed as part of the background or auxiliary hypotheses b. For example, the auxiliary b may describe the error characteristics of a device that measures the torque imparted to a quartz fiber, where the measured torque is used to assess the strength of the gravitational force between test masses. In that case b may say that for this kind of device the measurement errors are normally distributed about whatever value a given gravitational theory predicts, with some specified standard deviation that is characteristic of the device. This results in specific values r_i for the likelihoods, $P[e \mid h_i \cdot b \cdot c] = r_i$, for each of the various alternative gravitational theories h_i being tested.

On the other hand, the hypotheses being tested may themselves be statistical in nature. One of the simplest examples of statistical hypotheses and their role in likelihoods are hypotheses about the chance characteristic of coin-tossing. Let $h_{[r]}$ be a hypothesis that says a specific coin has a propensity r (e.g., 1/2) for coming up heads on normal tosses, and that such tosses are probabilistically independent of one another. Let c state that the coin is tossed n times in the normal way; and let e say that on these tosses the coin comes up heads m times. In cases like this the value of the likelihood of the outcome e on hypothesis h for condition c is given by the well-known binomial term:

$$P[e \mid h_{[r]} \cdot b \cdot c] = \frac{n!}{m! \times (n-m)!} \times r^m (1-r)^{n-m}.$$

There are, of course, more complex cases of likelihoods involving statistical hypotheses. Consider, for example, the hypothesis that plutonium 233 nuclei have a half-life of 20 minutes — i.e., the propensity for a Pu-233 nucleus to decay within a 20 minute period is 1/2. This hypothesis, h, together with background b about decay products and the efficiency of the equipment used to detect them (which may itself be an auxiliary statistical hypothesis), yields precisely calculable values for likelihoods $P[e_k \mid h \cdot b \cdot c]$ of possible outcomes of the experimental arrangement.

Likelihoods that arise from explicit statistical claims — either within the hypotheses being tested, or from explicit statistical background claims that tie the hypotheses to the evidence — are often called *direct inference likelihoods*. Such likelihoods are completely objective. So it seems reasonable to suppose that all support functions should agree on their values, just as all support functions agree on likelihoods when evidence is logically entailed. Direct inference likelihoods are *logical* in an extended, non-deductive sense. Indeed, some logicians have attempted to spell out the logic of *direct inferences* in terms of the logical form of the sentences involved.[6] But regardless of whether that project succeeds, it seems reasonable to take likelihoods of this sort to have highly objective or intersubjectively agreed values.

Not all likelihoods of interest in confirmational contexts are warranted deductively or by explicitly stated statistical claims. Nevertheless, the likelihoods that relate hypotheses to evidence in scientific contexts should often have objective or intersubjectively agreed values. So, although a variety of different support functions P_α, P_β ,…, P_γ, etc., may be needed to represent the differing "inductive proclivities" of the various members of a scientific community, all should agree, at least approximately, on the values of the likelihoods. For, likelihoods represent the empirical content of a hypothesis, what the hypothesis (together with background b) *probabilistically implies* about the evidence. Thus, the empirical objectivity of a science relies on a high degree of objectivity or intersubjective agreement among scientists on the numerical values of likelihoods.

To see the point more vividly, imagine what a science would be like if scientists disagreed widely about the values of likelihoods. Each practitioner interprets a theory to *say* quite different things about how likely it is that various possible evidence statements will turn out to be true. Whereas scientist α takes theory h_1 to probabilistically imply that event e is highly likely, his colleague β understands the empirical import of h_1 to say that e is very unlikely. And, conversely, α takes competing theory h_2 to probabilistically imply that e is quite unlikely, whereas β reads h_2 to say that e is very likely. So, for α the evidential outcome e supplies strong support for h_1 over h_2, because $P_\alpha[e \mid h_1 \cdot b \cdot c] \gg P_\alpha[e \mid h_2 \cdot b \cdot c]$. But his colleague β takes outcome e to show just the opposite — that h_2 is strongly supported over h_1 — because $P_\beta[e \mid h_1 \cdot b \cdot c] \ll P_\beta[e \mid h_2 \cdot b \cdot c]$. If this kind of thing were to occur often or for significant evidence claims in a scientific domain, it would make a shambles of the empirical objectivity of that science. It would completely undermine the empirical testability of its hypotheses and theories. Under such circumstances, although each scientist employs the same *theoretical sentences* to express a given theory h, each understands the *empirical import* of these sentences so differently that h as understood by α is an empirically different theory than h as understood by β. Thus, the empirical objectivity of the sciences requires that experts should be in close agreement about the values of the likelihoods.[7]

For now we will suppose that the likelihoods have objective or intersubjectively agreed values, common to all agents in a scientific community. Let us mark this agreement by dropping the subscript 'α', 'β', etc., from expressions that represent likelihoods. One might worry that this supposition is overly strong. There are many legitimate scientific contexts where, although scientists should have enough of a common understanding of the empirical import of hypotheses to assign quite similar values to likelihoods, precise agreement on the numerical values is unrealistic. This point is right. So later we will see how to relax the supposition that likelihood values agree precisely. But for now the main ideas behind probabilistic inductive logic will be more easily explained if we focus on those contexts were objective or intersubjectively agreed likelihoods are available. Later we will see that much the same logic continues to apply in contexts where the values of likelihoods may be somewhat vague, or where members of the scientific community disagree to some extent about their values.

An adequate treatment of the likelihoods calls for the introduction of one additional notational device. Scientific hypotheses are generally tested by a sequence of experiments or observations conducted over a period of time. To explicitly represent the accumulation of evidence, let the series of sentences c_1, c_2, …, c_n, describe the conditions under which a sequence of experiments or observations are conducted. And let the corresponding outcomes of these observations be represented by sentences e_1, e_2,…,e_n. We will abbreviate the conjunction of the first n descriptions of experimental or observation conditions as 'c^n', and abbreviate the conjunction of descriptions of their outcomes as 'e^n'. Then, for a stream of n observations or experiments and their outcomes, the likelihoods take form $P[e^n \mid h_i \cdot b \cdot c^n] = r$, for appropriate r between 0 and 1. In many cases the likelihood of the evidence stream will be equal to the product of the likelihoods of the individual outcomes:

$$P[e^n \mid h_i \cdot b \cdot c^n] = P[e_1 \mid h_i \cdot b \cdot c_1] \times \ldots \times P[e_n \mid h_i \cdot b \cdot c_n].$$

When this equality holds the individual bits of evidence are said to be *probabilistically independent on the hypothesis*. In what follows such *independence* will only be assumed in those places where it is explicitly invoked.

3.2 Posterior Probabilities and Prior Probabilities

In the probabilistic logic of evidential support the evaluation of a hypothesis on evidence is represented by its *posterior probability*, $P_\alpha[h_i \mid b \cdot c^n \cdot e^n]$. The posterior probability represents the net plausibility of the hypothesis resulting from the combination of the evidence together with any relevant non-evidential plausibility considerations (which should be packaged within b). The likelihoods are the means through which evidence contributes to posterior probabilities. But another factor, the *prior probability* of the hypothesis based on considerations expressed within b, $P_\alpha[h_i \mid b]$, also makes a contribution. It represents the weight of all non-evidential plausibility considerations on which posterior probabilities may depend. It turns out that posterior probabilities depend *only* on the values of (ratios of) likelihoods *and* on the values of (ratios of) prior probabilities.

To understand the role of *prior probabilities*, consider the *HIV test* example described in the previous section. What the physician and patient want to know is the value of the posterior probability $P_\alpha[h \mid b \cdot c \cdot e]$ that the patient has HIV, h, given the evidence of the positive test, $c \cdot e$, and given the error rates of the test, described within b. The value of this posterior probability depends on the likelihood (due to the error rates) of this patient obtaining a true-positive result, $P[e \mid h \cdot b \cdot c] = .99$, and of obtaining a false positive result, $P[e \mid {\sim}h \cdot b \cdot c] = .05$. In addition, the value of the of the posterior probability depends on how plausible it is that the patient has HIV before the test results are taken into account, $P_\alpha[h \mid b]$. In the context of medical diagnosis this prior probability is sometimes called the *base rate*. It represents the probability that the patient may have contracted HIV based on his risk group (i.e., whether he is an IV drug user, has unprotected sex with multiple partners, etc.). Such information should be explicitly stated, and represented within b as well. To see its importance, consider the following numerical results (which may be calculated using the formula called Bayes' Theorem, presented in the next section). If the base rate for the patient's risk group is relatively high, say $P_\alpha[h \mid b] = .10$, then the positive test result yields a probability for his having HIV of $P_\alpha[h \mid b \cdot c \cdot e] = .69$. However, if the patient is in a very low risk group, $P_\alpha[h \mid b] = .001$, then a positive test only raises the probability of HIV infection to $P_\alpha[h \mid b \cdot c \cdot e] = .02$. This posterior probability is much higher than the prior probability of .001, but should not worry the patient too much. This positive test result is more likely due to the false-positive rate of the test than to the presence of HIV. (This sort of test, with such a large false-positive rate, .05, is best used as a

screening test; a positive result should lead to a second, more rigorous, more expensive test.)

In the evidential evaluation of scientific theories, prior probabilities often represent assessments by agents of non-evidential, conceptually motivated *plausibility weightings* among hypotheses. However, because such plausibility assessments tend to vary among agents, critics often brand them as *merely subjective*, and take their role in probabilistic induction to be highly problematic. Bayesian inductivists counter that such assessments often play an important role in the sciences, especially when there is insufficient evidence to distinguish among some of the alternative hypotheses. And, they argue, the epithet *merely subjective* is unwarranted. Such plausibility assessments are often backed by extensive arguments that may draw on forceful conceptual considerations.

Consider, for example, the kind of plausibility arguments that have been brought to bear on the various interpretations of quantum theory (e.g., those related to the measurement problem). These arguments go to the heart of conceptual issues that were central to the development of the theory. Indeed, many of these issues were first raised by the scientists who made the greatest contributions to the theory's development, in the attempt to get a conceptual hold on the theory and its implications. Such arguments seem to play a legitimate role in the assessment of alternative views when distinguishing evidence has yet to be found.

More generally, scientists often bring plausibility arguments to bear in assessing their views. Although such arguments are seldom decisive, they may bring the scientific community into widely shared agreement, especially regarding the *implausibility* of some logically possible alternatives. This seems to be the primary epistemic role of the thought experiment. Thus, although prior probabilities may be subjective in the sense that agents may disagree on the relative strengths of plausibility arguments — and so disagree on the comparative plausibilities of various hypotheses — the priors used in scientific contexts should not represent *mere subjective whims*. Rather, they should be supported (or at least be supportable) by explicit arguments regarding how much more plausible one hypothesis is than another. The important role of plausibility assessments is apparent in such received bits of scientific wisdom as the old saw that *extraordinary claims require extraordinary evidence*. That is, it takes especially strong evidence, in the form of extremely high values for ratios of likelihoods, to overcome the extremely low plausibility values possessed by some hypotheses. We'll see precisely how this idea works in the next section, and return to it again in Section 3.5.

When sufficiently strong evidence becomes available, it turns out that prior plausibility assessments may be “washed out” or overridden by the evidence. We'll see how this works in Sections 4 and 5. Thus, prior plausibility assessments play their most important role when the kind of evidence for which hypotheses specify likelihoods is still fairly sparse. It will be shown that provided the value of the prior probability of a true hypothesis isn't assessed to be zero, as evidence accumulates the influence of the values of the prior probabilities will *very probably* fade away as evidence accumulates.

Some Bayesian logicists (e.g. Carnap) have maintained that posterior probabilities of hypotheses should be determined by logical form alone. The idea is that the likelihoods might reasonably be specified in terms of logical form; so if logical form might be made to determine the values of prior probabilities as well, then inductive logic would be fully “formal” in the same way that deductive logic is “formal”. Keynes and Carnap tried to implement this idea through syntactic versions of the principle of indifference — the idea that syntactically similar hypotheses should be assigned the same prior probability values. Carnap showed how to carry out this project in detail, but only for extremely simple formal languages. Most logicians now take the project to have failed because of a fatal flaw with the whole idea that reasonable prior probabilities can be made to depend on logical form alone. Semantic content should matter. Goodmanian grue-predicates provide one way to illustrate the point.[8] Furthermore, as suggested earlier, for this idea to apply to the evidential support of real scientific theories, scientists would have to assess the prior probabilities of each alternative theory based only on its syntactic structure. That seems an unreasonable way to proceed. Are we to evaluate the prior probabilities of alternative theories of gravitation, or of alternative quantum theories, by exploring only their syntactic structures, with absolutely no regard for their semantic content — with *no* regard for what they *say* about the world? This seems an extremely dubious approach to the evaluation of real scientific theories. Logical structure alone cannot, and should not suffice for determining reasonable prior probability values for real scientific theories. Moreover, real scientific hypotheses and theories are inevitably subject to plausibility considerations based on what they *say* about the world. Prior probabilities are much better suited to representating of the weight of such plausibility considerations, vague as they may be..

We will return to prior probabilities in a bit. But first let's see how likelihoods combine with prior probabilities to yield posterior probabilities for hypotheses.

3.3 Bayes' Theorem

Any *probabilistic inductive logic* that draws on the usual axioms of probability theory to represent how evidence supports hypotheses must be a *Bayesian inductive logic* in the broad sense. For, Bayes' Theorem is just a simple theorem of probability theory. Its importance is due to the relationship it expresses between hypotheses and evidence. The theorem shows how, through the likelihoods, evidence combines with prior plausibility assessments to produce posterior plausibility values for hypotheses. Thus, a logics of hypothesis evaluation of this sort is called a *Bayesian Confirmation Theory*.

Let's now examine several forms of Bayes' Theorem, each derivable from axioms 1–5. The simplest is this:

Bayes' Theorem: Simple Form

$$(8)\quad P_\alpha[h_i \mid b{\cdot}c^n{\cdot}e^n] = \frac{P[e^n \mid h_i{\cdot}b{\cdot}c^n] \times P_\alpha[h_i \mid b]}{P_\alpha[e^n \mid b{\cdot}c^n]} \times \frac{P_\alpha[c^n \mid h_i{\cdot}b]}{P_\alpha[c^n \mid b]}$$

$$= \frac{P[e^n \mid h_i{\cdot}b{\cdot}c^n] \times P_\alpha[h_i \mid b]}{P_\alpha[e^n \mid b{\cdot}c^n]} \quad \text{if } P_\alpha[c^n \mid h_i{\cdot}b] = P_\alpha[c^n \mid b].$$

This equation expresses the posterior probability of h_i, $P_\alpha[h_i \mid b{\cdot}c^n{\cdot}e^n]$, in terms of the *likelihood* of the evidence on the hypothesis (together with background and observation conditions), $P[e^n \mid h_i{\cdot}b{\cdot}c^n]$, the *prior probability* of the hypothesis (given background conditions), $P_\alpha[h_i \mid b]$, and the *simple probability* of the evidence (given background and observation conditions), $P_\alpha[e^n \mid b{\cdot}c^n]$. This latter probability is sometimes called the *expectedness of the evidence*.

This version of Bayes' Theorem also includes a term, $(P_\alpha[c^n \mid h_i{\cdot}b] \,/\, P_\alpha[c^n \mid b])$, that represents the ratio of the *likelihood of the experimental conditions* on the hypothesis and background to the *“likelihood” of the experimental conditions* on the background alone. Bayes' Theorem is usually expressed in a way that suppresses this factor by building c^n into the background b. However, if c^n is built into b, then technically b

must change as new evidence is accumulated. So it is better to make this factor explicit and see how to deal with it logically. Arguably the term ($P_\alpha[c^n \mid h_i{\cdot}b] / P_\alpha[c^n \mid b]$) should be 1, or be very near 1, since the truth of the hypothesis at issue should not significantly affect how likely it is that the experimental conditions are satisfied. If various alternative hypotheses assign significantly different likelihoods to the experimental conditions, then such conditions should more properly be included in the evidential outcomes e^n.

Both the *prior probability* of the hypothesis and the *expectedness* tend to be somewhat subjective factors in that various agents from the same scientific community may legitimately disagree on what values these factors should take. Bayesian logicians usually accept the subjectivity of the prior probabilities of hypotheses, but find the subjectivity of the *expectedness* to be more troubling. This is due at least in part to the the fact that in a Bayesian logic of evidential support the value of the expectedness cannot be independendent of likelihoods and the prior probabilities of hypotheses. That is, when for each member of a set of alternative hypotheses the likelihood $P[e^n \mid h_j{\cdot}b{\cdot}c^n]$ has an objective (or intersubjectively agreed) value, the *expectedness* is constrained by the following equation (where the sum ranges over a mutually exclusive and exhaustive set of alternative hypotheses $\{h_1, h_2, \ldots, h_m, \ldots\}$, which may be finite or infinite):

$$\begin{aligned} P_\alpha[e^n \mid b{\cdot}c^n] &= \textstyle\sum_j P[e^n \mid h_j{\cdot}b{\cdot}c^n] \times P_\alpha[h_j \mid b{\cdot}c^n] \\ &= \textstyle\sum_j P[e^n \mid h_j{\cdot}b{\cdot}c^n] \times P_\alpha[h_j \mid b] \\ &\quad \text{if } c^n \text{ is irrelevant to each hypothesis } h_j \text{ given } b. \end{aligned}$$

The first line implies that the value of the *expectedness* must at least lie between the largest and smallest of the various likelihood values based on specific hypotheses. The second line shows that the values for the prior probabilities together with the values of the likelihoods should uniquely determine the value for the *expectedness* of the evidence. This result reflects the intuitive idea that, according to an evidential support function, evidence claims are not "simply likely" to a certain degree on their own, independently of what any hypothesis has to say. Rather, the likelihoods of evidence claims are most fundamentally fixed relative to relevant hypotheses. Furthermore, the *expectedness* can only be calculated in this way in cases where *every* alternative hypothesis to h_i already figured out. Otherwise, although the *expectedness*is constrained in principle, but there is no way to figure out what its value should be.

However, this troublesome *expectedness of the evidence* term is easily sidestepped.

The subjective *expectedness* term may be circumvented by considering a ratio form of Bayes' Theorem, a form that compares hypotheses one pair at a time:

Bayes' Theorem: Ratio Form

$$\begin{aligned} (9)\quad \frac{P_\alpha[h_j \mid b{\cdot}c^n{\cdot}e^n]}{P_\alpha[h_i \mid b{\cdot}c^n{\cdot}e^n]} &= \frac{P[e^n \mid h_j{\cdot}b{\cdot}c^n]}{P[e^n \mid h_i{\cdot}b{\cdot}c^n]} \times \frac{P_\alpha[h_j \mid b]}{P_\alpha[h_i \mid b]} \times \frac{P_\alpha[c^n \mid h_j{\cdot}b]}{P_\alpha[c^n \mid h_i{\cdot}b]} \\ &= \frac{P[e^n \mid h_j{\cdot}b{\cdot}c^n]}{P[e^n \mid h_i{\cdot}b{\cdot}c^n]} \times \frac{P_\alpha[h_j \mid b]}{P_\alpha[h_i \mid b]} \quad \text{if } \frac{P_\alpha[c^n \mid h_j{\cdot}b]}{P_\alpha[c^n \mid h_i{\cdot}b]} = 1 \\ &= \frac{P[e_1 \mid h_j{\cdot}b{\cdot}c_1]}{P[e_1 \mid h_i{\cdot}b{\cdot}c_1]} \times \ldots \times \frac{P[e_n \mid h_j{\cdot}b{\cdot}c_n]}{P[e_n \mid h_i{\cdot}b{\cdot}c_n]} \times \frac{P_\alpha[h_j \mid b]}{P_\alpha[h_i \mid b]} \end{aligned}$$

if $P_\alpha[c^n \mid h_j{\cdot}b] / P_\alpha[c^n \mid h_i{\cdot}b] = 1$ and relative to each hypothesis the evidential events are probabilistic independent of one another.

The condition '$P_\alpha[c^n \mid h_j{\cdot}b] / P_\alpha[c^n \mid h_i{\cdot}b] = 1$' says that c^n is no more likely on $h_i{\cdot}b$ than on $h_j{\cdot}b$ — i.e., that neither hypothesis makes the occurrence of experimental or observation conditions more likely than the other.[9]

This Ratio Form of Bayes' Theorem expresses how much more plausible, on the evidence, one hypothesis is than another. Notice that the *likelihood ratios* carry the full import of the evidence. The evidence influences the evaluation of hypotheses in no other way. Alsi notice that the only element affecting the ratio of posterior probabilities that may not be fully objectively determinate is the ratio of prior probabilities.

This version of Bayes's Theorem shows that to evaluate the *posterior probability ratios of hypotheses*, the prior probabilities of hypotheses need not be evaluated absolutely; only their ratios are needed. That is, with regard to the priors, the Bayesian evaluation of hypotheses only relies on *how much more plausible* one hypothesis is than another (due to the considerations expressed within b). The Bayesian evaluation of hypotheses is *essentially comparative* in that only *ratios of likelihoods* and *ratios of prior probabilities* are ever really needed for the assessment of scientific hypotheses. Furthermore, we will soon see that the absolute values of the posterior probabilities entirely derive from the posterior probability ratios provided by the Ratio Form of Bayes' Theorem.

Let's consider a simple example of how the Ratio Form of the theorem is utilized. Suppose we possess a warped coin and want to determine its propensity for *heads* when tossed in the usual way. We may compare two hypotheses, $h_{[q]}$ and $h_{[r]}$, that propose that the propensity for the coin to come up *heads* on the usual kind of toss is q and r, respectively. Let c^n report that the coin is tossed n times in the normal way, and let e^n report a total m *heads*. Supposing that the outcomes of tosses are probabilistically independent relative to each of the two hypotheses, line 3 of Equation (9) yields the following equation, where the likelihood ratio is the ratio of the respective binomial terms:

$$\frac{P_\alpha[h_{[q]} \mid b\cdot c^n\cdot e^n]}{P_\alpha[h_{[r]} \mid b\cdot c^n\cdot e^n]} = \frac{q^m\,(1-q)^{n-m}}{r^m\,(1-r)^{n-m}} \times \frac{P_\alpha[h_{[q]} \mid b]}{P_\alpha[h_{[r]} \mid b]}$$

When, for instance, the coin is tossed $n = 100$ times and comes up *heads* $m = 72$ times, the evidence for hypothesis $h_{[1/2]}$ as compared to $h_{[3/4]}$ is given by the likelihood ratio $[(1/2)^{72}(1/2)^{28}]/[(3/4)^{72}(1/4)^{28}] = .000056269$. So, even if prior to the evidence, plausibility considerations (expressed within b) make it 100 times more plausible that the coin is fair than that it is warped towards heads with propensity 3/4 — i.e., even if $P_\alpha[h_{[1/2]} \mid b] / P_\alpha[h_{[3/4]} \mid b] = 100$ — the evidence provided by these tosses makes the posterior plausibility that the coin is fair only about $6/1000^{\text{th}}$ as plausible as the hypothesis that it is warped towards heads with propensity 3/4 — i.e., $P_\alpha[h_{[1/2]} \mid b\cdot c^n\cdot e^n] / P_\alpha[h_{[3/4]} \mid b\cdot c^n\cdot e^n] = .0056269$. Thus, such evidence *strongly refutes* the "fairness hypothesis" relative to the "3/4-*heads*-propensity hypothesis", provided the assessment of prior probabilities (i.e. prior plausibilities) doesn't make the latter hypothesis *too extremely implausible* to begin with. Notice, however, that *strong refutation* is not *absolute refutation*. Additional evidence could reverse the trend towards the strong refutation of the "fairness hypothesis".

This example employs repetitions of the same kind of experiment — repeated tosses of a coin. But the point holds more generally. If, as the evidence increases, the *likelihood ratios* $P[e^n \mid h_j\cdot b\cdot c^n] / P[e^n \mid h_i\cdot b\cdot c^n]$ approach 0, then the Ratio Form of Bayes' Theorem, Equation 9, shows that the posterior probability of h_j must approach 0 as well. The evidence comes to strongly refute h_j with little regard for its prior plausibility value. Indeed, Bayesian induction turns out to be a version of *eliminative induction*, and Equation 9 begins to illustrate this. For, suppose that h_i is the true hypothesis, and consider what happens to *each* of its false competitors, h_j. If enough evidence becomes available to drive each of the likelihood ratios $P[e^n \mid h_j\cdot b\cdot c^n] / P[e^n \mid h_i\cdot b\cdot c^n]$ toward 0 (as n increases), then Equation 9 says that each false h_j will become effectively refuted — each of their posterior probabilities approaches 0. As a result, the posterior probability of h_i must approach 1. The next two equations show precisely how this works.

If we sum the ratio versions of Bayes' Theorem in Equation 9 over all alternatives to hypothesis h_i (including the catch-all h_K, if we need one), we get the Odds Form of Bayes' Theorem. The *odds against A* given B is defined as $\Omega_\alpha[{\sim}A \mid B] = P_\alpha[{\sim}A \mid B] / P_\alpha[A \mid B]$. So, we have:

Bayes' Theorem: The Odds Form

$$\begin{aligned}
(10)\quad \Omega_\alpha[{\sim}h_i \mid b\cdot c^n\cdot e^n] &= \sum_{j\neq i} \frac{P_\alpha[h_j \mid b\cdot c^n\cdot e^n]}{P_\alpha[h_i \mid b\cdot c^n\cdot e^n]} + \frac{P_\alpha[h_K \mid b\cdot c^n\cdot e^n]}{P_\alpha[h_i \mid b\cdot c^n\cdot e^n]} \\
&= \sum_{j\neq i} \frac{P[e^n \mid h_j\cdot b\cdot c^n]}{P[e^n \mid h_i\cdot b\cdot c^n]} \times \frac{P_\alpha[h_j \mid b]}{P_\alpha[h_i \mid b]} \\
&\quad + \frac{P_\alpha[e^n \mid h_K\cdot b\cdot c^n]}{P[e^n \mid h_i\cdot b\cdot c^n]} \times \frac{P_\alpha[h_K \mid b]}{P_\alpha[h_i \mid b]}
\end{aligned}$$

where the factor following the '+' sign is only required in cases where a catch-all alternative hypothesis, h_K, is needed.

Notice that if a catch-all hypothesis is needed, the likelihood of evidence relative to it will not generally enjoy the same kind of objectivity as the likelihoods for *specific,* *positive* hypotheses. We leave the subscript α on the likelihood for the catch-all to indicate this lack of objectivity.

Although the catch-all hypothesis may lack objective likelihoods, the influence of the catch-all term in Bayes' theorem diminishes as additional *positive* hypotheses are articulated. That is, as new hypotheses are discovered they are "peeled off" of the catch-all. So, when a new hypothesis h_{u+1} is formulated and made explicit, the old catch-all h_K is replaced by a new catch-all, h_{K^*}, of form $(\sim h_1 \cdot \ldots \cdot \sim h_u \cdot \sim h_{u+1})$; and the prior probability for the new catch-all hypothesis is gotten by diminishing the prior of the old catch-all: $P_\alpha[h_{K^*} \mid b] = P_\alpha[h_K \mid b] - P_\alpha[h_{u+1} \mid b]$. Thus, the influence of the catch-all term should diminish towards 0 as new alternative hypotheses are made explicit.[10]

If increasing evidence drives the likelihood ratios comparing h_i with each competitor towards 0, then the odds against h_i, $\Omega_\alpha[\sim h_i \mid b \cdot c^n \cdot e^n]$, will approach 0 (provided that priors of catch-all terms, if needed, approach 0 as well as new alternative hypotheses are made explicit and peeled off). And, as $\Omega_\alpha[\sim h_i \mid b \cdot c^n \cdot e^n]$ approaches 0, the posterior probability of h_i goes to 1. The relationship between the odds against h_i and its posterior probability is this:

Bayes' Theorem: General Probabilistic Form

(11) $P_\alpha[h_i \mid b \cdot c^n \cdot e^n] = 1 / (1 + \Omega_\alpha[\sim h_i \mid b \cdot c^n \cdot e^n])$.

The odds against a hypothesis depends only on the values of *ratios of posterior probabilities*, which entirely derive from the Ratio Form of Bayes' Theorem. Thus we see that the individual value of the posterior probability of a hypothesis depends only on the *ratios of posterior probabilities*, which come from the Ratio Form of Bayes' Theorem. Thus, the Ratio Form of Bayes' Theorem completely captures the essential features of the Bayesian evaluation of hypothesis. It shows how the impact of evidence (in the form of likelihood ratios) combines with comparative plausibility assessments of hypotheses (in the form of ratios of prior probabilities) to provide a net assessment of the extent to which hypotheses are refuted or supported via contests with their rivals.

There is a result, a kind of *Bayesian Convergence Theorem*, that shows that if h_i (together with $b \cdot c^n$) is true, then the likelihood ratios $P[e^n \mid h_j \cdot b \cdot c^n] / P[e^n \mid h_i \cdot b \cdot c^n]$ comparing evidentially distinguishable alternative hypothesis h_j to h_i will *very*

probably approach 0 as evidence accumulates (i.e., as n increases). Let's call this result the *Likelihood Ratio Convergence Theorem*. When this theorem applies, Equation 9 shows that the posterior probability of false competitor h_j will very probably approach 0 as evidence accumulates, regardless of the value of its prior probability $P_\alpha[h_j \mid b]$. As this happens to each of h_i's false competitors, Equations 10 and 11 say that the posterior probability of the true hypothesis, h_i, will approach 1 as evidence increases.[11] Thus, Bayesian induction is at bottom a version of *induction by elimination*, where the elimination of alternatives comes by way of likelihood ratios approaching 0 as evidence accumulates. We will examine the *Likelihood Ratio Convergence Theorem* in detail in Section 5.[12]

For more on Bayes' Theorem see the entries on Bayes' Theorem and on Bayesian epistemology in this *Encyclopedia*.

3.4 Likelihood Ratios, Likelihoodism, and the Law of Likelihood

The versions of Bayes' Theorem provided by Equations 9-11 show that for probabilistic inductive logic the influence on posterior probabilities of hypotheses of the kind of empirical evidence for which hypotheses express likelihoods is completely captured by the ratios of likelihoods, $P[e^n \mid h_j \cdot b \cdot c^n] / P[e^n \mid h_i \cdot b \cdot c^n]$. The evidence $(c^n \cdot e^n)$ influences the posterior probabilities in no other way. So, the following "Law" is a consequence of the logic of probabilistic support functions.

> **General Law of Likelihood**:
> Given any pair of incompatible hypotheses h_i and h_j, whenever the likelihoods $P_\alpha[e^n \mid h_j \cdot b \cdot c^n]$ and $P_\alpha[e^n \mid h_i \cdot b \cdot c^n]$ are defined, the evidence $(c^n \cdot e^n)$ supports h_i over h_j, given b, *if and only if* $P_\alpha[e^n \mid h_i \cdot b \cdot c^n] > P_\alpha[e^n \mid h_j \cdot b \cdot c^n]$. The ratio of likelihoods $P_\alpha[e^n \mid h_i \cdot b \cdot c^n] / P_\alpha[e^n \mid h_j \cdot b \cdot c^n]$ measures the *strength of the evidence* for h_i over h_j given b.

Two features of this law require some explanation. As stated, the **General Law of Likelihood** does not presuppose that likelihoods of form $P_\alpha[e^n \mid h_j \cdot b \cdot c^n]$ and $P_\alpha[e^n \mid h_i \cdot b \cdot c^n]$ are always *defined*. This qualification is introduced to accommodate a conception of evidential support called *Likelihoodism*, which is especially influential among statisticians. Also, the likelihoods in the law are expressed with the subscript α attached to indicate that the law holds for each *inductive support function* P_α, even when the values of the likelihoods are not objective or agreed on by all agents in a given scientific community. These two features of the law are closely related, as we will see.

Each *probabilistic support function* satisfies the axioms of Section 2. According to these axioms the conditional probability of one sentence on another is always defined. So, in the context of the *inductive logic of support functions* the likelihoods are always defined, and the qualifying clause about this in the **General Law of Likelihood** is automatically satisfied. For *inductive support functions*, all of the versions of Bayes' theorem (Equations 8-11) continue to hold even when the likelihoods are not objective or intersubjectively agreed on by the scientific community. In many scientific contexts there will be general agreement on the values of likelihoods; but whenever such agreement fails the subscripts α, β, etc. must remain attached to the support function likelihoods to indicate this. Even so, the **General Law of Likelihood** continues to hold for each support function.

There is a view, or family of views, called *likelihoodism* that maintains that the inductive logician or statistician should only be concerned with whether the evidence provides *increased* or *decreased support* for one hypothesis over another, and only in cases where this evaluation is based on the ratios of *completely objective* likelihoods. (Prominent likelihoodists include Edwards (1972) and Royall (1997); also see Forster and Sober (2004) and Fitelson (2007).) When the likelihoods involved are objective, the ratios $P[e^n \mid h_j\cdot b\cdot c^n] / P[e^n \mid h_i\cdot b\cdot c^n]$ provide a *pure, objective measure* of how strongly the evidence supports h_i as compared to h_j, a measure that is "untainted" by prior plausibility considerations. According to likelihoodists, only this kind of *pure measure* is scientifically appropriate for the assessment of how evidence impacts hypotheses. (It should be noted that the *classical school of statistics*, assocoated with R.A. Fisher (1922) and with Neyman and Pearson (1967), reject the claim about the nature of evidential support expressed by the **General Law of Likelihood**.)

Likelihoodists maintain that it is not appropriate for statisticians to incorporate assumptions about prior probabilities of hypotheses into the assessment of evidential support. It is not their place to compute *recommended values* of posterior probabilities for the scientific community. When the results of experiments are made public, say in scientific journals, only objective likelihoods should be reported. The evaluation of the impact of objective likelihoods on agents' posterior probabilities depends on each

agent's individual *subjective* prior probability, which represents plausibility considerations that have nothing to do with the evidence. So, likelihoodists suggest, posterior probabilities should be left for individuals to compute (if they desire to do so).

The conditional probabilities between most pairs of sentences fail to be objectively defined in a way that suits likelihoodists. So, for likelihoodists, the general *logic of support functions* (captured by the axioms of Section 2) cannot represent an *objectivelogic* of evidential support for hypotheses. Because they eschew the logic of support functions, likelihoodist do not have Bayes' theorem available, and so cannot derive the **Law of Likelihood** from it. Rather, they must state the **Law of Likelihood** as an *axiom* of their inductive logic, an axiom that applies only when the likelihoods have *well-defined* objective values.

Likelihoodists tend to have a very strict conception of what it takes for likelihoods to be *well-defined*. They consider a likelihood to be *well-defined* only when it is (what we referred to earlier as) a *direct inference likelihood* — i.e., only when either, (1) the hypothesis (together with background and experimental conditions) logically entails the data, or (2) the hypothesis (together with background) logically entails an explicit *simple statistical hypothesis* that (together with experimental conditions) specifies precise probabilities for each of the events that make up the evidence.

Likelihoodists contrast *simple statistical hypotheses* with *composite statistical hypotheses*, which only entail vague, or imprecise, or *directional* claims about the statistical probabilities of evidential events. Whereas a *simple statistical hypothesis* might say, for example, "the chance of *heads* on tosses of the coin is precisely .65", a composite statistical hypothesis might say, "the chance of *heads* on tosses is either .65 or .75," or it may be a *directional hypothesis* that says, "the chance of *heads* on tosses is greater than .65." *Likelihoodists* maintain that *composite hypotheses* are not an appropriate basis for well-defined likelihoods. Such hypotheses represent a kind of disjunction of simple statistical hypotheses. The *direction hypothesis*, for instance, is essentially a disjunction of the various *simple statistical hypotheses* that assign specific values above .65 to the chances of heads on tosses. Likelihoods based on such hypotheses are not *appropriately objective* by the lights of the *likelihoodist* because they must in effect depend on factors that represent the degree to which the *composite hypothesis* supports each of the *simple statistical hypotheses* that it encompasses; and *likelihoodists* consider such factors *too subjective* to be permitted in a logic that should permit only objective likelihoods.[13]

Taking all of this into account, the version of the **Law of Likelihood** appropriate to *likelihoodists* may be stated as follows.

> **Special Law of Likelihood**:
> Given a pair of incompatible hypotheses h_i and h_j that imply simple statistical models regarding outcomes e^n given $(b \cdot c^n)$, the likelihoods $P[e^n \mid h_j \cdot b \cdot c^n]$ and $P[e^n \mid h_i \cdot b \cdot c^n]$ are well defined. For such likelihoods, the evidence $(c^n \cdot e^n)$ supports h_i over h_j, given b, *if and only if* $P[e^n \mid h_i \cdot b \cdot c^n] > P[e^n \mid h_j \cdot b \cdot c^n]$; the ratio of likelihoods $P[e^n \mid h_i \cdot b \cdot c^n] / P[e^n \mid h_j \cdot b \cdot c^n]$ measures the *strength of the evidence* for h_i over h_j given b.

Notice that when either version of the **Law of Likelihood** holds, the absolute size of a likelihood is irrelevant to the strength of the evidence. All that matters is the relative size of the likelihoods for one hypothesis as compared to another. That is, let c_1 and c_2 be the conditions for two distinct experiments having outcomes e_1 and e_2, respectively. Suppose that e_1 is 1000 times more likely on h_i (given $b \cdot c_1$) than is e_2 on h_i (given $b \cdot c_2$); and suppose that e_1 is also 1000 times more likely on h_j (given $b \cdot c_1$) than is e_2 on h_j (given $b \cdot c_2$) — i.e., suppose that $P_\alpha[e_1 \mid h_i \cdot b \cdot c_1] = 1000 \times P_\alpha[e_2 \mid h_i \cdot b \cdot c_1]$, and $P_\alpha[e_1 \mid h_j \cdot b \cdot c_1] = 1000 \times P_\alpha[e_2 \mid h_j \cdot b \cdot c_2]$. Which piece of evidence, $(c_1 \cdot e_1)$ or $(c_2 \cdot e_2)$, is stronger evidence with regard to the comparison of h_i to h_j? The **Law of Likelihood** implies both are equally strong. All that matters evidentially are the ratios of the likelihoods, and they are the same: $P_\alpha[e_1 \mid h_i \cdot b \cdot c_1] / P_\alpha[e_1 \mid h_j \cdot b \cdot c_1] = P_\alpha[e_2 \mid h_i \cdot b \cdot c_2] / P_\alpha[e_2 \mid h_j \cdot b \cdot c_2]$. Thus, the **General Law of Likelihood** implies the following principle.

> **General Likelihood Principle**:
> Suppose two different experiments or observations (or two sequences of them) c_1 and c_2 produce outcomes e_1 and e_2, respectively. Let $\{ h_1, h_2, \ldots\}$ be any set of alternative hypotheses. If there is a constant K such that for each hypothesis h_j from the set, $P_\alpha[e_1 \mid h_j \cdot b \cdot c_1] = K \times P_\alpha[e_2 \mid h_j \cdot b \cdot c_2]$, then the *evidential import* of $(c_1 \cdot e_1)$ for distinguishing among hypotheses in the set (given b) is precisely the same as the *evidential import* of $(c_2 \cdot e_2)$.

Similarly, the **Special Law of Likelihood** implies a corresponding **Special Likelihood Principle** that applies only to hypotheses that express simple statistical models.[14]

Throughout the remainder of this article we will not assume that likelihoods must be based on simple statistical hypotheses, as *likelihoodist* would have them. However, most of what will be said about likelihoods, especially the convergence result in Section 5, applies to *likelihoodist* likelihoods as well. We will, however, continue to suppose that likelihoods are *objective* in the sense that all members of the scientific community agree on their numerical values. In Section 6 we will see how even this supposition may be relaxed in scientific contexts where completely objective values for likelihoods are not realistically available.

3.5 On Prior Probability Assessments — and Representations of Vague and Diverse Plausibility Assessments

Given that a scientific community should largely agree on the values of the likelihoods, any significant disagreement among them with regard to the values of posterior probabilities of hypotheses should derive from disagreements over their assessments of values for the prior probabilities of those hypotheses. We saw in section 3.3 that the Bayesian logic of evidential support need only rely on assessments of *ratios of prior probabilities* — on how much more plausible one hypothesis is than another. Furthermore, presumably, in scientific contexts the comparative plausibility values for hypotheses should depend on explicit plausibility arguments (stated within b), not on privately help opinions. (It would be highly *unscientific* for a member of the scientific community to disregard or dismiss a hypothesis that other members take to be a reasonable proposal with only the comment: “don't ask me to give reasons, it's just my opinion”.) Even so, agents may be unable to specify *precisely* how much more strongly the available plausibility arguments support a hypothesis over an alternative; so prior probability ratios for hypotheses may be vague. Furthermore, agents in a scientific community may disagree about how strongly the available plausibility arguments support a hypothesis over a rival hypothesis; so prior probability ratios may be somewhat diverse as well.

Both the vagueness of prior plausibility ratio values for individual agents and the diversity of values among the community of agents can be represented formally by sets of probabilistic support functions, $\{P_\alpha, P_\beta, \ldots\}$, that agree on the values for the likelihoods but encompass a range of values for the (ratios of) prior probabilities of hypotheses. *Vagueness* and *diversity* are somewhat different issues, but they may be represented in much the same way. Let's consider each in turn.

Assessments of the prior plausibilities of hypotheses will often be vague — not subject to the kind of precise quantitative treatment that a Bayesian version of probabilistic inductive logic may seem to require for prior probabilities. So, it is sometimes objected, the kind of assessment of prior probabilities required to get the Bayesian algorithm going cannot be accomplished in practice. To see how Bayesian inductivists address this worry, first recall the Ratio Form of Bayes' Theorem, equation (9).

$$\frac{P_\alpha[h_j \mid b\cdot c^n\cdot e^n]}{P_\alpha[h_i \mid b\cdot c^n\cdot e^n]} = \frac{P[e^n \mid h_j\cdot b\cdot c^n]}{P[e^n \mid h_i\cdot b\cdot c^n]} \times \frac{P_\alpha[h_j \mid b]}{P_\alpha[h_i \mid b]}$$

Recall that this Ratio Form of the theorem captures the essential features of the logic of evidential support, even though it only provides a value for the ratio of the posterior probabilities. (We'll see why this is so in more detail in a moment.)

Notice that the ratio form of the theorem easily accommodates situations where we don't have precise numerical values for prior probabilities. It only depends on our ability to assess *how much more or less plausible* alternative hypothesis h_j is than hypothesis h_i — only the value of the ratio $P_\alpha[h_j \mid b] / P_\alpha[h_i \mid b]$ need be assessed; the values of the individual prior probabilities are not required. Such comparative plausibilities are much easier to assess than specific numerical prior plausibility values for individual hypotheses. When combined with the *ratio of likelihoods*, this *ratio of priors* suffices to yield an assessment of the *ratio of posterior plausibilities*, $P_\alpha[h_j \mid b\cdot c^n\cdot e^n] / P_\alpha[h_i \mid b\cdot c^n\cdot e^n]$.

Although such posterior ratios don't supply values for the individual posterior probabilities, they place a crucial constraint on the posterior support of hypothesis h_j, since

$$P_\alpha[h_j \mid b\cdot c^n\cdot e^n] < \frac{P_\alpha[h_j \mid b\cdot c^n\cdot e^n]}{P_\alpha[h_i \mid b\cdot c^n\cdot e^n]} = \frac{P[e^n \mid h_j\cdot b\cdot c^n]}{P[e^n \mid h_i\cdot b\cdot c^n]} \times \frac{P_\alpha[h_j \mid b]}{P_\alpha[h_i \mid b]}$$

This Ratio Form of Bayes' Theorem tolerates a good deal of vagueness or imprecision in assessments of the ratios of prior probabilities. In practice one need only assess bounds for these prior plausibility ratios to achieve meaningful results. Given a prior ratio in a specific interval,

$$q \leq P_\alpha[h_j \mid b] / P_\alpha[h_i \mid b] \leq r$$

a likelihood ratio $P[e^n \mid h_j\cdot b\cdot c^n] / P[e^n \mid h_i\cdot b\cdot c^n] = LR^n$ results in a posterior confirmation ratio in the interval

$$(LR^n\times q) \leq P_\alpha[h_j \mid b\cdot c^n\cdot e^n] / P_\alpha[h_i \mid b\cdot c^n\cdot e^n] \leq (LR^n\times r).$$

> Technically each probabilistic support function assigns a specific numerical value to each pair of sentences; so when we write an inequality like '$q \leq P_\alpha[h_j \mid b] / P_\alpha[h_i \mid b] \leq r$'we are really referring to a set of probability function P_α, a *vaguness set*, for which the inequality holds. Thus, technically, the Bayesian logic employs sets of probabilistic support functions to represent the vagueness in comparative plausibility values for hypotheses.

Observe that if the likelihood ratio values LR^n approach 0 as the amount of evidence e^n increases, the interval of values for the posterior probability ratio becomes tighter as the upper bound $(LR^n\times r)$ approaches 0. Furthermore, the absolute degree of support for h_j, $P_\alpha[h_j \mid b\cdot c^n\cdot e^n]$, must also approach 0.

This observation is really useful because it can be shown that when $h_i\cdot b\cdot c^n$ is true and h_j is empirically distinct from h_i, the continual pursuit of evidence is *very likely* to result in evidential outcomes e^n that (as n increases) yield values of likelihood ratios $P[e^n \mid h_j\cdot b\cdot c^n] / P[e^n \mid h_i\cdot b\cdot c^n] = LR^n$ that approach 0 as the amount of evidence increases. (I'll provide the details of this *Likelihood Ratio Convergence Theorem* in section 5.) When that happens, the upper bound on the posterior probability ratio also approaches 0, driving the posterior probability of h_j to approach 0, effectively refuting hypothesis h_j. Thus, false competitors of a true hypothesis will effectively be eliminated by increasing evidence. As this happens, equations (10) and (11) show that the posterior probability $P_\alpha[h_i \mid b\cdot c^n\cdot e^n]$ of the true hypothesis h_i approaches 1.

Thus, Bayesian inductive support for hypotheses is a form of eliminative induction, where the evidence effectively refutes false alternatives to the true hypothesis. The eliminative nature of Bayesian evidential support doesn't require precise values for prior probabilities. It only need draw on bounds on comparative plausibility ratios, and these bounds only play a significant role while evidence remains fairly sparse. If the true hypothesis is comparatively plausible (due to plausibility arguments contained in b), then plausibility assessments give it a leg-up over alternatives. If the true hypothesis is comparatively implausible, the plausibility assessments merely slow down the rate at which it comes to dominate its rivals, reflecting the idea that *extraordinary hypotheses require extraordinary evidence* (or an extraordinary accumulation of evidence) to overcome their initial implausibilities.

Thus, as evidence accumulates, the agent's vague initial plausibility assessments transform into quite sharp posterior probabilities that indicate the strong refutation or support of the various hypotheses. Intuitively this seems quite a reasonable way to represent how evidential support should work.

The various agents in a community may widely disagree over the non-evidential plausiblities of hypotheses. Bayesian inductivists may represent this kind of *diversity* across the community of agents as a collection of the agents' *vagueness sets*. Let's call such a collection a *diversity set*. That is, a *diversity set* is just a set of support functions P_α that cover the ranges of values for comparative plausibility assessments for pairs of competing hypotheses

$$q \le P_\alpha[h_j \mid b] \,/\, P_\alpha[h_i \mid b] \le r$$

as assessed by the scientific community on the basis of plausibility arguments and considerations (expressed within b).

So, although there may well be disagreement among agents regarding the ranges of comparative prior plausibilities of hypotheses, a probabilistic inductive logic may easily represent this diversity. Furthermore, if accumulating evidence drives the likelihood ratios to extremes, the range of functions in a *diversity set* will come to near agreement, near 0 or 1, on the values for posterior probabilities of hypotheses. So, not only does such evidence *firm up* each agent's vague initial plausibilities, it also brings the whole community into agreement on the *near refutation* of empirically distinct competitors of a true hypothesis.

Under what conditions might the likelihood ratios go to such extremes as evidence accumulates, effectively washing out vagueness and diversity? The *Likelihood Ratio Convergence Theorem* (discussed in detail in Section 5) implies that if a true hypothesis disagrees with false alternatives on the likelihoods of *possible outcomes* for a long enough stream of experiments or observations, then that evidence stream will very probably produce *actual outcomes* that drive the likelikood ratios of false alternatives as compared to the true hypothesis to approach 0. As this happens, almost any range of prior plausibility assessments will be driven to agreement on the posterior plausibilities for hypotheses. Thus, the accumulating evidence will very probably bring all support functions in the *vagueness* and *diversity sets* for a community of agents to near agreement on posterior plausibility values — near 0 for the false competitors, and near 1 for the true hypothesis (or for its disjunction with empirically indistinguishable alternatives).

One more point about prior probabilities and Bayesian convergence should be mentioned here. Some subjectivist versions of Bayesian induction seem to suggest that an agent's prior plausibility assessments for hypotheses should stay fixed once and for all, and that all plausibility updating should be brought about via the likelihoods in accord with Bayes' Theorem. Critics argue that this is unreasonable. The members of a scientific community may quite legitimately revise their (comparative) prior plausibility assessments for hypotheses from time to time as they rethink plausibility arguments and bring new considerations to bear. This seems a natural part of the conceptual development of a science. It turns out that such reassessments of priors poses no difficulty for probabilistic inductive logic as I've described it here. Reassessments may come about by the addition of explicit statements that supplement or modify the background information b, and they may also take the form of (non-Bayesian) transitions to new *vagueness sets* for individual agents and to new *diversity sets* for the community. The *logic* of Bayesian induction (as described here) has nothing to say about what values the prior plausibility assessments for hypotheses should have; and it places no restrictions on how they might change. Provided that the series of reassessments of prior plausibilities doesn't push the prior of the true hypothesis ever nearer to zero, the *Likelihood Ratio Convergence Theorem* implies that the evidence will very probably bring the posterior probabilities of empirically distinct rivals of the true hypothesis to approach 0 via decreasing likelihood ratios; and as this happens, the posterior probability of the true hypothesis will head towards 1.

4. Bayesian Estimation and Convergence for Enumerative Inductions

In this section we'll see that for the special case of enumerative inductions probabilistic inductive logic satisfies the Criterion of Adequacy (CoA) stated at the beginning of this article. That is, under some plausible conditions, given a reasonable amount of evidence, the *degree* to which that evidence comes to *support* a hypothesis through enumerative induction is very likely to approach 1 for true hypotheses. We will now see precisely how this works.

Recall that in enumerative inductions the idea is to infer the proportion, or *relative frequency*, of an attribute in a population from how frequently the attribute occurs in a sample of the population. Examples 1 and 2 at the beginning of the article describe two such inferences. Enumerative induction is only a rather special case of inductive inference. However, such inferences are very common, and so worthy of careful attention. They arise, for example, in the context of polling and in many other cases where a population frequency is estimated from a sample. We will establish conditions under which such inferences give rise to highly objective posterior probabilities, posterior probabilities that are extremely stable over a wide range of reasonable prior plausibility assessments. That is, we will consider all of the inductive support functions in an agent's *vagueness set V* or in a community's *diversity set D*. We will see that under some very weak suppositions about the make up of V or of D, a reasonable amount of data will bring all of the support functions in these sets to agree that the posterior degree of support for a particular frequency hypothesis is very close to 1. And, we will see, it is very likely these support functions will converge to agreement on a true hypothesis.

4.1 Convergence to Agreement

Suppose we want to know the frequency with which attribute A occurs among members of population B. We randomly select a sample S from B consisting of n members, and find that it contains m members that exhibit attribute A.[15] On the basis of this evidence, what is the posterior probability p of the hypothesis that the true proportion (or frequency) of As among Bs is within a given region R around the sample proportion m/n? And to what extent does that bound depend on the prior probabilities of the various possible alternative frequency hypotheses. More generally, for a given *vagueness* or *diversity set*, what bounds can we place on the values of p.

Put more formally, we are asking for what values of p and q does the following inequality hold:

$$P_\alpha[(m/n)-q < F[A,B] < (m/n)+q \mid b \cdot F[A,S]=m/n \cdot \mathrm{Rnd}[S,B,A] \cdot \mathrm{Size}[S]=n] > p?$$

It turns out that we need only a very weak supposition about the values of prior probabilities of support functions P_α in *vagueness* or *diversity sets* to legitimize such inferences, an supposition that almost always holds in the context of enumerative inductions.

> **Boundedness Assumption for Estimation:**
> There is a region R of possible values near the sample frequency m/n (e.g., R is the region between $(m/n)-q$ and $(m/n)+q$, for some *margin of error q* of interest) such that no frequency hypothesis outside of region R is overwhelmingly more initially plausible than those frequency hypotheses inside of region R.

What does it mean for *no frequency hypothesis outside of region R to be overwhelmingly more initially plausible than those frequency hypotheses inside of region R* (where R is some specific region in which the sample frequence, $F[A,S]=m/n$, lies)? The main idea is that there is some (perhaps very large) bound K on how much more plausible frequency hypotheses outside of region R may be than those frequency hypotheses inside of region R. We state this condition carefully by considering two kinds of cases, depending on whether or not the population B is known to be bounded in size by some specific (perhaps overly large) integer u. (The first case will be simpler because it doesn't suppose that the support functions involved may be characterized by probability density functions, while the second case does suppose this.)

> Case 1. Suppose the size of the population B is finite. We need not know how big B is. We merely suppose that for some positive integer u that is at least as large as the size of B, but might well be many times larger, the following condition holds for all support functions P_α in the *vagueness* or *diversity set* under consideration.
>
> There is some specific positive factor K (possibly very large, perhaps as large as 1000, or larger) such that for any pair of hypotheses of form $F[A,B] = v/u$ inside region R and of form $F[A,B] = w/u$ outside of region R (where u, v, and w are non-negative integers), the hypothesis outside of region R is no more than K times more plausible than the hypothesis within region R (given plausibility consideration within b) — i.e., for all ratios v/u inside region R and all ratios v/u outside region R, $P_\alpha[F[A,B]{=}w/u \mid b] / P_\alpha[F[A,B]{=}v/u \mid b] \le K$.

For Case 1 we also assume (as seems reasonable) that in the absence of information about the observed sample frequency, the claim ‘Random$[S,B,A]$ · Size$[S]{=}n$’, that the sample is randomly selected and of size n, should be irrelevant to the initial plausibilities of possible population frequencies — i.e. we suppose that $P_\alpha[F[A,B]{=}k/u \mid \text{Rnd}[S,B,A] \cdot \text{Size}[S]{=}n \cdot b] = P_\alpha[F[A,B]{=}k/u \mid b]$ for each integer k from 0 through u.

> Case 2. Alternatively, suppose that there is no positive integer u at least as large as the size of population B that satisfies the conditions of case 1. But suppose that the prior probabilities of the various competing hypotheses can be represented (at least very nearly) by a probability density function $p_\alpha[F[A,B]{=}r \mid b]$— i.e., for any specific values v and u, the value of $P_\alpha[\, v < F[A,B] < u \mid b] = \int_v^u p_\alpha[F[A,B]{=}r \mid b]\, dr$, or at least very nearly so. Then we just need the following condition to be satisfied by all support functions P_α in the *vagueness* or *diversity set* under consideration.
>
> There is some specific positive factor K (possibly very large, perhaps as large as 1000, or larger) such that for any pair of hypotheses of form $F[A,B] = r$ inside region R and of form $F[A,B] = s$ outside of region R (where r and s are non-negative real numbers no larger than 1), the value of the probability density function for the hypothesis outside of region R is no more than K times larger than the value of the probability density function for the hypothesis within region R (given plausibility consideration within b), where the density function within region R is never less than some (perhaps very tiny) positive lower bound — i.e., for all values r inside region R and all values s outside region R, $p_\alpha[F[A,B]{=}s \mid b] / p_\alpha[F[A,B]{=}r \mid b] \le K$, where for all r within region R, $p_\alpha[F[A,B]{=}r \mid b] \ge g$ for some small $g > 0$.

For Case 2 we also assume (as seems reasonable) that in the absence of information about the observed sample frequency, the claim ‘Random$[S,B,A]$ · Size$[S]{=}n$’, that the sample is randomly selected and of size n, should be irrelevant to the initial plausibilities of possible population frequencies — i.e. in particular, we suppose that for each probability density function p_α under consideration, $p_\alpha[F[A,B]{=}q \mid \text{Rnd}[S,B,A] \cdot \text{Size}[S]{=}n \cdot b] = p_\alpha[F[A,B]{=}q \mid b]$ for real numbers q from 0 through 1.

When either of these two Cases hold, let us say that for the support functions P_α in the *vagueness* or *diversity sets* under consideration, the prior probabilities are *K bounded with resprct to region R*. Then we have the following theorem about enumerative inductions, which shows that the posterior probability that the true frequency must lie within a small region R around the sample frequency m/n quicky approaches 1 as the sample size n becomes large.

Theorem: Frequency Estimation Theorem:[16]
Suppose, for all support functions P_α in the *vagueness* or *diversity* set under consideration, the prior probabilities are *K bounded with respect to region R*, where region R contains the fraction m/n (for positive integer n and non-negitive integer $m \le n$). Then, for all support functions P_α in the *vagueness* or *diversity set*,

$$P_\alpha[\, F[A,B]\in R \mid b \cdot F[A,S]{=}m/n \cdot \text{Rnd}[S,B,A] \cdot \text{Size}[S]{=}n]$$

$$\ge\ 1 / (1 + K\times[(1/\beta[R, m{+}1, n{-}m{+}1]) - 1]).$$

> For any given region R containing sample frequencies m/n, this lower bound approaches 1 rapidly as n increases.

The expression '$\beta[R, m+1, n-m+1]$' represents the beta-distribution function with parameters $m+1$ and $n-m+1$ evaluated over region R. By definition $\beta[R, m+1, n-m+1] = \int_R r^m (1-r)^{n-m}\, dr / \int_0^1 r^m (1-r)^{n-m} dr$. When region R contains an interval around m/n, the value of this function is a fraction that approaches 1 for large n. In a moment we will see some concrete illustrations of the implications of this theorem for specific values of m and n and specific regions R.

The values of the beta-distribution function may be easily computed using a version of the function supplied with most mathematics and spreadsheet programs. The version of the function supplied by such programs usually takes the form BETADIST(x, y, z), which computes the value of the beta distribution from 0 up to to x, and where y and z are the "parameters of the distribution". For our purposes, where the sample S of n selections from B yields m that exhibit As, these parameters need to be $m+1$ and $n-m+1$. So if the region R of interest (wherein the sample frequency m/n lies) is between the values v and u (where v is the lower bound on region R and u is the upper bound on region R), then the program should be asked to compute the value of $\beta[R, m+1, n-m+1] = \int_v^u r^m (1-r)^{n-m}\, dr / \int_0^1 r^m (1-r)^{n-m} dr$ by having it compute BETADIST[u, m+1, n−m+1]− BETADIST[v, m+1, n−m+1]. So, to have your mathematics or spreadsheet program compute a lower bound on the value of

$$P_\alpha[v \leq F[A,B] \leq u \mid b \cdot F[A,S]=m/n \cdot \mathrm{Rnd}[S,B,A] \cdot \mathrm{Size}[S]=n]$$

for a given upper bound K (on how much more initially plausible it is that the true population frequency lies *outside* the region between v and u than it is that the true polulation frequency lies inside that region), you may be able to simply paste the following expression into your program and then plug in desired values for K, u, v, m, n in this expression:

```
1 / (1 + K*((1/(BETADIST(u, m+1, n-m+1)-BETADIST(v, m+1, n-m+1)) - 1))
```

In many real cases it will *not be initially more plausible* that the true frequency value lies outside of the *region of interest* between v and u than that it lies inside that region. In such cases set the value of K to 1. However, you will find that for any moderately large sample size n, this function yields very similar values for all plausible values of K you might try out, even when the values of K are quite large. (We'll see examples of this fact in the computed tables below.)

This theorem implies that for large samples the values of prior probabilities don't matter much. Given such evidence, a vary wide range of inductive support functions P_α will come to agree on high posterior probabilities that the proportion of attribute A in population B is very close to the sample frequency. Thus, all support functions in such *vagueness* or *diversity* sets come to near agreement. Let us look at several numerical examples to make clear how strong this result really is.

The first section of this article provided two examples of enumerative inductive inferences. Consider Example 1. Let 'B' represent the population of all ravens. Let 'A' represent the class of black ravens. Now consider those hypotheses of form '$F[A,B] = r$' for r in the interval between .99 and 1. This collection of hypotheses includes the claim that "all ravens are black" together with those alternative hypotheses that claim the frequency of being black among ravens is within .01 of 1. The alternatives to these hypotheses are just those that assert '$F[A,B] = s$' for values of s below .99.

Suppose none of the support functions represented in the *vagueness* or *diversity* set under consideration rates the prior plausibility of any of the hypotheses '$F[A,B] = s$' with s less than .99 to be *more than twice as plausible* as the hypotheses '$F[A,B] = r$' for which r is between .99 and 1. That is, suppose, for each P_α in the *vagueness* or *diversity* set under consideration, the prior plausibility $P_\alpha[F[A,B] = s \mid b]$ for hypotheses with s below .99 is never more than $K = 2$ times greater than the prior plausibility $P_\alpha[F[A,B] = r \mid b]$ for hypotheses with r between .99 and 1. Then, on the evidence of 400 ravens selected randomly with respect to color, the theorem yields the following bound for all P_α in the *vagueness* or *diversity* set:

$$P_\alpha[F[A,B] > .99 \mid b \cdot F[A,S] = 1 \cdot \mathrm{Rnd}[S,B,A] \cdot \mathrm{Size}[S] = 400] \;\geq\; .9651.$$

The following table describes similar results for other upper bounds K on values of prior probability ratios and other sample sizes n:

Table 1: Values of lower bound p on the posterior probability

$m/n = 1$ $F[A,B] > .99$	Sample-Size = n (number of As in Sample of Bs = $m = n$)			
Prior Ratio K ↓	400	800	1600	3200
1	0.9822	0.9997	1.0000	1.0000
2	0.9651	0.9994	1.0000	1.0000
5	0.9170	0.9984	1.0000	1.0000
10	0.8468	0.9968	1.0000	1.0000
100	0.3560	0.9691	1.0000	1.0000
1,000	0.0524	0.7581	0.9999	1.0000
10,000	0.0055	0.2386	0.9990	1.0000
100,000	0.0006	0.0304	0.9898	1.0000
1,000,000	0.0001	0.0031	0.9068	1.0000
10,000,000	0.0000	0.0003	0.4931	1.0000

$P_\alpha[F[A,B]>.99 \mid b \cdot F[A,S]=1 \cdot \mathrm{Rnd}[S,B,A] \cdot \mathrm{Size}[S]=n] \geq p$, for a range of Sample-Sizes n (from 400 to 3200), when the prior probability of any specific frequency hypothesis outside the region between .99 and 1 is no more than K times more than the lowest prior probability for any specific frequency hypothesis inside of the region between .99 and 1.

(All probabilities with entries ‘1.0000’ in this table and the next actually have values slightly less than one, but nearly equal 1.0000 to four significant decimal places.)

To see what the table tells us, consider the third to last row. It represents what happens when a *vagueness* or *diversity* set contains at least some support functions that assign prior probabilities (i.e. prior plausibilities) nearly *one hundred thousand* times higher to some hypotheses asserting frequencies *not* between .99 and 1 than it assigns to hypotheses asserting frequencies between .99 and 1. The table shows that even in such cases, a random sample of 1600 black ravens will, nevertheless, pull the posterior plausibility level that “the true frequency is above .99” to a value above .9898, for every support function in the set. And if the *vagueness* or *diversity* set contains support functions that assign even more extreme priors, say, priors that are nearly *ten million* times higher for some hypotheses asserting frequencies below .99 than for hypotheses within .99 of 1 (the table's last row), this poses no great problem for convergence-to-agreement. A random sample of 3200 black ravens will yield posterior probabilities

(i.e. posterior plausibilities) indistinguishable from 1 for the claim that “more than 99% of all ravens are black.”

Strong support can be gotten for an even narrower range of hypotheses about the percentage of black birds among the ravens. But a larger sample size is needed for this. For an additional example, see the supplementary document

Tighter Bounds on the Margin of Error.

Now consider the second example of an enumerative induction provided at the beginning of this article, involving the poll about the presidential preferences of voters. The posterior probabilities for this example follow a pattern similar to that of the first example. Let ‘B’ represent the class of all registered voters on February 20, 2004, and let ‘A’ represent those who prefer Kerry to Bush. In sample S (randomly drawn from B with respect to A) consisting of 400 voters, 248 report preference for Kerry over Bush — i.e., $F[A,B] = 248/400 = .62$. Suppose, as seems reasonable, that none of the support functions in the *vagueness* or *diversity* set under consideration rates the hypotheses ‘$F[A,B] = r$’ for values of r outside the interval .62±.05 as *more* initially plausible than they rate alternative frequency hypotheses having values of r inside this interval. That is, suppose, for each P_α under consideration, the prior probabilities $P_\alpha[F[A,B] = s \mid b]$ when s is *not* within .62±.05 is never more than $K = 1$ times as great as the prior probabilities $P_\alpha[F[A,B] = r \mid b]$ for hypotheses having r within .62±.05. Then, the theorem yields the following lower bound on the posterior plausibility ratings, for all P_α in the *vagueness* or *diversity* set under consideration:

$$P_\alpha[.57 < F[A,B] < .67 \mid b \cdot F[A,S]=.62 \cdot \mathrm{Rnd}[S,B,A] \cdot \mathrm{Size}[S]=400] \;\geq\; .9614.$$

The following table gives similar results for other sample sizes, and for upper bounds on ratios of prior probabilities that may be much larger than 1. In addition, this table shows what happens when we tighten up the interval around the frequency hypotheses being supported to .62±.025 — i.e., it shows the bounds p on support for the hypothesis $.595 < F[A,B] < .645$ as well:

Table 2: Values of lower bound p on the posterior probability

m/n = .62	$F[A,B]$ = $m/n \pm q$	Sample-Size = n (number of As in Sample of Bs = m: where m/n = .62)					
Prior Ratio K ↓	q = .05 or .025	400 (248)	800 (496)	1600 (992)	3200 (1984)	6400 (3968)	12800 (7936)
1	.05 →	0.9614	0.9965	1.0000	1.0000	1.0000	1.0000
	.025 →	0.6982	0.8554	0.9608	0.9964	1.0000	1.0000
2	.05 →	0.9256	0.9930	0.9999	1.0000	1.0000	1.0000
	.025 →	0.5364	0.7474	0.9246	0.9929	0.9999	1.0000
5	.05 →	0.8327	0.9827	0.9998	1.0000	1.0000	1.0000
	.025 →	0.3163	0.5420	0.8306	0.9825	0.9998	1.0000
10	.05 →	0.7133	0.9661	0.9996	1.0000	1.0000	1.0000
	.025 →	0.1879	0.3717	0.7103	0.9656	0.9996	1.0000
100	.05 →	0.1992	0.7402	0.9963	1.0000	1.0000	1.0000
	.025 →	0.0226	0.0559	0.1969	0.7371	0.9962	1.0000
1,000	.05 →	0.0243	0.2217	0.9639	1.0000	1.0000	1.0000
	.025 →	0.0023	0.0059	0.0239	0.2190	0.9637	1.0000
10,000	.05 →	0.0025	0.0277	0.7277	0.9999	1.0000	1.0000
	.025 →	0.0002	0.0006	0.0024	0.0273	0.7261	0.9999
100,000	.05 →	0.0002	0.0028	0.2109	0.9994	1.0000	1.0000
	.025 →	0.0000	0.0001	0.0002	0.0028	0.2096	0.9994
1,000,000	.05 →	0.0000	0.0003	0.0260	0.9940	1.0000	1.0000
	.025 →	0.0000	0.0000	0.0000	0.0003	0.0258	0.9943
10,000,000	.05 →	0.0000	0.0000	0.0027	0.9433	1.0000	1.0000
	.025 →	0.0000	0.0000	0.0000	0.0000	0.0026	0.9457

$P_\alpha[.62-q < F[A,B] < .62+q \mid F[A,S] = .62 \cdot \mathrm{Rnd}[S,B,A] \cdot \mathrm{Size}[S] = n] \geq p$, for two values of q (.05 and .025) and a range of Sample-Sizes n (from 400 to 12800), when the prior probability of any specific frequency hypothesis outside of .62 ± q is no more than K times more than the lowest prior probability for any specific frequency hypothesis inside of .62 ± q.

Notice that even if the *vagueness* or *diversity* set includes prior plausibilities nearly *ten million* times higher for hypotheses asserting frequency values *outside of* .62±.025 than for hypotheses asserting frequencies *within* .62±.025, a random sample of 12800 registered voters will, nevertheless, bring about a posterior plausibility value greater than .9457 for the claim that "the true frequency of preference for Kerry over Bush among all registered voters is within .62±.025", for all support functions P_α in the set.

4.2 Convergence to the Truth

The Frequency Estimation Theorem is a Bayesian Convergence-to-Agreement result. It does not, on its own, show that the Criterion of Adequacy (CoA) is satisfied. The theorem shows, for enumerative inductions, that as evidence accumulates, diverse support functions will come to near agreement on high posterior support strengths for those hypotheses expressing population frequencies near the sample frequency. But, it does not show that the true hypothesis is among them — it does not show that the sample frequency is near the true population frequency. So, it does not show that these converging support functions converge on strong support for the true hypothesis, as a CoA result is supposed to do.

However, there is such a CoA result close at hand. It is a *Weak Law of Large Numbers* result that establishes that *each* frequency hypothesis of form '$F[A,B] = r$' implies, via *direct inference likelihoods*, that randomly selected sample data is highly likely to result in sample frequencies very close to the value r that *it* claims to be the *true frequency*. Of course *each* frequency hypothesis says that the sample frequency will be near *its own* frequency value; but only the true hypothesis says this truthfully. Add this result to the previous theorem and we get that, for large sample sizes, it is very likely that a sample frequency will occur that yields a very high degree of support for the true hypothesis. Thus the CoA is satisfied.

Here is the needed result.

Theorem: Weak Law of Large Numbers for Enumerative Inductions.

Let r be any frequency between 0 and 1.

For $r = 0$, $P[F[A,S]=0 \mid F[A,B]=0 \cdot \text{Rnd}[S,B,A] \cdot \text{Size}[S]=n] = 1$.

For $r = 1$, $P[F[A,S]=1 \mid F[A,B]=1 \cdot \text{Rnd}[S,B,A] \cdot \text{Size}[S]=n] = 1$.

For $0 < r < 1$, let q be any real number such that r is in the region,
$0 < (r-q) < r < (r+q) < 1$.

> Given a specific q (which identifies a specific small region of interest around r), for each given positive integer n that's large enough to permit it, we define associated non-negative integers v and u such that $v < u$, where by definition:
>
> > v is the non-negative integer for which v/n is the smallest fraction greater than $(r-q)$, and
> >
> > u is the non-negative integer for which u/n is the largest fraction less than $(r+q)$.
>
> Then,
>
> $$P[r-q < F[A,S] < r+q \mid F[A,B]=r \cdot \text{Rnd}[S,B,A] \cdot \text{Size}[S]=n]$$
>
> $$= \sum_{m=v}^{u} \frac{n!}{m! \times (n-m)!} \times r^m (1-r)^{n-m}$$
>
> $$\approx 1 - 2 \times \Phi[-q/(r\times(1-r))/n)^{1/2}] \geq 1 - 2 \times \Phi[-2\times q\times n^{1/2}],$$
>
> which goes to 1 quickly as n increases.

Here $\Phi[x]$ is the area under the Standard Normal Distribution up to point x. The first equality is a version of the binomial theorem. The approximation of the binomial formula by the normal distribution is guaranteeed by the Central Limit Theorem. This approximation is very close for n near 20, and gets extremely close as n gets larger.

Notice that the degree of support probability in this theorem is a *direct inference likelihood* — all support functions should agree on these values.[17]

This Weak Law result together with the Simple Estimation Theorem yields the promised CoA result: for large sample sizes, it is very likely that a sample frequency will occur that has a value very near the true frequency; and whenever such a sample frequency does occur, it yields a very high degree of support for the true frequency hypothesis.

This result only applies to enumerative inductions. In the next section we establish a CoA result that applies much more generally. It applies to the inductive support of hypotheses in any context where competing hypotheses are empirically distinct enough to disagree, at least a little, on the likelihoods of possible evidential outcomes.

5. The Likelihood Ratio Convergence Theorem

In this section we will investigate the **Likelihood Ratio Convergence Theorem**. This theorem shows that under certain reasonable conditions, when hypothesis h_i (in conjunction with auxiliaries in b) is true and an alternative hypothesis h_j is empirically distinct from h_i on some possible outcomes of experiments or observations described by conditions c_k, then it is *very likely* that a long enough sequence of such experiments and observations c^n will produce a sequence of outcomes e^n that yields likelihood ratios $P[e^n \mid h_j{\cdot}b{\cdot}c^n] / P[e^n \mid h_i{\cdot}b{\cdot}c^n]$ that approach 0 as evidence accumulates (i.e., as n increases). The theorem places an explicit lower bound on the “rate of probable convergence” of these likelihood ratios towards 0. That is, it puts a lower bound on how likely it is, if h_i is true, that a stream of outcomes will occur that yields likelihood ratio values against h_j as compared to h_i that lie within any specified small distance from 0.

The theorem itself does not require the full apparatus of Bayesian probability functions. It draws only on likelihoods. Neither the statement of the theorem nor its proof employ prior probabilities of any kind. Likelihoodists and Bayesian inductivists agree that when the ratios $P[e^n \mid h_j{\cdot}b{\cdot}c^n] / P[e^n \mid h_i{\cdot}b{\cdot}c^n]$ approach 0 for increasing n, the evidence goes strongly against h_j as compared to h_i. So even a *likelihoodist* who eschews the use of Bayesian prior probabilities may embrace this result.

For Bayesians, the *Likelihood Ratio Convergence Theorem* further implies the likely *convergence to agreement* near 0 of the posterior probabilities of false competitors of a true hypothesis. When the ratios $P[e^n \mid h_j{\cdot}b{\cdot}c^n] / P[e^n \mid h_i{\cdot}b{\cdot}c^n]$ approach 0 for increasing n, the Ratio Form of Bayes' Theorem, Equation 9, says that the posterior probability of h_j must also approach 0 as evidence accumulates, regardless of the value of its prior probability. So, support functions in collections representing vague prior plausibilities for an individual agent (i.e., a *vagueness* set) and representing the diverse range of priors for a community of agents (i.e., a *diversity* set) will very likely come to agree on the near 0 posterior probability of empirically distinct false rivals of a true hypothesis. And as the posterior probabilities of false competitors fall, the posterior probability of the true hypothesis heads towards 1. Thus, the theorem establishes that the inductive logic of probabilistic support functions satisfies the Criterion of Adequacy (CoA).

The *Likelihood Ratio Convergence Theorem* overcomes many of the objections raised by critics of Bayesian convergence results. First, this theorem does not employ *second-order probabilities*; it says noting about the probability of a probability. It only concerns the probability of a particular disjunctive sentence that expresses a disjunction of various possible sequences of experimental or observational outcomes. The theorem does not require evidence to consist of sequences of events that, according to the hypothesis, are identically distributed (like repeated tosses of a die). Although the result is most easily expressed in cases where the sequence of outcomes are probabilistically independent relative to each hypothesis, a version of the theorem also holds when the individual outcomes of the evidence stream are not probabilistically independent on the hypotheses. The result does not rely on countable additivity. And the explicit lower bounds it provides on convergence means that there is no need to wait for the infinte long run before convergence occurs (as some critics seem to think).

It is sometimes claimed that Bayesian convergence results only work when an agent locks in values for the prior probabilities of hypotheses once and for all, and updates posterior probabilities from there only by conditioning on evidence via Bayes Theorem. The *Likelihood Ratio Convergence Theorem*, however, applies even if agents revise their prior probability assessments over time. Such non-Bayesian shifts

from one support function (or *vagueness* set) to another may arise from new plausibility arguments or from reassessments of the strengths of old ones. The *Likelihood Ratio Convergence Theorem* itself only involves the values of likelihoods. So, provided such reassessments don't push the prior probability of the true hypothesis towards 0 *too rapidly*, the theorem implies that the posterior probabilities of each empirically distinct false competitor will *very probably* approach 0 as evidence increases.[18]

5.1 The Space of Possible Outcomes of Experiments and Observations

To specify the details of the *Likelihood Ratio Convergence Theorem* we'll need a few additional notational conventions and definitions. Here they are.

For a given sequence of n experiments or observations c^n, consider the set of those possible sequences of outcomes that would result in likelihood ratios for h_j over h_i that are less than some chosen small number $\varepsilon > 0$. This set is represented by the expression:

$$\{e^n : P[e^n \mid h_j{\cdot}b{\cdot}c^n] / P[e^n \mid h_i{\cdot}b{\cdot}c^n] < \varepsilon\}.$$

Placing the disjunction symbol '$\vee$' in front of this expression yields an expression:

$$\vee\{e^n : P[e^n \mid h_j{\cdot}b{\cdot}c^n] / P[e^n \mid h_i{\cdot}b{\cdot}c^n] < \varepsilon\},$$

that we'll use to represent the disjunction of all outcome sequences in this set. So,

$$\vee\{e^n : P[e^n \mid h_j{\cdot}b{\cdot}c^n] / P[e^n \mid h_i{\cdot}b{\cdot}c^n] < \varepsilon\}$$

is just a particular sentence that says, in effect, "one of the sequences of outcomes of the first n experiments or observations will occur that makes the likelihood ratio for h_j over h_i less than ε."

The *Likelihood Ratio Convergence Theorem* says that under certain conditions (covered in detail below), the likelihood of a disjunctive sentence of this sort, given that '$h_i{\cdot}b{\cdot}c^n$' is true,

$$P[\vee\{e^n : P[e^n \mid h_j{\cdot}b{\cdot}c^n]/P[e^n \mid h_i{\cdot}b{\cdot}c^n] < \varepsilon\} \quad | \quad h_i{\cdot}b{\cdot}c^n],$$

must be at least $1-(\psi/n)$, for some explicitly calculable term ψ. Thus, the true hypothesis h_i probabilistically implies that as the amount of evidence, n, increases, it becomes highly likely (as close to 1 as you please) that one of the outcome sequences e^n will occur that yields a likelihood ratio $P[e^n \mid h_j{\cdot}b{\cdot}c^n] / P[e^n \mid h_i{\cdot}b{\cdot}c^n]$ less than ε; and this holds for any specific value of ε you may choose. As this happens, the posterior probability of h_i's false competitor, h_j, must approach 0, as required by the Ratio Form of Bayes' Theorem, Equation 9.

The term ψ in the lower bound of this probability depends on a measure of the empirical distinctness of the hypotheses for the proposed sequence of experiments and observations. To specify this measure we need to contemplate the collection of possible outcomes of each experiment or observation. So, consider some sequence of experimental or observational conditions described by sentences $c_1,c_2,\ldots,c_n$. Corresponding to each condition c_k there will be some range of possible alternative outcomes. Let $O_k = \{o_{k1},o_{k2},\ldots,o_{kw}\}$ be a set of statements describing the alternative possible outcomes for condition c_k. (The number of alternative outcomes will usually differ for distinct experiments $c_1,\ldots,c_n$; so, the value of w depends on c_k.) For each hypothesis h_j, the alternative outcomes of c_k in O_k are mutually exclusive and exhaustive, so we have:

$$P[o_{ku}{\cdot}o_{kv} \mid h_j{\cdot}b{\cdot}c_k] = 0 \text{ and } \sum_{u=1}^{w} P[o_{ku} \mid h_j{\cdot}b{\cdot}c_k] = 1.$$

We now let expressions like 'e_k' act as variables that range over the possible outcomes of c_k— i.e., e_k ranges over the members of O_k. As before, 'c^n' denotes the conjunction of the first n test conditions, $(c_1{\cdot}c_2{\cdot}\ldots{\cdot}c_n)$, and '$e^n$' represents possible sequences of corresponding outcomes, $(e_1{\cdot}e_2{\cdot}\ldots{\cdot}e_n)$. Let's use the expression 'E^n' to represent the set of all possible outcome sequences that may result from the sequence of conditions c^n. So, for each hypothesis h_j (including h_i), $\sum_{e^n \in E^n} P[e^n \mid h_j{\cdot}b{\cdot}c^n] = 1$.

Everything introduced in this subsection is mere notational convention. No substantive suppositions (other than the axioms of probability theory) have yet been introducted. The version of the *Likelihood Ratio Convergence Theorem* I'll present below does,

however, draw on one substantive supposition, although a rather weak one. The next subsection will discuss that supposition in detail.

5.2 Probabilistic Independence

In most scientific contexts the outcomes in a stream of experiments or observations are *probabilistically independent* of one another relative to each hypothesis under consideration, or can at least be divided up into probabilistically independent parts. For our purposes *probabilistic independence of evidential outcomes on a hypothesis* divides neatly into two types.

Definition: Independent Evidence Conditions:

(1) A sequence of outcomes e^k is **condition-independent** of a condition for an additional experiment or observation c_{k+1}, given $h{\cdot}b$ and its own conditions c^k, *if and only if*

$P[e^k \mid h{\cdot}b{\cdot}c^k{\cdot}c_{k+1}] = P[e^k \mid h{\cdot}b{\cdot}c^k]$.

(2) An individual outcome e_k is **result-independent** of a sequence of other observations and their outcomes $(c^{k-1}{\cdot}e^{k-1})$, given $h{\cdot}b$ and its own condition c_k, *if and only if*

$P[e_k \mid h{\cdot}b{\cdot}c_k{\cdot}(c^{k-1}{\cdot}e^{k-1})] = P[e_k \mid h{\cdot}b{\cdot}c_k]$.

When these two conditions hold, the likelihood for an evidence sequence may be decomposed into the product of the likelihoods for individual experiments or observations. To see how the two *independence conditions* affect the decomposition, first consider the following formula, which holds even when neither *independence condition* is satisfied:

$$(12) \quad P[e^n \mid h_j{\cdot}b{\cdot}c^n] = \prod_{k=1}^{n} P[e_k \mid h_j{\cdot}b{\cdot}c^n{\cdot}e^{k-1}].$$

When *condition-independence* holds, the likelihood of the whole evidence stream parses into a product of likelihoods that *probabilistically depend* on only past observation conditions and their outcomes. They do not depend on the conditions for other experiments whose outcomes are not yet specified. Here is the formula:

$$(13)\quad P[e^n \mid h_j\cdot b\cdot c^n] = \prod_{k=1}^{n} P[e_k \mid h_j\cdot b\cdot c_k\cdot (c^{k-1}\cdot e^{k-1})].$$

Finally, whenever both *independence conditions* are satisfied we have the following relationship between the likelihood of the evidence stream and the likelihoods of individual experiments or observations:

$$(14)\quad P[e^n \mid h_j\cdot b\cdot c^n] = \prod_{k=1}^{n} P[e_k \mid h_j\cdot b\cdot c_k].$$

(For proofs of Equations 12-14, see the supplementary document: Immediate Consequences of Independent Evidence Conditions.)

In scientific contexts the evidence can almost always be divided into parts that satisfy both clauses of the *Independent Evidence Condition* with respect to each alternative hypothesis. To see why, let us consider each independence condition more carefully.

Condition-independence says that the mere addition of a new observation condition c_{k+1}, *without specifying one of its outcomes*, does not alter the likelihood of the outcomes e^k of other experiments c^k. To appreciate the significance of this condition, imagine what it would be like if it were violated. Suppose hypothesis h_j is some statistical theory, say, for example, a quantum theory of superconductivity. The conditions expressed in c^k describe a number of experimental setups, perhaps conducted in numerous labs throughout the world, that test a variety of aspects of the theory (e.g., experiments that test electrical conductivity in different materials at a range of temperatures). An outcome sequence e^k describes the results of these experiments. The violation of *condition-independence* would mean that merely adding to $h_j\cdot b\cdot c^k$ a statement c_{k+1} describing how an additional experiment has been set up, but with no mention of its outcome, changes how likely the evidence sequence e^k is taken to be. What $(h_j\cdot b)$ *says* via likelihoods about the outcomes e^k of experiments c^k differs as a result of merely supplying a description of another experimental arrangement, c_{k+1}. *Condition-independence*, when it holds, rules out such strange effects.

Result-independence says that the description of previous test conditions *together with their outcomes* is irrelevant to the likelihoods of outcomes for additional experiments. If this condition were widely violated, then in order to specify the most informed likelihoods for a given hypothesis one would need to include information about volumes of past observations and their outcomes. What a hypothesis says about future cases would depend on how past cases have gone. Such *dependence* had better not happen on a large scale. Otherwise, the hypothesis would be fairly useless, since its empirical import in each specific case would depend on taking into account volumes of past observational and experimental results. However, even if such dependencies occur, provided they are not too pervasive, *result-independence* can be accommodated rather easily by packaging each collection of *result-dependent* data together, treating it like a single extended experiment or observation. The *result-independence condition* will then be satisfied by letting each term 'c_k' in the statement of the independence condition represent a conjunction of test conditions for a collection of *result-dependent* tests, and by letting each term 'e_k' (and each term 'o_{ku}') stand for a conjunction of the corresponding *result-dependent* outcomes. Thus, by packaging *result-dependent* data together in this way, the *result-independence* condition is satisfied by those (conjunctive) statements that describe the separate, *result-independent* chunks.[19]

The version of the *Likelihood Ratio Convergence Theorem* we will examine depends only on the *Independent Evidence Conditions* (together with the axioms of probability theory). It draws on no other assumptions. Indeed, an even more general version of the theorem can be established that draws on neither of the *Independent Evidence Conditions*. However, the *Independent Evidence Conditions* will be satisfied in almost all scientific contexts, so little will be lost by assuming them. (And the presentation will run more smoothly if we side-step the added complications needed to explain the more general result.)

From this point on let us assume that the following versions of the *Independent Evidence Conditions* hold.

Assumption: Independent Evidence Assumptions. For each hypothesis h and background b under consideration, we assume that the experiments and observations can be packaged into condition statements, $c_1, \ldots, c_k, c_{k+1}, \ldots$, and possible outcomes in a way that satisfies the following conditions:

(1) Each sequence of possible outcomes e^k of a sequence of conditions c^k is **condition-independent** of additional conditions c_{k+1} — i.e., $P[e^k \mid h{\cdot}b{\cdot}c^k{\cdot}c_{k+1}] = P[e^k \mid h{\cdot}b{\cdot}c^k]$.

(2) Each possible outcome e_k of condition c_k is **result-independent** of sequences of other observations and possible outcomes $(c^{k-1}{\cdot}e^{k-1})$ — i.e., $P[e_k \mid h{\cdot}b{\cdot}c_k{\cdot}(c^{k-1}{\cdot}e^{k-1})] = P[e_k \mid h{\cdot}b{\cdot}c_k]$.

We now have all that is needed to begin to state the *Likelihood Ratio Convergence Theorem*.

5.3 Likelihood Ratio Convergence when Falsifying Outcomes are Possible

The *Likelihood Ratio Convergence Theorem* comes in two parts. The first part applies only to experiments or observations c_k in the total evidence stream c^n for which some of the possible outcomes have 0 probability of occurring according to hypothesis h_j but have non-0 likelihood of occurring according to h_i. Such outcomes are highly desirable. If they occur, the likelihood ratio comparing h_j to h_i will become 0, and h_j will be *falsified. Crucial experiments* are a special case of this — the case where for at least one possible outcome o_{ku}, $P[o_{ku} \mid h_i{\cdot}b{\cdot}c_k] = 1$ and $P[o_{ku} \mid h_j{\cdot}b{\cdot}c_k] = 0$. In the more general case h_i together with b says that one of the outcomes of c_k is at least minimally probable, whereas h_j says that outcome is impossible — i.e., $P[o_{ku} \mid h_i{\cdot}b{\cdot}c_k] > 0$ and $P[o_{ku} \mid h_j{\cdot}b{\cdot}c_k] = 0$. It will be convenient to define a term for this situation.

Definition: Full Outcome Compatibility. Let's call h_j *fully outcome-compatible* with h_i on experiment or observation c_k *just when* for each of its possible outcomes e_k, if $P[e_k \mid h_i{\cdot}b{\cdot}c_k] > 0$, then $P[e_k \mid h_j{\cdot}b{\cdot}c_k] > 0$. Equivalently, h_j is *fails to be fully outcome-compatible* with h_i on experiment or observation c_k *just when* for at least one of its possible outcomes e_k, $P[e_k \mid h_i{\cdot}b{\cdot}c_k] > 0$ but $P[e_k \mid h_j{\cdot}b{\cdot}c_k] = 0$.

The first part of the *Likelihood Ratio Convergence Theorem* applies to that part of the total stream of evidence (i.e. that subsequence of the total evidence stream) on which hypothesis h_j *fails to be fully outcome-compatible* with hypothesis h_i; the second part of the theorem applies to the remaining part of the total stream of evidence, that subsequence of the total evidence stream on which h_j is *fully outcome-compatible* with h_i for each experiment and observation. It turns out that these two kinds of cases must be treated differently. (This is due to the way in which the *expected information content* of for distinguishing between the two hypotheses will be measured for experiments and observations that are *fully outcome compatible*; this measure of information content blows up (becomes infinite) for experiments and observations that *fail to be fully outcome compatible*). Thus, the following part of the convergence theorem applies to just that part of the total stream of evidence that consists of experiments and observations that *fail to be fully outcome compatible* for the pair of hypotheses involved. Here, then, is the first part of the theorem.

Likelihood Ratio Convergence Theorem 1 — The Falsification Theorem: Suppose that the total stream of evidence c^n contains precisely m experiments or observations on which h_j *fails to be fully outcome-compatible* with h_i. And suppose that the *Independent Evidence Conditions* hold for evidence stream c^n with respect to each of these two hypotheses. Furthermore, suppose there is a lower bound $\delta > 0$ such that for each c_k on which h_j *fails to be fully outcome-compatible* with h_i, $P[\vee\{o_{ku} : P[o_{ku} \mid h_j{\cdot}b{\cdot}c_k] = 0\} \mid h_i{\cdot}b{\cdot}c_k] \geq \delta$ — i.e. h_i together with $b{\cdot}c_k$ *says*, with likelihood at least as large as δ, that one of the outcomes will occur that h_j *says* cannot occur. Then,

$$P[\vee\{e^n : P[e^n \mid h_j{\cdot}b{\cdot}c^n]/P[e^n \mid h_i{\cdot}b{\cdot}c^n] = 0\} \quad \mid \quad h_i{\cdot}b{\cdot}c^n]$$

$$= \quad P[\vee\{e^n : P[e^n \mid h_j{\cdot}b{\cdot}c^n] = 0\} \quad \mid \quad h_i{\cdot}b{\cdot}c^n]$$

$$\geq \quad 1-(1-\delta)^m,$$

which approaches 1 for large m. (For proof see the supplementary document Proof of the Falsification Theorem.)

In other words, we only suppose that for each of m observations, c_k, (drawn from the total stream of all n observations, c^n), h_i *says* observation c_k has at least a small likelihood δ of producing one of the outcomes o_{ku} that h_j *says* is impossible. If the number m of such experiments or observations is large enough (or if the lower bound δ on the likelihoods of getting such outcomes is large enough), and if h_i (together with $b{\cdot}c^n$) is true, then it is highly likely that one of the outcomes held to be impossible by h_j will actually occur. If one of these outcomes does occur, then the likelihood ratio for h_j as compared to over h_i will become 0. According to Bayes' Theorem, when this happen, h_j is absolutely refuted by the evidence — its posterior probability becomes 0.

The Falsification Theorem is very commonsensical. First, notice that if there is a *crucial experiment* in the evidence stream, the theorem is completely obvious. That is, suppose for the specific experiment c_k (in evidence stream c^n) there are two incompatible possible outcomes o_{kv} and o_{ku} such that $P[o_{kv} \mid h_j{\cdot}b{\cdot}c_k] = 1$ and $P[o_{ku} \mid h_i{\cdot}b{\cdot}c_k] = 1$. Then, clearly, $P[\vee\{o_{ku}: P[o_{ku} \mid h_j{\cdot}b{\cdot}c_k] = 0\} \mid h_i{\cdot}b{\cdot}c_k] = 1$, since o_{ku}

is one of the o_{ku} such that $P[o_{ku} \mid h_j{\cdot}b{\cdot}c_k] = 0$. So where there is a crucial experiment available, the theorem applies with $m = 1$ and $\delta = 1$.

The theorem is equally commonsensical for cases where no crucial experiment is available. To see what it says in such cases, consider an example. Let h_i be some theory that implies a specific rate of proton decay, but a rate so low that there is only a very small probability that any particular proton will decay in a given year. Consider an alternative theory h_j that implies that protons *never* decay. If h_i is true, then for a persistent enough sequence of observations (i.e., if proper detectors can be built and trillions of protons kept under observation for long enough), eventually a proton decay will almost surely be detected. When this happens, the likelihood ratio becomes 0. Thus, the posterior probability of h_j becomes 0.

It is instructive to plug some specific values into the formula given by the Falsification Theorem, to see what the convergence rate might look like. For example, the theorem tells us that if we compare any pair of hypotheses h_i and h_j on an evidence stream c^n that contains at least $m = 19$ observations or experiments having $\delta \geq .10$ for the likelihood of yielding a *falsifying outcome*, then the likelihood (on $h_i{\cdot}b{\cdot}c^n$) of obtaining an outcome sequence e^n that yields likelihood-ratio $P[e^n \mid h_j{\cdot}b{\cdot}c^n] / P[e^n \mid h_i{\cdot}b{\cdot}c^n] = 0$, will be at least as large as $1-(1-.1)^{19} = .865$. (The reader is invited to try other values of δ and m as well.)

A comment about the *need for* and *usefulness of* such convergence theorems is in order, now that we've seen one. Given some specific pair of scientific hypotheses h_i and h_j one may directly compute the likelihood, given ($h_i{\cdot}b{\cdot}c^n$), that a proposed sequence of experiments or observations c^n will result in one of the sequences of outcomes that yield low likelihood ratios. So, given a specific pair of hypotheses and a proposed sequence of experiments, we don't need a general *Convergence Theorem* to tell us the likelihood of obtaining refuting evidence. The specific hypotheses h_i and h_j tell us this *themselves*. They tell us the likelihood of obtaining each specific outcome stream, including those that refute the competitor or produce a very small likelihood ratio for it. Furthermore, after we've actually performed an experiment and recorded its outcome, all that matters is the actual ratio of likelihoods for that outcome. Convergence theorems become moot.

The point of the Likelihood Ratio Convergence Theorem (both the Falsification Theorem and the part of the theorem still to come) is to assure us *in advance of the consideration of any specific pair of hypotheses* that if the possible evidence streams that test hypotheses have certain characteristics which reflect the empirical distinctness of the hypotheses, then it is highly likely that one of the sequences of outcomes will occur that yields a very small likelihood ratio. These theorems provide relatively meager, but finite lower bounds on how quickly such convergence is likely to be. Thus, they show that the CoA is satisfied in advance of our using the logic to test specific pairs of hypotheses against one another.

5.4 Likelihood Ratio Convergence When No Falsifying Outcomes are Possible

The Falsification Theorem applies whenever the evidence stream includes possible outcomes that may *falsify* the alternative hypothesis. However, it only takes account of the influence of the possibly falsifying experiments or observations. It completely ignores the influence of any experiments or observations in the evidence stream on which hypothesis h_j is *fully outcome-compatible* with hypothesis h_i. We now turn to a theorem that applies to those evidence streams (or to parts of evidence streams) consisting only of experiments and observations on which hypothesis h_j is *fully outcome-compatible* with hypothesis h_i. Evidence streams of this kind contain no *possibly falsifying* outcomes. In such cases the only outcomes of an experiment or observation c_k for which hypothesis h_j may specify 0 likelihoods are those for which hypothesis h_i specifies 0 likelihoods as well.

Hypotheses whose connection with the evidence is entirely statistical in nature will inevitably be *fully outcome-compatible* on the entire evidence stream. So, evidence streams of this kind are undoubtedly much more common in practice than those containing possibly falsifying outcomes. Furthermore, whenever an entire stream of evidence contains some mixture of experiments and observations on which the hypotheses are *not fully outcome compatible* along with others on which they are *fully outcome compatible*, we may treat the experiments and observations for which *full outcome compatiblity* holds as a seperate subsequence of the entire evidence stream, to see the likely impact of that part of the evidence in producing values for likelihood ratios.

To cover evidence streams (or subsequences of evidence streams) consisting entirely of experiments or observations on which h_j is *fully outcome-compatible* with hypothesis h_i we will first need to identify a useful way to measure the degree to which hypotheses are empirically distinct from one another on such evidence. Consider some particular sequence of outcomes e^n that results from observations c^n. The likelihood ratio $P[e^n \mid h_j{\cdot}b{\cdot}c^n] / P[e^n \mid h_i{\cdot}b{\cdot}c^n]$ itself measures the extent to which the outcome sequence distinguishes between h_i and h_j. But as a measure of the power of evidence to distinguish among hypotheses, raw likelihood ratios provide a rather lopsided scale, a scale that ranges from 0 to infinity with the midpoint, where e^n doesn't distinguish at all between h_i and h_j, at 1. So, rather than using raw likelihood ratios to measure the ability of e^n to distinguish between hypotheses, it proves more useful to employ a symmetric measure. The logarithm of the likelihood ratio provides such a measure.

> **Definition**: **QI — the Quality of the Information**.
> For each experiment or observation c_k, define *the quality of the information* provided by possible outcome o_{ku} for distinguishing h_j from h_i, given b, as follows (where henceforth we take "logs" to be base-2):
>
> $$QI[o_{ku} \mid h_i/h_j \mid b{\cdot}c_k] = \log[P[o_{ku} \mid h_i{\cdot}b{\cdot}c_k]/P[o_{ku} \mid h_j{\cdot}b{\cdot}c_k]].$$
>
> Similarly, for the sequence of experiments or observations c^n, define *the quality of the information* provided by possible outcome e^n for distinguishing h_j from h_i, given b, as follows:
>
> $$QI[e^n \mid h_i/h_j \mid b{\cdot}c^n] = \log[P[e^n \mid h_i{\cdot}b{\cdot}c^n]/P[e^n \mid h_j{\cdot}b{\cdot}c^n]].$$
>
> That is, QI is the base-2 logarithm of the likelihood ratio for h_i over that for h_j.

So, we'll measure the *Quality of the Information* an outcome would yield in distinguishing between two hypotheses as the base-2 logarithm of the likelihood ratio. This is clearly a symmetric measure of the outcome's evidential strength at distinguishing between the two hypotheses. On this measure hypotheses h_i and h_j assign the same likelihood value to a given outcome o_{ku} *just when*QI$[o_{ku} \mid h_i/h_j \mid b{\cdot}c_k] = 0$. Thus, QI measures information on a logarithmic scale that is symmetric about the natural no-information midpoint, 0. This measure is set up so that *positive information* favors h_i over h_j, and *negative information* favors h_j over h_i.

Given the *Independent Evidence Assumptions* with respect to each hypothesis, it's easy to show that the QI for a sequence of outcomes is just the sum of the QIs of the individual outcomes in the sequence:

$$(15) \quad \mathrm{QI}[e^n \mid h_i/h_j \mid b\cdot c^n] = \sum_{k=1}^{n} \mathrm{QI}[e_k \mid h_i/h_j \mid b\cdot c_k].$$

Probability theorists measure the *expected value* of a quantity by first multiplying each of its *possible values* by their probabilities of occurring, and then summing these products. Thus, the *expected value* of QI is given by the following formula:

Definition: **EQI — the Expected Quality of the Information**.
We adopt the convention that if $P[o_{ku} \mid h_i\cdot b\cdot c_k] = 0$, then the term $\mathrm{QI}[o_{ku} \mid h_i/h_j \mid b\cdot c_k] \times P[o_{ku} \mid h_i\cdot b\cdot c_k] = 0$. This convention will make good sense in the context of the following definition because, whenever the outcome o_{ku} has 0 probability of occurring according to h_i (together with $b\cdot c_k$), it makes good sense to give it 0 impact on the ability of the evidence to distinguish between h_j and h_i when h_i (together with $b\cdot c_k$) is true. Also notice that the *full outcome-compatiblity* of h_j with h_i on c_k means that whenever $P[e_k \mid h_j\cdot b\cdot c_k] = 0$, we must have $P[e_k \mid h_i\cdot b\cdot c_k] = 0$ as well; so whenever the denominator would be 0 in the term $\mathrm{QI}[o_{ku} \mid h_i/h_j \mid b\cdot c_k] = \log[P[o_{ku} \mid h_i\cdot b\cdot c_k]/P[o_{ku} \mid h_j\cdot b\cdot c_k]]$, the the convention just described makes the term $\mathrm{QI}[o_{ku} \mid h_i/h_j \mid b\cdot c_k] \times P[o_{ku} \mid h_i\cdot b\cdot c_k] = 0$. Thus the following notion is well-defined:

For h_j *fully outcome-compatible* with h_i on experiment or observation c_k, define

$$\mathrm{EQI}[c_k \mid h_i/h_j \mid b] = \sum_u \mathrm{QI}[o_{ku} \mid h_i/h_j \mid b\cdot c_k] \times P[o_{ku} \mid h_i\cdot b\cdot c_k].$$

Also, for h_j *fully outcome-compatible* with h_i on each experiment and observation in the sequence c^n, define

$$\mathrm{EQI}[c^n \mid h_i/h_j \mid b] = \sum_{e^n\in E^n} \mathrm{QI}[e^n \mid h_i/h_j \mid b\cdot c^n] \times P[e^n \mid h_i\cdot b\cdot c^n].$$

The EQI of an experiment or observation is the *Expected Quality of its Information* for distinguishing h_i from h_j when h_i is true. It is a measure of the expected evidential strength of the possible outcomes of an experiment or observation at distinguishing between the hypotheses when h_i (together with $b\cdot c$) is true. Whereas QI measures the ability of each particular outcome or sequence of outcomes to empirically distinguish hypotheses, EQI measures the tendency of experiments or observations to produce distinguishing outcomes. It can be shown that EQI tracks empirical distinctness in a very precise way. We return to this in a moment.

It is easily seen that the EQI for a sequence of observations c^n is just the sum of the EQIs of the individual observations c_k in the sequence:

(16) $$\mathrm{EQI}[c^n \mid h_i/h_j \mid b] = \sum_{k=1}^{n} \mathrm{EQI}[c_k \mid h_i/h_j \mid b].$$

(For proof see the supplementary document Proof that the EQI for c^n is the sum of the EQI for the individual c_k.)

This suggests that it may be useful to average the values of the EQI[$c_k \mid h_i/h_j \mid b$] over the number of observations n to obtain a measure of the *average expected quality of the information* among the experiments and observations that make up the evidence stream c^n.

Definition: **The Average Expected Quality of Information**
For h_j *fully outcome-compatible* with h_i on each experiment and observation in the evidence stream c^n, define the average expected quality of information, $\overline{\mathrm{EQI}}$, from c^n for distinguishing h_j from h_i, given $h_i \cdot b$, as follows:

$$\begin{aligned} \overline{\mathrm{EQI}}[c^n \mid h_i/h_j \mid b] &= \mathrm{EQI}[c^n \mid h_i/h_j \mid b] \div n \\ &= (1/n) \times \sum_{k=1}^{n} \mathrm{EQI}[c_k \mid h_i/h_j \mid b]. \end{aligned}$$

It turns out that the value of EQI[$c_k \mid h_i/h_j \mid b$] cannot be less than 0; and it will be greater just in case h_i is *empirically distinct* from h_j on at least one outcome o_{ku} — i.e., just in case it is *empirically distinct* in the sense that $P[o_{ku} \mid h_i \cdot b \cdot c_k] \neq P[o_{ku} \mid h_j \cdot b \cdot c_k]$. The same goes for the average, $\overline{\mathrm{EQI}}[c^n \mid h_i/h_j \mid b]$.

Theorem: Nonnegativity of EQI.

EQI[$c_k \mid h_i/h_j \mid b$] ≥ 0; and EQI[$c_k \mid h_i/h_j \mid b$] > 0 *if and only if* for at least one of its possible outcomes o_{ku}, $P[o_{ku} \mid h_i \cdot b \cdot c_k] \neq P[o_{ku} \mid h_j \cdot b \cdot c_k]$.

As a result, $\overline{\mathrm{EQI}}[c^n \mid h_i/h_j \mid b] \geq 0$; and $\overline{\mathrm{EQI}}[c^n \mid h_i/h_j \mid b] > 0$ *if and only if* at least one experiment or observation c_k has at least one possible outcome o_{ku} such that $P[o_{ku} \mid h_i \cdot b \cdot c_k] \neq P[o_{ku} \mid h_j \cdot b \cdot c_k]$.

(For proof, see the supplementary document The Effect on EQI of Partitioning the Outcome Space More Finely — Including Proof of the Nonnegativity of EQI.)

In fact, the more finely one partitions the outcome space $O_k = \{o_{k1},\ldots,o_{kv},\ldots,o_{kw}\}$ into distinct outcomes that differ on likelihood ratio values, the larger EQI becomes.[20] This shows that EQI tracks empirical distinctness in a precise way. The importance of the *Non-negativity of EQI* result for the *Likelihood Ratio Convergence Theorem* will become clear in a moment.

We are now in a position to state the second part of the *Likelihood Ratio Convergence Theorem*. It applies to all evidence streams not containing *possibly falsifying outcomes* for h_j when h_i holds — i.e., it applies to all evidence streams for which h_j is *fully outcome-compatible* with h_i on each c_k in the stream.

Likelihood Ratio Convergence Theorem 2 — The Non-Falsifying Refutation Theorem.

Suppose the evidence stream c^n contains only experiments or observations on which h_j is *fully outcome-compatible* with h_i — i.e. suppose that for each condition c_k in sequence c^n, for each of its possible outcomes possible outcomes o_{ku}, either $P[o_{ku} \mid h_i \cdot b \cdot c_k] = 0$ or $P[o_{ku} \mid h_j \cdot b \cdot c_k] > 0$. In addition (as a slight strengthening of the previous supposition), for some $\gamma > 0$ a number smaller than $1/e^2$ ($\approx .135$; where 'e' is the base of the natural logarithm), suppose that for each possible outcome o_{ku} of each observation condition c_k in c^n, either $P[o_{ku} \mid h_i \cdot b \cdot c_k] = 0$ or $P[o_{ku} \mid h_j \cdot b \cdot c_k] / P[o_{ku} \mid h_i \cdot b \cdot c_k] \geq \gamma$. And suppose that the *Independent Evidence Conditions* hold for evidence stream c^n with respect to each of these hypotheses. Now, choose any positive $\varepsilon < 1$, as small as you like, but large enough (for the number of observations n being contemplated) that the value of $\overline{\mathrm{EQI}}[c^n \mid h_i/h_j \mid b] > -(\log \varepsilon)/n$. Then:

$$P[\vee\{e^n : P[e^n \mid h_j \cdot b \cdot c^n] / P[e^n \mid h_i \cdot b \cdot c^n] < \varepsilon\} \quad \mid \quad h_i \cdot b \cdot c^n]$$

$$> \quad 1 \quad - \quad \frac{1}{n} \times \frac{(\log \gamma)^2}{(\overline{\mathrm{EQI}}[c^n \mid h_i/h_j \mid b] + (\log \varepsilon)/n\ )^2}$$

(For proof see the supplementary document Proof of the Non-Falsifying Refutation Theorem.)

This theorem provides sufficient conditions for the *likely refutation* of false alternatives via exceeding small likelihood ratios. The conditions under which this happens characterize the degree to which the hypotheses involved are empirically distinct from one another. The theorem says that when these conditions are met, according to hypothesis h_i (taken together with $b \cdot c^n$), the likelihood is near 1 that that one of the outcome sequence e^n will occur for which the likelihood ratio is smaller than ε (for any value of ε you may choose). The likelihood of getting such an evidential outcome e^n is quite coles to 1 — i.e. no more than the amount $(1/n) \times (\log \gamma)^2 / (\overline{\mathrm{EQI}}[c^n \mid h_i/h_j \mid b] + (\log \varepsilon)/n)^2$ below 1. (Notice that this amount below 1 goes to 0 as n increases.)

It turns out that in almost every case (for almost any pair of hypotheses) the actual likelihood of obtaining such evidence (i.e. evidence that has a likelihood ratio value less than ε) will be *much closer* to 1 than this factor indicates.[21] Thus, the theorem provides an overly cautious lower bound on the likelihood of obtaining small likelihood ratios. It shows that the larger the value of $\overline{\mathrm{EQI}}$ for an evidence stream, the more likely that stream is to produce a sequence of outcomes that yield a very small likelihood ratio value. But even if $\overline{\mathrm{EQI}}$ remains quite small, a long enough evidence stream, n, of such low-grade evidence will, nevertheless, almost surely produce an outcome sequence having a very small likelihood ratio value.[22]

Notice that the antecedent condition of the theorem, that "either $P[o_{ku} \mid h_i \cdot b \cdot c_k] = 0$ or $P[o_{ku} \mid h_j \cdot b \cdot c_k] / P[o_{ku} \mid h_i \cdot b \cdot c_k] \geq \gamma$, for some $\gamma > 0$ but less than $1/e^2$ ($\approx .135$)", does not favor hypothesis h_i over h_j in any way. The condition only rules out the possibility that some outcomes might furnish *extremely strong* evidence *against* h_j relative to h_i — by making $P[o_{ku} \mid h_i \cdot b \cdot c_k] = 0$ or by making $P[o_{ku} \mid h_j \cdot b \cdot c_k] / P[o_{ku} \mid h_i \cdot b \cdot c_k]$ less than some quite small γ. This condition is only needed because our measure of evidential distinguishability, QI, blows up when the ratio $P[o_{ku} \mid h_j \cdot b \cdot c_k] / P[o_{ku} \mid h_i \cdot b \cdot c_k]$ is extremely small. Furthermore, this condition is really no restriction at all on possible experiments or observations. If c_k has some possible outcome sentence o_{ku} that would make $P[o_{ku} \mid h_j \cdot b \cdot c_k] / P[o_{ku} \mid h_i \cdot b \cdot c_k] < \gamma$ (for a given small γ of interest), one may disjunctively lump o_{ku} together with some other outcome sentence o_{kv} for c_k. Then, the antecedent condition of the theorem will be satisfied, but with the sentence '$(o_{ku} \vee o_{kv})$'treated as a single outcome. It can be proved that the only effect of such "disjunctive lumping" is to make $\overline{\mathrm{EQI}}$ smaller than it would otherwise be (whereas larger values of $\overline{\mathrm{EQI}}$ are more desirable). If the *too strongly refuting* disjunct o_{ku} actually occurs when the experiment or observation c_k is conducted, all the better, since the result is to yield a likelihood ratio $P[o_{ku} \mid h_j \cdot b \cdot c_k] / P[o_{ku} \mid h_i \cdot b \cdot c_k]$ smaller than γ on that particular evidential outcome. We merely failed to take this *more strongly refuting* possibility into account when computing our lower bound on the likelihood that refutation via likelihood ratios would occur.

The point of the two *Convergence Theorems* explored in this section is to assure us, in advance of the consideration of any specific pair of hypotheses, that if the possible evidence streams that test them have certain characteristics which reflect their evidential distinguishability, it is highly likely that outcomes yielding small likelihood ratios will result. These theorems provide finite lower bounds on how quickly convergence is likely to occur, bounds that show one need not wait for convergence through some infinitely long run. Indeed, for any evidence sequence on which the probability distributions are at all well behaved, the *actual likelihood* of obtaining outcomes that yield small likelihood ratio values will inevitably be *much higher* than the lower bounds given by Theorems 1 and 2.

In sum, according to Theorems 1 and 2, each hypothesis h_i *says*, via likelihoods, that given enough observations, *it* is very likely to dominate its empirically distinct rivals in a contest of likelihood ratios. The true hypothesis speaks truthfully about this, and its competitors lie. Even a sequence of observations with an extremely low *average expected quality of information* is very likely to do the job if that evidential sequence is long enough. Thus (by Equation 9), as evidence accumulates, the *degree of support* for false hypotheses will very probably approach 0, indicating that they are probably false; and as this happens, (by Equations 10 and 11) the degree of support for the true hypothesis will approach 1, indicating its probable truth. Thus, the **Criterion of Adequacy** (CoA) is satisfied.

6. When the Likelihoods are Vague or Diverse

Up to this point we have been supposing that likelihoods possess objective or agreed numerical values. Although this supposition is often satisfied in scientific contexts, there are important settings where it is unrealistic, where hypotheses only support vague likelihood values, and where there is enough ambiguity in what hypotheses *say* about evidential claims that the scientific community cannot agree on precise values for the likelihoods of evidential claims.[23] Let us now see how the supposition of precise, agreed likelihood values may be relaxed in a reasonable way.

Recall why agreement, or near agreement, on precise values for likelihoods is so important to the scientific enterprise. To the extent that members of a scientific community disagree on the likelihoods, they disagree about the empirical content of their hypotheses, about what each hypothesis *says* about how the world is likely to be. This can lead to disagreement about which hypotheses are refuted or supported by a given body of evidence. Similarly, to the extent that the values of likelihoods are only vaguely implied by hypotheses as understood by an individual agent, that agent may be unable to determine which of several hypotheses is refuted or supported by a given body of evidence.

We have seen, however, that the values of individual likelihoods are not what is most crucial to the way evidence impacts hypotheses. Rather, as Equations 9-11 show, it is *ratios of likelihoods* that do the heavy lifting. So, even if two support functions P_α and P_β disagree on the values of individual likelihoods, they may, nevertheless, largely agree on the refutation or support that accrues to various rival hypotheses, provided that the following condition is satisfied:

> **Directional Agreement Condition**:
> The likelihood ratios due to each of a pair of support functions P_α and P_β are said to *agree in direction* (with respect to the possible outcomes of experiments or observations relevant to a pair of hypotheses) *just in case*
>
> - whenever possible outcome sequence e^n makes $P_\alpha[e^n \mid h_j \cdot b \cdot c^n] / P_\alpha[e^n \mid h_i \cdot b \cdot c^n] < 1$, it also makes $P_\beta[e^n \mid h_j \cdot b \cdot c^n] / P_\beta[e^n \mid h_i \cdot b \cdot c^n] < 1$;
> - whenever possible outcome sequence e^n makes $P_\alpha[e^n \mid h_j \cdot b \cdot c^n] / P_\alpha[e^n \mid h_i \cdot b \cdot c^n] > 1$, it also makes $P_\beta[e^n \mid h_j \cdot b \cdot c^n] / P_\beta[e^n \mid h_i \cdot b \cdot c^n] > 1$;
> - each of these likelihood ratios is either extremely close to 1 for both of these support functions or for neither of these support functions. [24]

When this condition holds, the evidence will support h_i over h_j according to P_α just in case it does so for P_β as well, although the strength of support may differ. Furthermore, the *rate* at which the likelihood ratios increase or decrease on a stream of evidence may differ for the two support functions, but the impact of the cumulative evidence should ultimately affect their refutation or support in much the same way.

When likelihoods are vague or diverse, we may take the approach we employed for *vague* and *diverse* prior plausibility assessments. We may extend the *vagueness sets* for individual agents to include a collection of inductive support functions that cover the range of values for likelihood ratios of evidence claims that the hypotheses apparently support (as well as covering the ranges of prior comparative support strengths for hypotheses due to plausibility arguments within b). Similarly, we may extend the *diversity sets* for communities of agents to include support functions that cover the ranges of likelihood ratio values (along with ranges of prior comparative support strengths for hypotheses) drawn from the *vagueness sets* of members of the scientific community.

This broading of *vagueness* and *diversity* sets to accommodate vague and diverse likelihood values makes no trouble for the *convergence to truth results* for hypotheses. For, provided that the *Directional Agreement Condition* is satisfied by all support functions in an extended *vagueness* or *diversity set* under consideration, the *Likelihood Ratio Convergence Theorem* applies to the whole range of support functions in that set. The the proof of the theorem doesn't depend on the supposition that likelihoods are objective or have intersubjectively agreed values. It applies to each individual support function P_α. The only problem with applying this result across a range of support functions is that when their values for likelihoods differ, function P_α may disagree with P_β on which of the hypotheses is favored by a given sequence of evidence. That can happen because different support functions may represent the evidential import of hypotheses differently, by specifying different likelihood values for the very same evidence claims. So, an evidence stream that favors h_i according to P_α may instead favor h_j according to P_β. However, when the *Directional Agreement Condition* holds for a given collection of support functions, this cannot happen. *Directional Agreement* means that the evidential import of hypotheses is similar enough for P_α and $P_\beta$that a sequence of outcomes may favor a hypothesis according to P_α only if it does so for P_β as well.

Thus, when the *Directional Agreement Condition* holds for all support functions in a *vagueness* or *diversity* set extended to include vague or diverse likelihoods, if enough evidentially distinguishing experiments or observations can be performed, all support functions in the extended *vagueness* or *diversity* set will very probably come to agree that the likelihood ratios for empirically distinct false competitors of a true hypothesis are extremely small. As that happens, the community comes to agree on the refutation of these competitors, and the true hypothesis rises to the top of the heap.[25]

What if the true hypothesis has evidentially equivalent rivals? Their posterior probabilities must rise as well. In that case we are only assured that the disjunction of the true hypothesis with its evidentially equivalent rivals will be driven to 1 as evidence lays low its evidentially distinct rivals. The true hypothesis will itself approach 1 only if either it has no evidentially equivalent rivals, or whatever equivalent rivals it has are laid low by plausibility arguments of the kind that don't depend on the evidential likelihoods.

Bibliography

- Boole, G., 1854. *The Laws of Thought*, London: MacMillan. Republished in 1958 by Dover: New York.
- Bovens, L., Hartmann, S., 2003. *Bayesian Epistemology*, Oxford: Oxford University Press.
- Carnap, R., 1950. *Logical Foundations of Probability*, Chicago: University of Chicago Press.
- –—, 1952. *The Continuum of Inductive Methods*, Chicago: University of Chicago Press.
- –—, 1963. “Replies and Systematic Expositions”, in *The Philosophy of Rudolf Carnap*, P. A. Schilpp (ed.), Open Court, Illinois: La Salle.
- Chihara, C., 1987. “Some Problems for Bayesian Confirmation Theory”, *British Journal for the Philosophy of Science* 38, 551-560.
- Christensen, D., 1999. “Measuring Evidence”, *Journal of Philosophy* 96, 437-61.
- –—, 2004. *Putting Logic in its Place: Formal Constraints on Rational Belief*, Oxford: Oxford University Press.
- De Finetti, B., 1937. “La Prévision: Ses Lois Logiques, Ses Sources Subjectives”, *Annales de l’Institut Henri Poincaré*, 7, 1-68; translated as “Foresight. Its Logical Laws, Its Subjective Sources”, in *Studies in Subjective Probability*, H. E. Kyburg, Jr. and H. E. Smokler (eds.), Robert E. Krieger Publishing Company, 1980.
- Dowe, D., Gardner, S., Oppy, G., 2007. “Bayes, Not Bust: Why Simplicity is No Problem for Bayesians”, *British Journal for the Philosophy of Science* 58, 709-754.
- Dubois, D., Prade, H., 1980. *Fuzzy Sets and Systems*, New York: Academic Press.
- –—, 1990. "An Introduction to Possibilistic and Fuzzy Logics", in G. Shafer and J. Pearl (eds.), *Readings in Uncertain Reasoning*, 742-761. Morgan Kaufmann.
- Duhem, P., 1906. *La theorie physique. Son objet et sa structure*, Paris: Chevalier et Riviere; translated by P.P. Wiener, *The Aim and Structure of Physical Theory*, Princeton, NJ: Princeton University Press, 1954.

- Earman, J., 1992. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*, Cambridge, MA: MIT Press.
- Edwards, A.W.F., 1972. *Likelihood*, Cambridge: Cambridge University Press.
- Edwards, W., Lindman, H., and Savage, L. J., 1963. “Bayesian Statistical Inference for Psychological Research”, *Psychological Review* 70, 193-242.
- Eells, E., 1985. “Problems of Old Evidence”, *Pacific Philosophical Quarterly*, 66, 283-302.
- –––, 2005. “Confirmation Theory”, in J. Pfeifer and S. Sarkar (eds.), *The Philosophy of Science: An Encyclopedia*, Oxford: Routledge.
- Eells, E., Fitelson, B., 2000. “Measuring Confirmation and Evidence”, *Journal of Philosophy* 97, 663-672.
- Field, H., 1977. “Logic, Meaning, and Conceptual Role”, *Journal of Philosophy*, 74, 379-409.
- Fisher, R.A., 1922. “On the Mathematical Foundations of Theoretical Statistics”, *Philosophical Transactions of the Royal Society, series A* , 309-368.
- Fitelson, B., 1999. “The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity”, *Philosophy of Science* 66, S362-378.
- –––, 2001. “A Bayesian Account of Independent Evidence with Applications”, *Philosophy of Science*, 68, S123-S140.
- –––, 2002. “Putting the Irrelevance Back Into the Problem of Irrelevant Conjunction”, *Philosophy of Science* 69, 611-622.
- –––, 2005. ”Inductive Logic“, in J. Pfeifer and S. Sarkar (eds.), *The Philosophy of Science: An Encyclopedia*, Oxford: Routledge.
- –––, 2006. ”Logical Foundations of Evidential Support“, *Philosophy of Science* 73, 500-512.
- –––, 2007. ”Likelihoodism, Bayesianism, and Relational Confirmation“, *Synthese* 156, 473-489.
- Fitelson, B., Hawthorne, J., forthcoming. “How Bayesian Confirmation Theory Handles the Paradox of the Ravens”, Eells and Fetzer (eds.), *The Place of Probability in Science* Open Court. [Preprint available from the author (PDF)]
- Forster, M., Sober, E., 2004. “Why Likelihood”, in Taper and Lele (eds.), *The nature of Scientific Evidence*, Chicago: University of Chicago Press.
- Friedman, N. , Halpern, J.Y., 1995. “Plausibility Measures: A User's Guide”, in *Proc. UAI 95*, 175-184.
- Gaifman, H., Snir, M., 1982. “Probabilities Over Rich Languages, Testing and Randomness”, *Journal of Symbolic Logic* 47, 495-548.
- Gillies, D., 2000. *Philosophical Theories of Probability*, London: Routledge.

- Glymour, C., 1980. *Theory and Evidence*, Princeton: Princeton University Press.
- Goodman, N., 1983. *Fact, Fiction, and Forecast*, 4th ed., Cambridge, MA: Harvard University Press.
- Hacking, I., 1965. *Logic of Statistical Inference*, Cambridge: Cambridge University Press.
- –––, 1975. *The emergence of probability: a Philosophical Study*, Cambridge: Cambridge University Press.
- –––, 2001. *An Introduction to Probability and Inductive Logic*, Cambridge: Cambridge University Press.
- Hájek, A., 2003a. “What Conditional Probability Could Not Be”, *Synthese* 137, 273-323.
- –––, 2003b. “Interpretations of the Probability Calculus”, in the *Stanford Encyclopedia of Philosophy*, (Summer 2003 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2003/entries/probability-interpret/>
- –––, 2005. “Scotching Dutch Books?” *Philosophical Perspectives* 19 (Epistemology), 139-151.
- –––, 2007. “The Reference Class Problem is Your Problem Too”, *Synthese* 156, 563-585.
- Halpern, J. Y., 2003. *Reasoning About Uncertainty*, Cambridge, MA: MIT Press.
- Harper, W., 1976. “Rational Belief Change, Popper Functions and Counterfactuals”, in W. Harper and C. Hooker (eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, vol. I*, Dordrecht: Reidel, 73-115.
- Hawthorne, J., 1993. “Bayesian Induction *Is* Eliminative Induction”, *Philosophical Topics*, 21, 99-138.
- –––, 1994.“On the Nature of Bayesian Convergence”, *PSA 1994*, v. 1, 1994; pp. 241-249.
- –––, 2005. “Degree-of-Belief and Degree-of-Support: Why Bayesians Need Both Notions”, *Mind*, 114, 277-320.
- –––, 2009. “The Lockean Thesis and the Logic of Belief”, in F. Huber and C. Schmidt-Petri (eds.), *Degrees of Belief*, *Synthese Library*, Vol. 342, 2009, Springer, pp. 49-74.
- Hawthorne, J., Bovens, L,, 1999. “The Preface, the Lottery, and the Logic of Belief”, *Mind*, 108, 241-264.
- Hawthorne J., Fitelson, B., 2004. “Re-solving Irrelevant Conjunction With Probabilistic Independence”, *Philosophy of Science*, 71, 505-514.

- Hellman, G., 1997. “Bayes and Beyond”, *Philosophy of Science* 64, 191-221.
- Hempel, C., 1945. “Studies in the Logic of Confirmation”, *Mind* 54, 1-26, 97-121.
- Horwich, P., 1982. *Probability and Evidence*, Cambridge: Cambridge University Press.
- Howson, C., 1997. “A Logic of Induction”, *Philosophy of Science* 64, 268-290.
- –––, 2000. *Hume's Problem: Induction and the Justification of Belief*, Oxford: Oxford University Press.
- –––, 2002. “Bayesianism in Statistics“, in R. Swinburne (ed.), *Bayes's Theorem* The British Academy: Oxford University Press, 39-71.
- –––, 2007. “Logic With Numbers”, *Synthese* 156, 491-512.
- Howson, C., Urbach, P., 2005. *Scientific Reasoning: The Bayesian Approach*, 3rd edition, La Salle, IL: Open Court.
- Huber, F., 2005a. “Subjective Probabilities as Basis for Scientific Reasoning?” *British Journal for the Philosophy of Science* 56, 101-116.
- –––, 2005b. “What Is the Point of Confirmation?” *Philosophy of Science* 72, 1146-1159.
- Jaynes, E. T., 1968. “Prior Probabilities”, *I.E.E.E. Transactions on Systems Science and Cybernetics*, SSC-4, 227-241.
- Jeffrey, R. C., 1983. *The Logic of Decision, 2nd edition.* Chicago: University of Chicago Press.
- –––, 1987. “Alias Smith and Jones: The Testimony of the Senses”, *Erkenntnis* 26, 391-399.
- –––, 1992. *Probability and the Art of Judgment*, New York: Cambridge University Press.
- –––, 2004. *Subjective Probability: The Real Thing*, Cambridge: Cambridge University Press.
- Jeffreys, H., 1939. *Theory of Probability*, Oxford University Press.
- Joyce, J., 1998. “A Nonpragmatic Vindication of Probabilism”, *Philosophy of Science*, 65 (4) 575-603.
- –––, 1999. *The Foundations of Causal Decision Theory*, New York: Cambridge University Press.
- –––, 2003. “Bayes' Theorem”, in the *Stanford Encyclopedia of Philosophy*, (Summer 2003 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2003/entries/bayes-theorem/>
- –––, 2004. “Bayesianism”, in A. Mele and P. Rawling (eds.), *The Oxford Handbook of Rationality*, Oxford: Oxford University Press,
- –––, 2005. “How Degrees of Belief Reflect Evidence”, *Philosophical Perspectives* 19, 153-179.
- Kaplan, M., 1996. *Decision Theory as Philosophy*, Cambridge: Cambridge University Press.
- Kelly, K., Schulte, O., Juhl, C., 1997. “Learning Theory and the Philosophy of Science”, *Philosophy of Science* 64, 245-267.
- Keynes, J. M., 1921. *A Treatise on Probability*, Macmillan and Co.
- Kolmogorov, A. N., 1956. *Foundations of the Theory of Probability*, 2nd ed., New York: Chelsea Publishing Company.
- Koopman, B., 1940. “The Bases of Probability”, *Bulletin of the American Mathematical Society*, 46, 763-774. Reprinted in H. Kyburg and H. Smokler (eds.), 1980, *Studies in Subjective Probability*, 2nd ed. Huntington, N.Y.: Krieger Publ. Co.
- Kyburg, H. E., Jr., 1974. *The Logical Foundations of Statistical Inference*, Dordrecht: Reidel.
- –––, 1977. “Randomness and the Right Reference Class”, *Journal of Philosophy*, 74: 501-520.
- –––, 2006. “Belief, Evidence, and Conditioning”, *Philosophy of Science*, 73, 42-65.
- Lange, M., 1999. “Calibration and the Epistemological Role of Bayesian Conditionalization”, *Journal of Philosophy* 96, 294-324.
- –––, 2002. “Okasha on Inductive Scepticism”, *The Philosophical Quarterly*, 52, 226-232.
- Laudan, L., 1997. “How About Bust? Factoring Explanatory Power Back into Theory Evaluation”, *Philosophy of Science* 64, 206-216.
- Lenhard J., 2006. “Models and Statistical Inference: The Controversy Between Fisher and Neyman-Pearson“, *British Journal for the Philosophy of Science* 57: 69-91.
- Levi, I., 1967. *Gambling with Truth*, New York: Knopf.
- –––, 1977. “Direct Inference,” *Journal of Philosophy*, 74, 5-29.
- –––, 1978. “Confirmational Conditionalization”, *Journal of Philosophy*, 75, 730-737.
- –––, 1980. *The Enterprise of Knowledge*, Cambridge, Mass.: MIT Press.
- Lewis, D., 1980. “A Subjectivist's Guide to Objective Chance”, in Richard C. Jeffrey, (ed.), *Studies in Inductive Logic and Probability, vol. 2*, Berkeley: University of California Press, 263-293.
- Maher, P., 1993. *Betting on Theories*, Cambridge: Cambridge University Press.

- —–, 1996. “Subjective and Objective Confirmation”, *Philosophy of Science* 63, 149-174.
- —–, 1997. “Depragmatized Dutch Book Arguments”, *Philosophy of Science*, 64, 291-305.
- —–, 1999. “Inductive Logic and the Ravens Paradox”, *Philosophy of Science* 66, 50-70.
- —–, 2004. “Probability Captures the Logic of Scientific Confirmation”, in C. Hitchcock (ed.), *Contemporary Debates in Philosophy of Science*, Oxford: Blackwell, 69-93.
- —–, 2005. “Confirmation Theory”, *The Encyclopedia of Philosophy* 2nd ed., D. M. Borchert (ed.), Macmillan.
- —–, 2006a, “The Concept of Inductive Probability”, *Erkenntnis* 65,185-206.
- —–, 2006b. “A Conception of Inductive Logic”, *Philosophy of Science*, 73, 513-523.
- —–, forthcoming. “Bayesian Probability”, *Synthese*. [Preprint available from the author (PDF)]
- Mayo, D., 1996. *Error and the Growth of Experimental Knowledge*, Chicago: University of Chicago Press.
- —–, 1997. “Duhem's Problem, the Bayesian Way, and Error Statistics, or "What's Belief Got to do with It?"”, *Philosophy of Science* 64, 222-244.
- Mayo D., Spanos, A. 2006. “Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction“, *British Journal for the Philosophy of Science* 57, 323-357.
- McGee, V., 1994. “Learning the Impossible”, in E. Eells and B. Skyrms (eds.), *Probability and Conditionals*, New York: Cambridge University Press, 179-200.
- McGrew, T., 2003. “Confirmation, Heuristics, and Explanatory Reasoning”, *Brithish Journal for the Philosophy of Science* 54, 553-567.
- McGrew, L., McGrew, T., 2008. “Foundationalism, Probability, and Mutual Support”, *Erkenntnis* 68, 55-77.
- Neyman, J., Pearson E. S., 1967. *Joint Statistical Papers*, Cambridge: Cambridge University Press.
- Norton, J. D., 2003. “A Material Theory of Induction”, *Philosophy of Science*, 70, 647-670.
- —–, 2007. “Probability Disassembled”, *British Journal for the Philosophy of Science*, 58, 141-171.
- Okasha, S., 2001. “What Did Hume Really Show About Induction?”, *The Philosophical Quarterly* 51, 307-327.

- Popper, K., 1968. *The Logic of Scientific Discovery*, 3rd edition, London: Hutchinson.
- Quine, W.V., 1953. “Two Dogmas of Empiricism”, in *From a Logical Point of View*, New York: Harper Torchbooks.
  Routledge Encyclopedia of Philosophy, Version 1.0, London: Routledge
- Ramsey, F. P., 1926. “Truth and Probability”, in *Foundations of Mathematics and other Essays*, R. B. Braithwaite (ed.), Routledge & P. Kegan,1931, 156-198; reprinted in *Studies in Subjective Probability*, H. Kyburg and H. Smokler (eds.), 2nd ed., R. E. Krieger Publishing Company, 1980, 23-52; reprinted in *Philosophical Papers*, D. H. Mellor (ed.) Cambridge: University Press, Cambridge, 1990.
- Reichenbach, H., 1938. *Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge*, Chicago: University of Chicago Press.
- Renyi, A., 1970. *Foundations of Probability*, Holden-Day, Inc.
- Rosenkrantz, R.D., 1981. *Foundations and Applications of Inductive Probability*, Atascadero, CA: Ridgeview Publishing.
- Roush, S., 2004. “Positive Relevance Defended”, *Philosophy of Science* 71, 110-116.
- —–, 2005. “Problem of Induction”, in J. Pfeifer and S. Sarkar (eds.), *The Philosophy of Science: An Encyclopedia*, Oxford: Routledge.
- Royall, R. 1997. *Statistical Evidence: A Likelihood Paradigm*, New York: Chapman & Hall/CRC.
- Salmon, W., 1966. *The Foundations of Scientific Inference*, University of Pittsburgh Press.
- —–, 1975. “Confirmation and Relevance”, in H. Feigl and G. Maxwell (eds.), *Induction, Probability, and Confirmation, Minnesota Studies in the Philosophy of Science*, 6, Minneapolis: University of Minnesota Press, 3-36.
- Savage, L. J., 1954. *The Foundations of Statistics*, John Wiley (2nd ed., New York: Dover 1972).
- Savage, L.J., et al., 1962. *The Foundations of Statistical Inference*, London: Methuen.
- Schlesinger, G., 1991. *The Sweep of Probability*, Notre Dame: Notre Dame University Press.
- Seidenfeld, T., 1978, “Direct Inference and Inverse Inference”, *Journal of Philosophy* 75, 709-730.

- ——, 1992. "R.A. Fisher's Fiducial Argument and Bayes' Theorem", *Statistical Science* 7, 358-368.
- Shafer, G., 1976. *A Mathematical Theory of Evidence*, Princeton Univ. Press.: Princeton, NJ.
- ——, 1990. "Perspectives on the Theory and Practice of Belief Functions", *International Journal of Approximate Reasoning*, 3, 1-40.
- Skyrms, B., 1984. *Pragmatics and Empiricism*, New Haven: Yale University Press.
- ——, 1990. *The Dynamics of Rational Deliberation*, Cambridge, Mass.: Harvard University Press.
- ——, 2000. *Choice and Chance*, 4th edition. Wadsworth, Inc.
- Sober, E., 2002. "Bayesianism − Its Scope and Limits", in R. Swinburne (ed.), *Bayes's Theorem* The British Academy: Oxford University Press, 21-38.
- Spohn, W., 1988. "Ordinal Conditional Functions: A Dynamic Theory of Epistemic States." In W. Harper and B. Skyrms (eds.), *Causation in Decision, Belief Change, and Statistics*, vol. 2, 105-134. Dordrecht: Reidel.
- Strevens, M., 2004. "Bayesian Confirmation Theory: Inductive Logic, or Mere Inductive Framework?" *Synthese* 141, 365-379.
- Suppes, P., 2007. "Where do Bayesian Priors Come From?" *Synthese* 156, 441-471.
- Swinburne, R., 2002. *Bayes' Theorem*, Oxford: Oxford University Press.
- Talbot, W., 2001. "Bayesian Epistemology", in the *Stanford Encyclopedia of Philosophy* (Fall 2001 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2001/entries/epistemology-bayesian/>
- Teller, P., 1976. "Conditionalization, Observation, and Change of Preference", in W. Harper and C.A. Hooker (eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, Dordrecht: D. Reidel.
- Van Fraassen, B. C., 1983. "Calibration: A Frequency Justification for Personal Probability ", in R.S. Cohen and L. Laudan (eds.), *Physics, Philosophy, and Psychoanalysis: Essays in Honor of Adolf Grunbaum*, Dordrecht: Reidel.
- Venn, J., 1876. *The Logic of Chance*, 2nd ed., Macmillan and co; reprinted, New York, 1962.
- Vineberg, S., 2005. "Dutch Book Argument", in J. Pfeifer and S. Sarkar (eds.), *The Philosophy of Science: An Encyclopedia*, Oxford: Routledge.
- Vranas, P., 2004. "Hempel's Raven Paradox: A Lacuna in the Standard Bayesian Solution", *British Journal for the Philosophy of Science* 55, 545-560.

- Weatherson, B. 1999. "Begging the Question and Bayesianism", *Studies in History and Philosophy of Science* 30, 687-697.
- Williamson, J., 2007, "Inductive Influence", *British Journal for Philosophy of Science* 58, 689-708.
- Zadeh, L., 1965. "Fuzzy Sets", *Information and Control*, 8, 338-353.
- ——, 1978. "Fuzzy Sets as a Basis for a Theory of Possibility", *Fuzzy Sets and Systems*, vol. 1, 3-28.

Other Internet Resources

- Confirmation and Induction. Really nice overview by Franz Huber on the *Internet Encyclopedia of Philosophy*.
- Inductive Logic, (in PDF), by Branden Fitelson, *Philosophy of Science: An Encyclopedia*, (J. Pfeifer and S. Sarkar, eds.), Routledge. An extensive encyclopedia article on inductive logic.
- Teaching Theory of Knowledge: Probability and Induction. A very extensive outline of issues in Probability and Induction, each topic accompanied by a list of relevant books and articles (without links), compiled by Brad Armendt and Martin Curd.
- Probabilistic Confirmation Theory and Bayesian Reasoning. An annotated bibliography of influential works compiled by Timothy McGrew.
- Sherlock Holmes and Probabilistic Induction, by Soshichi Uchii (Philosophy and History of Science, University of Kyoto)
- Bayesian Networks Without Tears, (in PDF), by Eugene Charniak (Computer Science and Cognitive Science, Brown University). An introductory article on Bayesian inference.
- Miscellany of Works on Probabilistic Thinking. A collection of on-line articles on Subjective Probability and probabilistic reasoning by Richard Jeffrey and by several other philosophers writing on related issues.
- Fitelson's course on Confirmation Theory. Main page of Branden Fitelson's course on Confirmation Theory. The Syllabus provides an extensive list of links to readings. The Notes, Handouts, & Links page has Fitelson's weekly course notes and some links to useful internet resources on confirmation theory.
- Fitelson's course on Probability and Induction. Main page of Branden Fitelson's course on Probability and Induction. The Syllabus provides an extensive list of links to readings on the subject. The Notes & Handouts page has Fitelson's

powerpoint slides for each of his lectures and some links to handouts for the course. The Links page contains links to some useful internet resources.

Related Entries

Bayes' Theorem | epistemology: Bayesian | probability, interpretations of

Acknowledgments

Thanks to Jim Joyce and Edward Zalta for many valuable comments and suggestions. The editors and author would also like to thank Greg Stokley for carefully reading the entry and identifying several typographical errors.

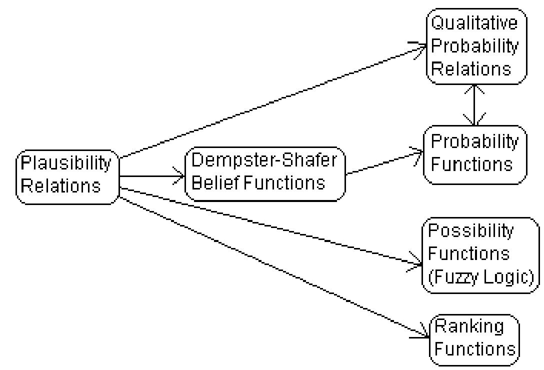
James Hawthorne <*hawthorne@ou.edu*>

Author & Citation Info | Friends PDF Preview | InPho Search | PhilPapers Bibliography

Supplement to Inductive Logic

Some Prominent Approaches to the Represention of Uncertain Inferences

The following figure indicates some relationships among six of the most prominent approaches. The arrows point from more general to less general representation schemes. For example, the *Dempster-Shafer* represention contains the *probability functions* as a special case.



Representations of Uncertainty

These representations are often described as measures on events, or states, or propositions, or sets of possibilities. But deductive logics are usually described in terms of statements or sentences of a language. So let's follow suit.

Plausibility relations (Friedman and Halpern, 1995) constitute the most general of these representations. They satisfy the weakest axioms, the weakest constraints on the logic of uncertainty. For a *plausibility relation* $\subseteq$ between sentences, an expression '$A \subseteq B$', says that A is no more plausible than B (i.e., B is at least as plausible as A, maybe more plausible). The axioms for plausibility relations say that tautologies are more plausible than contradictions, any two logically equivalent sentences are plausibility-related to other sentence in precisely the same way, a sentence is no more plausible than the sentences it logically entails, and the *no more plausible than* relation is transitive. These axioms make plausibility relations *weak partial orderings* on the relative plausibility of sentences. They permit some sentences to be incomparable — neither more plausible, nor less plausible, nor equally plausible to one another.

Qualitative probability relations are *plausibility relations* for which the ordering is *total* — i.e. any two sentences are either equally plausible, or one is more plausible than the other. This *total ordering* is established by one additional axiom. *Qualitative probability relations* also satisfy a second additional axiom that says that when a sentence S is logically incompatible with A and with B, then $A \subseteq B$ holds *just in case* (A or S) $\subseteq$ (B or S) holds as well. When *qualitative probability relations* are defined on a language with a rich enough vocabulary and satisfy one additional axiom, they can be shown to be *representable* by *probability functions* — i.e., given any *qualitative probability relation* $\subseteq$, there is a unique probability function P such that $A \subseteq B$ just in case $P[A] \leq P[B]$. So quantitative *probability* may be viewed as essentially just a way of placing a numerical measure on sentences that uniquely emulates the *is no more plausible than* relation specfied by *qualitative probability*. (See (Koopman, 1940), (Savage, 1954), (Hawthorne and Bovens, 1999), (Hawthorne, 2009).)

Probability (i.e., *quantitative probability*) is a measure of *plausibility* that assigns a number between 0 and 1 to each sentence. Intuitively, the *probability* of a sentence S, $P[S] = r$, says that *S is plausible to degree r*, or that *the rational degree of confidence (or belief) that S is true is r*. The axioms for *probabilities* basically require two things. First, tautologies get probability 1. Second, when A and B contradict each other, the probability of the disjunction (A or B) must be the sum of the probabilities of A and of B individually. It is primarily in regard to this second axiom that *probability* differs from each of the other quantitative measures of uncertainty.

Like *probability*, *Dempster-Shafer belief functions* (Shafer, 1976, 1990) measure *appropriate belief strengths* on a scale between 0 and 1, with contradictions and tautologies at the respective extremes. But whereas the *probability* of a disjunction of incompatible claims must equal the sum of the parts, *Dempster-Shafer belief functions* only require such disjunctions be *believed at least as strongly as* the sum of the *belief strengths* of the parts. So these functions are a generalization of *probability*. By simply tightening up the *Dempster-Shafer* axiom about how disjunctions are related to the their

parts we get back a restricted class of *Dempster-Shafer functions* that just is the class of *probability functions*. *Dempster-Shafer* functions are primarily employed as a logic of the evidential support for hypotheses. In that realm they are a generalization of the idea of evidential support embodied by *probabilistic inductive logic*. There is some controversy as to whether such a generalization is useful or desirable, or whether simple *probability* is too narrow to represent important evidential relationships captured by some *Dempster-Shafer functions*.

There is a sense in which the other two quantative measures of uncertainty, *possibility functions* and *ranking functions*, are definable in terms of formulas employing the *Dempster-Shafer functions*. But this is not the best way to understand them. *Possibility functions* (Zadeh, 1965, 1978), (Dubois and Prade, 1980, 1990) are generally read as representing *the degree of uncertainty* in a claim, where such uncertainty is often attributed to vaguness or fuzziness. These functions are formally like *probability functions* and *Dempster-Shafer functions*, but they subscribe to a simpler addition rule: the *degree of uncertainty* of a disjunction is the greater of the *degrees of uncertainty* of the parts. Similarly, the *degree of uncertainty* of a conjunction is the smaller of the *uncertainties* of the parts.

Ranking functions (Spohn, 1988) supply a measure of how surprising it would be if a claim turned out to be true, rated on a scale from 0 (not at all surprizing) to infinity. Tautologies have *rank* 0 and contradictions are infinitely surprizing. Logically equivalent claims have the same *rank*. The *rank* of a disjunction is equal to the *rank* of the lower ranking disjunct. These functions may be used to represent a kind of *order-of-magnitude* reasoning about the plausibility of various claims.

See (Halpern, 2003) for a good comparative treatment of all of these approaches.

Here are the axioms for the *Plausibility Relations* and the *Qualitative Probability Relations*.

Axioms for the *Plausibility Relations*
Each *plausibility relation* $\subseteq$ satisfies the following axioms:

1. if T is a tautology and K is a contradiction, it is not the case that $T \subseteq K$;
2. if A is logically equivalent to B and C is logically equivalent to D, and $A \subseteq C$, then $B \subseteq D$;

3. if A logically entails B, then $A \subseteq B$;
4. if $A \subseteq B$ and $B \subseteq C$, then $A \subseteq C$.

Two sentences are defined as *equally plausible*, $A = B$, just when $A \subseteq B$ and $B \subseteq A$. One sentence is defined as *less plausible* than another, $A \subset B$, just when $A \subseteq B$ but not $B \subseteq A$.

Axioms for the *Qualitative Probability Relations*
To get the *qualitative probability relations* we add the axioms

5. $A \subseteq B$ or $B \subseteq$ A;
6. if '(S and A)' and '(S and B)' are both logical contradictions, then $A \subseteq B$ just in case (A or S) $\subseteq$ (B or S).

The typical axioms for *quantitative probability* are as follows:

i. for all sentences S, $0 \leq P[S] \leq 1$;
ii. if S is a tautology, then $P[S] = 1$;
iii. if '(A and B)' is a logical contradiction, then $P[A \text{ or } B] = P[A] + P[B]$.

Axioms 1-6 for the *qualitative probability relations* are probabilistically sound with respect to the quantitative probability functions. That is, for each given probability function P, define a relation $\subseteq$ such that $A \subseteq B$ *just in case* $P[A] \leq P[B]$. Then $\subseteq$ must satisfy axioms 1-6. However, not every qualitative probability relation that satisfies axioms 1-6 may be represented by a probability function. To get that we must add one further axiom.

Let's say that a qualitative probability relation $\subseteq$ is *fine-grained* just in case it satisfies the following axiom:

(7) if $A \subset B$, then there is some tautology consisting of n sentences, (S_1 or S_2 or … or S_n), where each distinct S_i and S_j are inconsistent with one another, such that for each of the S_i, (A or S_i) $\subset B$.

For each *fine-grained* qualitative probability relation $\subseteq$ there is a unique probability function P such that $A \subseteq B$ *just in case* $P[A] \leq P[B]$.

Now call a qualitative probability relation $\subseteq$ *properly extendable* just in case it can be extended to a *fine-grained* qualitative probability relation defined on a larger language (i.e., a language containing additional sentences). Then for every *properly extendable* qualitative probability relation $\subseteq$ there is a probability function P such that $A \subseteq B$ *just in case* $P[A] \leq P[B]$. In general a given *properly extendable* qualitative probability relation may have many such representing probability functions, corresponding to different ways of extending it to *fine-grained* qualitative probability relations.

Thus, the quantitative probability functions may be viewed as just useful ways of representing *properly extendable* qualitiative probability relations on a convenient numerical scale.

[Back to Text]

James Hawthorne <*hawthorne@ou.edu*>

Author & Citation Info | Friends PDF Preview |InPho Search |PhilPapers Bibliography

Supplement to Inductive Logic

Tighter Bounds on the Margin of Error

If we want strong support for hypotheses claiming more than 99.9% of all ravens are black, the following extension of Table 1 applies.

Table 1.2: Values of lower bound p on the posterior probability

$m/n = 1$ $F[A,B] > .999$	Sample-Size = n (number of As in Sample of Bs = $m = n$)						
Prior Ratio: K ↓	400	800	1600	3200	6400	12800	25600
1	0.3305	0.5513	0.7985	0.9593	0.9983	1.0000	1.0000
2	0.1980	0.3805	0.6645	0.9219	0.9967	1.0000	1.0000
5	0.0899	0.1973	0.4421	0.8252	0.9918	1.0000	1.0000
10	0.0470	0.1094	0.2838	0.7023	0.9837	1.0000	1.0000
100	0.0049	0.0121	0.0381	0.1909	0.8578	0.9997	1.0000
1,000	0.0005	0.0012	0.0039	0.0231	0.3763	0.9973	1.0000
10,000	0.0000	0.0001	0.0004	0.0024	0.0569	0.9733	1.0000
100,000	0.0000	0.0000	0.0000	0.0002	0.0060	0.7849	1.0000
1,000,000	0.0000	0.0000	0.0000	0.0000	0.0006	0.2674	1.0000
10,000,000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0352	0.9999

$P_{\alpha}[F[A,B] > .999 \mid b \cdot F[A,S]=1 \cdot \text{Random}[S,B,A] \cdot \text{Size}[S]=n] \geq p$, for a range of Sample-Sizes n (from 400 to 25600), when the prior probability of any specific frequency hypothesis outside the region between .999 and 1 is no more than K times more than the lowest prior probability for any specific frequency hypothesis inside of the region between .999 and 1.

The lower right corner of the table shows that even when the*vagueness* or *diversity* sets include support functions with prior plausibilities up to *ten million* times higher for hypotheses asserting frequency values below .999 than for hypotheses making frequency claims between .999 and 1, a sample of 25600 black ravens will,

nevertheless, pull the posterior plausibility above .9999 that “the true frequency is over .999” for every support function in the set.

[Back to Text]

Author & Citation Info | Friends PDF Preview |InPho Search |PhilPapers Bibliography

Supplement to Inductive Logic

Immediate Consequences of the Independent Evidence Conditions

When neither *independence condition* holds, we at least have:

$$
\begin{aligned}
P[e^n \mid h_j \cdot b \cdot c^n] &= P[e_n \mid h_j \cdot b \cdot c^n \cdot e^{n-1}] \times P[e^{n-1} \mid h_j \cdot b \cdot c^n] \\
&= \ldots \\
&= \prod_{k=1}^{n} P[e_k \mid h_j \cdot b \cdot c^n \cdot e^{k-1}]
\end{aligned}
$$

When *condition-independence* holds we have:

$$
\begin{aligned}
P[e^n \mid h_j \cdot b \cdot c^n] &= P[e_n \mid h_j \cdot b \cdot c_n \cdot (c^{n-1} \cdot e^{n-1})] \times P[e^{n-1} \mid h_j \cdot b \cdot c_n \cdot c^{n-1}] \\
&= P[e_n \mid h_j \cdot b \cdot c_n \cdot (c^{n-1} \cdot e^{n-1})] \times P[e^{n-1} \mid h_j \cdot b \cdot c^{n-1}] \\
&= \ldots \\
&= \prod_{k=1}^{n} P[e_k \mid h_j \cdot b \cdot c_k \cdot (c^{k-1} \cdot e^{k-1})]
\end{aligned}
$$

If we add *result-independence* to*condition-independence*, the occurrences of‘$(c^{k-1} \cdot e^{k-1})$’may be removed from the previous formula, giving:

$$
P[e^n \mid h_j \cdot b \cdot c^n] = \prod_{k=1}^{n} P[e_k \mid h_j \cdot b \cdot c_k]
$$

[Back to Text]

Author & Citation Info | Friends PDF Preview |InPho Search |PhilPapers Bibliography

Supplement to Inductive Logic

Proof of the Falsification Theorem

Likelihood Ratio Convergence Theorem 1—The Falsification Theorem:
Suppose the evidence stream c^n contains precisely m experiments or observations on which h_j is*not fully outcome-compatible* with h_i. And suppose that the *Independent Evidence Conditions* hold for evidence stream c^n with respect to each of these hypotheses. Furthermore, suppose there is a lower bound $\delta > 0$ such that for each c_k on which h_j is*not fully outcome-compatible* with h_i,
$P[\vee\{o_{ku} : P[o_{ku} \mid h_j{\cdot}b{\cdot}c_k] = 0\} \mid h_i{\cdot}b{\cdot}c_k] \geq \delta$ — i.e. h_i (together with $b{\cdot}c_k$) *says*, via a likelihood with value no smaller than δ, that one of the outcomes will occur that h_j *says* cannot occur). Then,

$$P[\vee\{e^n : P[e^n \mid h_j{\cdot}b{\cdot}c^n]/P[e^n \mid h_i{\cdot}b{\cdot}c^n] = 0\} \mid h_i{\cdot}b{\cdot}c^n]$$

$$= P[\vee\{e^n : P[e^n \mid h_j{\cdot}b{\cdot}c^n] = 0\} \mid h_i{\cdot}b{\cdot}c^n]$$

$$\geq 1-(1-\delta)^m,$$

which approaches 1 for large m.

Proof

First notice that according to the supposition of the theorem, for each of the m experiments or observations c_k on which h_j is *not fully outcome-compatible* with h_i we have

$$(1-\delta) \geq P[\vee\{o_{ku} : P[o_{ku} \mid h_j{\cdot}b{\cdot}c_k] > 0\} \mid h_i{\cdot}b{\cdot}c_k]$$

$$= \sum_{\{o_{ku}\in O_k:\, P[o_{ku} \mid h_j\cdot b\cdot c_k] > 0\}} P[o_{ku} \mid h_i{\cdot}b{\cdot}c_k].$$

And for each of the other c_k in the evidence stream c^n — i.e. for each of the c_k on which h_j is *fully outcome-compatible* with h_i,

$$P[\vee\{o_{ku} : P[o_{ku} \mid h_j{\cdot}b{\cdot}c_k] > 0\} \mid h_i{\cdot}b{\cdot}c_k] = 1.$$

Then, we may iteratively decompose $P[\vee\{e^n : P[e^n \mid h_j{\cdot}b{\cdot}c^n] > 0\} \mid h_i{\cdot}b{\cdot}c^n]$ into it's components as follows:

$$P[\vee\{e^n : P[e^n \mid h_j{\cdot}b{\cdot}c^n] > 0\} \mid h_i{\cdot}b{\cdot}c^n]$$

$$= \sum_{\{e^n: P[e^n \mid h_j\cdot b\cdot c^n] > 0\}} P[e^n \mid h_i{\cdot}b{\cdot}c^n]$$

$$= \sum_{\{e^n: P[e_n \mid h_j\cdot b\cdot c_n\cdot c^{n-1}\cdot e^{n-1}] \times P[e^{n-1} \mid h_i\cdot b\cdot c_n\cdot c^{n-1}] > 0\}} P[e_n \mid h_j{\cdot}b{\cdot}c_n{\cdot}c^{n-1}{\cdot}e^{n-1}] \times P[e^{n-1} \mid h_i{\cdot}b{\cdot}c_n{\cdot}c^{n-1}]$$

$$= \sum_{\{e^n: P[e_n \mid h_j\cdot b\cdot c_n] \times P[e^{n-1} \mid h_i\cdot b\cdot c^{n-1}] > 0\}} P[e_n \mid h_j{\cdot}b{\cdot}c_n] \times P[e^{n-1} \mid h_i{\cdot}b{\cdot}c^{n-1}]$$

$$= \sum_{\{e^n: P[e_n \mid h_j\cdot b\cdot c_n] > 0 \,\&\, P[e^{n-1} \mid h_i\cdot b\cdot c^{n-1}] > 0\}} P[e_n \mid h_j{\cdot}b{\cdot}c_n] \times P[e^{n-1} \mid h_i{\cdot}b{\cdot}c^{n-1}]$$

$$= \sum_{\{e^{n-1}:\, P[e^{n-1} \mid h_j\cdot b\cdot c^{n-1}] > 0\}} \sum_{\{o_{nu}\in O_n: P[o_{nu} \mid h_j\cdot b\cdot c_n] > 0\}} P[o_{nu} \mid h_i{\cdot}b{\cdot}c_n] \times P[e^{n-1} \mid h_i{\cdot}b{\cdot}c^{n-1}]$$

$$= \sum_{\{e^{n-1}:\, P[e^{n-1} \mid h_j\cdot b\cdot c^{n-1}] > 0\}} P[\vee\{o_{nu}: P[o_{nu} \mid h_j{\cdot}b{\cdot}c_n] > 0\} \mid h_i{\cdot}b{\cdot}c_n] \times P[e^{n-1} \mid h_i{\cdot}b{\cdot}c^{n-1}]$$

$$\leq (1-\gamma) \times \sum_{\{e^{n-1}:\, P[e^{n-1} \mid h_j\cdot b\cdot c^{n-1}] > 0\}} P[e^{n-1} \mid h_i{\cdot}b{\cdot}c^{n-1}],$$

if c_n is an observation on which h_j is *not fully outcome-compatible* with h_i

or

$$= \sum_{\{e^{n-1}:\, P[e^{n-1} \mid h_j\cdot b\cdot c^{n-1}] > 0\}} P[e^{n-1} \mid h_i{\cdot}b{\cdot}c^{n-1}],$$

if c_n is an observation on which h_j is *fully outcome-compatible* with h_i

…

continuining this process of decomposing terms of form$\sum_{\{e^k:\, P[e^k \mid h_j\cdot b\cdot c^k] > 0\}} P[e^k \mid h_i{\cdot}b{\cdot}c^k]$ (in each disjunct of the 'or' above, using the same decomposition process shown in the six lines preceding that disjunction), and realizing that according to the supposition of the theorem, this decomposition leads to terms of the form of the first disjunct exactly m times, we get

…

$$\leq (1-\gamma) = (1-\gamma)^m.$$

$$\prod_{k=1}^{m}$$

So,

$$\begin{aligned} &P[\vee\{e^n : P[e^n \mid h_j\cdot b\cdot c^n] = 0\} \mid h_i\cdot b\cdot c^n] \\ &\quad = 1 - P[\vee\{e^n : P[e^n \mid h_j\cdot b\cdot c^n] > 0\} \mid h_i\cdot b\cdot c^n] \geq 1-(1-\gamma)^m. \end{aligned}$$

We also have,

$$\begin{aligned} &P[\vee\{e^n : P[e^n \mid h_j\cdot b\cdot c^n]/P[e^n \mid h_i\cdot b\cdot c^n] = 0\} \mid h_i\cdot b\cdot c^n] \\ &\quad = P[\vee\{e^n : P[e^n \mid h_j\cdot b\cdot c^n] = 0\} \mid h_i\cdot b\cdot c^n], \end{aligned}$$

because

$$\begin{aligned} &P[\vee\{e^n : P[e^n \mid h_j\cdot b\cdot c^n]/P[e^n \mid h_i\cdot b\cdot c^n] > 0\} \mid h_i\cdot b\cdot c^n] \\ &\quad = \textstyle\sum_{\{e^n:\, P[e^n \mid h_j\cdot b\cdot c^n]/P[e^n \mid h_i\cdot b\cdot c^n] > 0\}} P[e^n \mid h_i\cdot b\cdot c^n] \\ &\quad = \textstyle\sum_{\{e^n:\, P[e^n \mid h_j\cdot b\cdot c^n] > 0\ \&\ P[e^n \mid h_i\cdot b\cdot c^n] > 0\}} P[e^n \mid h_i\cdot b\cdot c^n] \\ &\quad = \textstyle\sum_{\{e^n:\, P[e^n \mid h_j\cdot b\cdot c^n] > 0\}} P[e^n \mid h_i\cdot b\cdot c^n] \\ &\quad = P[\vee\{e^n : P[e^n \mid h_j\cdot b\cdot c^n] > 0\} \mid h_i\cdot b\cdot c^n]. \end{aligned}$$

[Back to Text]

Author & Citation Info | Friends PDF Preview |InPho Search |PhilPapers Bibliography

Supplement to Inductive Logic

Proof that the EQI for c^n is the sum of EQI for the individual c_k

Theorem: The EQI Decomposition Theorem:

When the *Independent Evidence Conditions* are satisfied,

$$\mathrm{EQI}[c^n \mid h_i/h_j \mid b] = \sum_{k=1}^{n} \mathrm{EQI}[c_k \mid h_i/h_j \mid b].$$

[Back to Text]

Proof:

$$\begin{aligned} &\mathrm{EQI}[c^n \mid h_i/h_j \mid b] \\ &= \textstyle\sum_{\{e^n\}} \mathrm{QI}[e^n \mid h_i/h_j \mid b\cdot c^n] \times P[e^n \mid h_i\cdot b\cdot c^n] \\ &= \textstyle\sum_{\{e^n\}} \log[P[e^n \mid h_i\cdot b\cdot c^n]/P[e^n \mid h_j\cdot b\cdot c^n]] \times P[e^n \mid h_i\cdot b\cdot c^n] \\ &= \textstyle\sum_{\{e^{n-1}\}} \sum_{\{e_n\}} (\log[P[e_n \mid h_i\cdot b\cdot c_n\cdot(c^{n-1}\cdot e^{n-1})]/P[e_n \mid h_j\cdot b\cdot c_n\cdot(c^{n-1}\cdot e^{n-1})]] \\ &\qquad + \log[P[e^{n-1} \mid h_i\cdot b\cdot c_n\cdot c^{n-1}]/P[e^{n-1} \mid h_j\cdot b\cdot c_n\cdot c^{n-1}]]) \times \\ &\qquad P[e_n \mid h_i\cdot b\cdot c_n\cdot(c^{n-1}\cdot e^{n-1})] \times P[e^{n-1} \mid h_i\cdot b\cdot c_n\cdot c^{n-1}] \\ &= \textstyle\sum_{\{e^{n-1}\}} \sum_{\{e_n\}} (\log[P[e_n \mid h_i\cdot b\cdot c_n]/P[e_n \mid h_j\cdot b\cdot c_n]] \\ &\qquad + \log[P[e^{n-1} \mid h_i\cdot b\cdot c^{n-1}]/P[e^{n-1} \mid h_j\cdot b\cdot c^{n-1}]]) \times \\ &\qquad P[e_n \mid h_i\cdot b\cdot c_n] \times P[e^{n-1} \mid h_i\cdot b\cdot c^{n-1}] \\ &= (\textstyle\sum_{\{e_n\}} \log[P[e_n \mid h_i\cdot b\cdot c_n]/P[e_n \mid h_j\cdot b\cdot c_n]] \times P[e_n \mid h_i\cdot b\cdot c_n] \times \\ &\qquad \textstyle\sum_{\{e^{n-1}\}} P[e^{n-1} \mid h_i\cdot b\cdot c^{n-1}]) \\ &\quad + (\textstyle\sum_{\{e^{n-1}\}} \log[P[e^{n-1} \mid h_i\cdot b\cdot c^{n-1}]/P[e^{n-1} \mid h_j\cdot b\cdot c^{n-1}]] \times P[e^{n-1} \mid h_i\cdot b\cdot c^{n-1}] \times \\ &\qquad \textstyle\sum_{\{e_n\}} P[e_n \mid h_i\cdot b\cdot c_n]) \end{aligned}$$

$= \text{EQI}[c_n \mid h_i/h_j \mid b] + \text{EQI}[c^{n-1} \mid h_i/h_j \mid b]$

$= \ldots$ (iterating this decomposition process)

$= \sum_{k=1}^{n} \text{EQI}[c_k \mid h_i/h_j \mid b].$

James Hawthorne<*hawthorne@ou.edu*>

Author & Citation Info | Friends PDF Preview |InPho Search |PhilPapers Bibliography

Supplement to Inductive Logic

The Effect on EQI of Partitioning the Outcome Space More Finely — Including Proof of the Nonnegativity of EQI

Given some experiment or observation (or series of them) c, is there any special advantage to parsing the space of possible outcomes O into more alternatives rather than fewer alternatives? Couldn't we do as well at evidentially evaluating hypotheses by parsing the space of outcomes into just a few alternatives — e.g., one possible outcome that h_i says is very likely and h_j says is rather unlikely, one that h_i says is rather unlikely and h_j says is very likely, and perhaps a third outcome on which h_i and h_j pretty much agree? The answer is "*No !*". Parsing the space of outcomes into a larger number of empirically distinct possible outcomes always provides a better measure of evidential support.

To see this intuitively, suppose some outcome description o can be parsed into two distinct outcome descriptions, o_1 and o_2, where o is equivalent to $(o_1 \vee o_2)$, and suppose that h_i differs from h_j much more on the likelihood of o_1 than on the likelihood of o_2. Then, intuitively, when o is found to be true, whichever of the more precise descriptions, o_1 or o_2, is true should make a difference as to how strong the comparative support for the two hypotheses turns out to be. Reporting whichever of o_1 or o_2 occurs will be more informative than simply reporting o. That is, if the outcome of the experiment is only described as o, relevant information is lost.

It turns out that EQI measures how well possible outcomes can distinguish between hypotheses in a way that reflects the intuition that a finer partition of the possible outcomes is more informative. The numerical value of EQI is always made larger by parsing the outcome space more finely, provided that the likelihoods for outcomes in the finer parsing differ at least a bit from some of the likelihoods for outcomes of the less refined parsing. This is important for our main convergence result because in that theorem we want the average value of EQI for the whole sequence of experiments and observations to be positive, and the larger the better.

The following **Partition Theorem** implies the **Nonnegativity of EQI** result as well. It show that each EQI[c_k | h_i/h_j | b] must be non-negative; and it will be positive *just in case* for at least one possible outcome o_{ku}, $P[o_{ku} \mid h_j \cdot b \cdot c_k] \neq P[o_{ku} \mid h_i \cdot b \cdot c_k]$. This theorem will also show that EQI[c_k | h_i/h_j |b] generally becomes larger whenever the outcome space is partitioned more finely. It follows immediately that the average value of EQI for a sequence of experiments or observations, $\overline{\text{EQI}}[c^n \mid h_i/h_j \mid b]$, averaged over the sequence of observations c^n, is non-negative, and must be positive if for even one of the c_k that contribute to it, at least one possible outcome o_{ku} distinguishes between the two hypotheses by making $P[o_{ku} \mid h_j \cdot b \cdot c_k] \neq P[o_{ku} \mid h_i \cdot b \cdot c_k]$.

Partition Theorem:
For any positive real numbers r_1, r_2,s_1, s_2:

(1) if $r_1/s_1 > (r_1+r_2)/(s_1+s_2)$, then
$(r_1+r_2)\times\log[(r_1+r_2)/(s_1+s_2)] < r_1\times\log[r_1/s_1] + r_2\times\log[r_2/s_2]$;

and

(2) if $r_1/s_1 = (r_1+r_2)/(s_1+s_2)$, then
$r_1\times\log[r_1/s_1] + r_2\times\log[r_2/s_2] = (r_1+r_2)\times\log[(r_1+r_2)/(s_1+s_2)]$.

To prove this theorem first notice that

$$\begin{aligned} r_1/s_1 = (r_1+r_2)/(s_1+s_2) &\text{ iff } r_1s_1 + r_1s_2 = s_1r_1 + s_1r_2 \\ &\text{ iff } r_1/s_1 = r_2/s_2. \end{aligned}$$

We'll draw on this little result immediately below. It is clearly relevant to the antecedent of case (2) of the theorem we want to prove.

We establish case (2) first. Suppose the antecedent of case (2) holds. Then, from the little result just proved, we have

$$\begin{aligned} & r_1 \log[r_1/s_1] + r_2 \log[r_2/s_2] \\ &= r_1 \log[(r_1+r_2)/(s_1+s_2)] + r_2 \log[(r_1+r_2)/(s_1+s_2)] \\ &= (r_1 + r_2) \log[(r_1+r_2)/(s_1+s_2)]. \end{aligned}$$

That establishes case (2).

To get case (1), consider the following function of p:

$f(p) = p \log[p/u] + (1-p) \log[(1-p)/v]$,
where we only assume that $u > 0$, $v > 0$, and $0 < p < 1$.

This function has its minimum value when $p = u/(u+v)$. (This is easily verified by setting the derivative of $f(p)$ with respect to p equal to 0 to find the minimum value of $f(p)$; and it is easy to verified that this is a minimum rather than a maximum value.) At this minimum, where $p = u/(u+v)$, we have

$$\begin{aligned} f(p) &= -u/(u+v) \log[u+v] - v/(u+v) \log[u+v] \\ &= -\log[u+v]. \end{aligned}$$

Thus, for all values of p other than $u/(u+v)$,

$$\begin{aligned} -\log[u+v] &< f(p) \\ &= p \log[p/u] + (1-p) \log[(1-p)/v]. \end{aligned}$$

That is, if $p \neq u/(u+v)$, $-\log[u+v] < p \log[p/u] + (1-p) \log[(1-p)/v]$.

Now, let $p = r_1/(r_1+r_2)$, let $u = s_1/(r_1+r_2)$, and let $v = s_2/(r_1+r_2)$. Plugging into the previous formula, and multiplying both sides by (r_1+r_2), we get:

if
$r_1/(r_1+r_2) \neq s_1/(s_1+s_2)$ (i.e., equivalently, if$r_1/s_1 \neq (r_1+r_2)/(s_1+s_2)$),
then
$\log[(r_1+r_2)/(s_1+s_2)] < [r_1/(r_1+r_2)] \log[r_1/s_1] + (1-[r_1/(r_1+r_2)]) \log[r_2/s_2]$
(i.e. equivalently, $(r_1+r_2) \log[(r_1+r_2)/(s_1+s_2)] < r_1 \log[r_1/s_1] + r_2\log[r_2/s_2]$).

Thus, from the two equivalents, we've proved case 2:
if
$r_1/s_1 \neq (r_1+r_2)/(s_1+s_2)$),

then
$(r_1+r_2)\log[(r_1+r_2)/(s_1+s_2)] < r_1\log[r_1/s_1] + r_2\log[r_2/s_2])$.

This completes the proof of the theorem.

To apply this result to $\mathrm{EQI}[c_k \mid h_i/h_j \mid b]$ recall that

$$\begin{aligned}&\mathrm{EQI}[c_k \mid h_i/h_j \mid b]\\ &\quad = \textstyle\sum_{\{u:\, P[o_{ku} \mid h_j\cdot b\cdot c_k] > 0\}} \log[P[o_{ku} \mid h_i\cdot b\cdot c_k]/P[o_{ku} \mid h_j\cdot b\cdot c_k]]\\ &\qquad\qquad \times P[o_{ku} \mid h_i\cdot b\cdot c_k].\end{aligned}$$

Suppose c_k has m alternative outcomes o_{ku} on which both

$$P[o_{ku} \mid h_j\cdot b\cdot c_k] > 0 \ \text{ and } \ P[o_{ku} \mid h_i\cdot b\cdot c_k] > 0.$$

Let's label their likelihoods relative to h_i (i.e., their likelihoods $P[o_{ku} \mid h_i\cdot b\cdot c_k]$) as $r_1, r_2, \ldots, r_m$. And let's label their likelihoods relative to h_j as $s_1, s_2, \ldots, s_m$. In terms of this notation,

$$\mathrm{EQI}[c_k \mid h_i/h_j \mid b] = \sum_{u=1}^{m} r_u\times\log[r_u/s_u].$$

Notice also that $(r_1+r_2+r_3+\ldots+r_m) = 1$ and $(s_1+s_2+s_3+\ldots+s_m) = 1$.

Now, think of $\mathrm{EQI}[c_k \mid h_i/h_j \mid b]$ as generated by applying the theorem in successive steps:

$$\begin{aligned}0 &= 1\times\log[1/1]\\ &= (r_1+r_2+r_3+\ldots+r_m)\times\log[(r_1+r_2+r_3+\ldots+r_m)/(s_1+s_2+s_3+\ldots+s_m)]\\ &\le r_1\times\log[r_1/s_1] + (r_2+r_3+\ldots+r_m)\times\log[(r_2+r_3+\ldots+r_m)/(s_2+s_3+\ldots+s_m)]\\ &\le r_1\times\log[r_1/s_1] + r_2\times\log[r_2/s_2] + (r_3+\ldots+r_m)\times\log[(r_3+\ldots+r_m)/(s_3+\ldots+s_m)]\\ &\le \ldots\\ &\le \sum_{u=1}^{m} r_u\times\log[r_u/s_u]\\ &= \mathrm{EQI}[c_k \mid h_i/h_j \mid b].\end{aligned}$$

The theorem also says that *at each step* equality holds just in case

$$r_u/s_u = (r_u+r_{u+1}+\ldots+r_m)/(s_u+s_{u+1}+\ldots+s_m),$$

which itself holds just in case

$$r_u/s_u = (r_{u+1}+\ldots+r_m)/(s_{u+1}+\ldots+s_m).$$

So,

$$\mathrm{EQI}[c_k \mid h_i/h_j \mid b] = 0$$

just in case

$$\begin{aligned}1 &= (r_1+r_2+r_3+\ldots+r_m)/(s_1+s_2+s_3+\ldots+s_m)\\ &= r_1/s_1\\ &= (r_2+r_3+\ldots+r_m)/(s_2+s_3+\ldots+s_m)\\ &= r_2/s_2\\ &= (r_3+\ldots+r_m)/(s_3+\ldots+s_m)\\ &= r_3/s_3\\ &= \ldots\\ &= r_m/s_m.\end{aligned}$$

That is,

$$\mathrm{EQI}[c_k \mid h_i/h_j \mid b] = 0$$

just in case for all o_{ku} such that $P[o_{ku} \mid h_j \cdot b] > 0$ and $P[o_{ku} \mid h_i \cdot b] > 0$,

$$P[o_{ku} \mid h_i \cdot b \cdot c_k]/P[o_{ku} \mid h_j \cdot b \cdot c_k] = 1.$$

Otherwise,

$$\mathrm{EQI}[c_k \mid h_i/h_j \mid b] > 0;$$

and for each successive step in partitioning the outcome space to generate $\mathrm{EQI}[c_k \mid h_i/h_j \mid b]$, if

$$r_u/s_u \neq (r_u+r_{u+1}+\ldots+r_m)/(s_u+s_{u+1}+\ldots+s_m),$$

we have the strict inequality:

$$(r_u+r_{u+1}+\ldots+r_m) \times \log[(r_u+r_{u+1}+\ldots+r_m)/(s_u+s_{u+1}+\ldots+s_m)] < r_u \times \log[r_u/s_u] + (r_{u+1}+\ldots+r_m) \times \log[(r_{u+1}+\ldots+r_m)/(s_{u+1}+\ldots+s_m)].$$

So each such division of $(o_{ku} \vee o_{ku+1} \vee \ldots \vee o_{km})$ into two separate tatements, o_{ku} and $(o_{ku+1} \vee \ldots \vee o_{km})$, adds a strictly positive contribution to the size of $\mathrm{EQI}[c_k \mid h_i/h_j \mid b]$ *just when* $P[o_{ku} \mid h_i \cdot b \cdot c_k] / P[o_{ku} \mid h_j \cdot b \cdot c_k] \neq P[(o_{ku+1} \vee \ldots \vee o_{km}) \mid h_i \cdot b \cdot c_k] / P[(o_{ku+1} \vee \ldots \vee o_{km}) \mid h_j \cdot b \cdot c_k]$.

[Back to Text]

Author & Citation Info | Friends PDF Preview |InPho Search |PhilPapers Bibliography

Supplement to Inductive Logic

Proof of the Non-Falsifying Refutation Theorem

The proof of Convergence Theorem 2 requires the introduction of one more concept, that of the *variance in the quality of information* for a sequence of experiments or observations, $\mathrm{VQI}[c^n \mid h_i/h_j \mid b]$. The quality of the information QI from a specific outcome sequence e^n may vary somewhat from the *expectedquality of information* for conditions c^n. A common statistical measure of how widely individual values tend to vary from an expected value is given by the *expected squared distance from the expected value*, which is called the *variance*.

> **Definition: VQI — the Variance in the Quality of Information.**
> For h_j outcome-compatible with h_i on c_k, define
> $\mathrm{VQI}[c_k \mid h_i/h_j \mid b] =$
> $$\sum_u (\mathrm{QI}[o_{ku} \mid h_i/h_j \mid b \cdot c_k] - \mathrm{EQI}[c_k \mid h_i/h_j \mid b])^2 \times P[o_{ku} \mid h_i \cdot b \cdot c_k].$$
> For a sequence c^n of observations on which h_j is outcome-compatible with h_i, define
> $\mathrm{VQI}[c^n \mid h_i/h_j \mid b] =$
> $$\sum_{\{e^n\}} (\mathrm{QI}[e^n \mid h_i/h_j \mid b \cdot c^n] - \mathrm{EQI}[c^n \mid h_i/h_j \mid b])^2 \times P[e^n \mid h_i \cdot b \cdot c^n].$$

Clearly VQI will be positive unless h_i and h_j agree on the likelihoods of all possible outcome sequences in the evidence stream, in which case both $\mathrm{EQI}[c^n \mid h_i/h_j \mid b]$ and $\mathrm{VQI}[c^n \mid h_i/h_j \mid b]$ equal 0.

When both *Independent Evidence Conditions* hold, $\mathrm{VQI}[c^n \mid h_i/h_j \mid b]$ decompose into the sum of the VQI for individual experiments or observations c_k.

Theorem: The VQI Decomposition Theorem for Independent Evidence on Each Hypothesis:
Suppose both *condition independence*and *result-independence* hold. Then

$$\mathrm{VQI}[c^n \mid h_i/h_j \mid b] = \sum_{k=1}^{n} \mathrm{VQI}[c_k \mid h_i/h_j \mid b].$$

For the Proof, we employ the following abbreviations:

$$\begin{aligned}
\mathrm{Q}[e_k] &= \mathrm{QI}[e_k \mid h_i/h_j \mid b\cdot c_k] \\
\mathrm{Q}[e^k] &= \mathrm{QI}[e^k \mid h_i/h_j \mid b\cdot c^k] \\
\mathrm{E}[c_k] &= \mathrm{EQI}[c_k \mid h_i/h_j \mid b] \\
\mathrm{E}[c^k] &= \mathrm{EQI}[c^k \mid h_i/h_j \mid b] \\
\mathrm{V}[c_k] &= \mathrm{VQI}[c_k \mid h_i/h_j \mid b] \\
\mathrm{V}[c^k] &= \mathrm{VQI}[c^k \mid h_i/h_j \mid b]
\end{aligned}$$

The equation stated by the theorem may be derived as follows:

$$\begin{aligned}
&\mathrm{V}[c^n] \\
&= \textstyle\sum_{\{e^n\}} (\mathrm{Q}[e^n] - \mathrm{E}[c^n])^2 \times P[e^n \mid h_i\cdot b\cdot c^n] \\
&= \textstyle\sum_{\{e^n\}} ((\mathrm{Q}[e_n]+\mathrm{Q}[e^{n-1}]) - (\mathrm{E}[c_n]+\mathrm{E}[c^{n-1}]))^2 \\
&\qquad\qquad \times P[e_n \mid h_i\cdot b\cdot c_n]\times P[e^{n-1} \mid h_i\cdot b\cdot c^{n-1}] \\
&= \textstyle\sum_{\{e^{n-1}\}} \sum_{\{e_n\}} ((\mathrm{Q}[e_n]-\mathrm{E}[c_n]) + (\mathrm{Q}[e^{n-1}]-\mathrm{E}[c^{n-1}]))^2 \\
&\qquad\qquad \times P[e_n \mid h_i\cdot b\cdot c_n]\times P[e^{n-1} \mid h_i\cdot b\cdot c^{n-1}] \\
&= \textstyle\sum_{\{e^{n-1}\}} \sum_{\{e_n\}} ((\mathrm{Q}[e_n]-\mathrm{E}[c_n])^2 + (\mathrm{Q}[e^{n-1}]-\mathrm{E}[c^{n-1}])^2 \;+ \\
&\qquad 2\times(\mathrm{Q}[e_n]-\mathrm{E}[c_n])\times(\mathrm{Q}[e^{n-1}]-\mathrm{E}[c^{n-1}])) \times P[e_n \mid h_i\cdot b\cdot c_n]\times P[e^{n-1} \mid h_i\cdot b\cdot c^{n-1}] \\
&= \textstyle\sum_{\{e^{n-1}\}} \sum_{\{e_n\}} (\mathrm{Q}[e_n]-\mathrm{E}[c_n])^2 \times P[e_n \mid h_i\cdot b\cdot c_n]\times P[e^{n-1} \mid h_i\cdot b\cdot c^{n-1}] \;+ \\
&\quad \textstyle\sum_{\{e^{n-1}\}} \sum_{\{e_n\}}(\mathrm{Q}[e^{n-1}]-\mathrm{E}[c^{n-1}])^2 \times P[e_n \mid h_i\cdot b\cdot c_n]\times P[e^{n-1} \mid h_i\cdot b\cdot c^{n-1}] \;+ \\
&\quad \textstyle\sum_{\{e^{n-1}\}} \sum_{\{e_n\}} \quad 2\times(\mathrm{Q}[e_n]-\mathrm{E}[c_n])\cdot(\mathrm{Q}[e^{n-1}]-\mathrm{E}[c^{n-1}]) \;\times \\
&\qquad\qquad P[e_n \mid h_i\cdot b\cdot c_n] \times P[e^{n-1} \mid h_i\cdot b\cdot c^{n-1}] \\
&= \\
&\quad \mathrm{V}[c_n] + \mathrm{V}[c^{n-1}] \;+ \\
&\qquad 2\times\textstyle\sum_{\{e^{n-1}\}} \sum_{\{e_n\}}(\mathrm{Q}[e_n]\times\mathrm{Q}[e^{n-1}] - \mathrm{Q}[e_n]\times\mathrm{E}[c^{n-1}] - \mathrm{E}[c_n]\times\mathrm{Q}[e^{n-1}] \;+ \\
&\qquad\qquad \mathrm{E}[c_n]\times\mathrm{E}[c^{n-1}]) \times P[e_n \mid h_i\cdot b\cdot c_n]\times P[e^{n-1} \mid h_i\cdot b\times c^{n-1}] \\
&= \mathrm{V}[c_n] + \mathrm{V}[c^{n-1}] \;+ \\
&\qquad 2\times (\textstyle\sum_{\{e^{n-1}\}} \sum_{\{e_n\}} \mathrm{Q}[e_n]\times\mathrm{Q}[e^{n-1}]\times P[e_n \mid h_i\cdot b\cdot c_n]\times P[e^{n-1} \mid h_i\cdot b\times c^{n-1}] \;- \\
&\quad \textstyle\sum_{\{e^{n-1}\}} \sum_{\{e_n\}} \mathrm{Q}[e_n]\times\mathrm{E}[c^{n-1}]\times P[e_n \mid h_i\cdot b\cdot c_n]\times P[e^{n-1} \mid h_i\cdot b\times c^{n-1}] \;- \\
&\quad \textstyle\sum_{\{e^{n-1}\}} \sum_{\{e_n\}} \mathrm{E}[c_n]\times\mathrm{Q}[e^{n-1}]\times P[e_n \mid h_i\cdot b\cdot c_n]\times P[e^{n-1} \mid h_i\cdot b\times c^{n-1}] \;+ \\
&\quad \textstyle\sum_{\{e^{n-1}\}} \sum_{\{e_n\}} \mathrm{E}[c_n]\times\mathrm{E}[c^{n-1}] \times P[e_n \mid h_i\cdot b\cdot c_n]\times P[e^{n-1} \mid h_i\cdot b\times c^{n-1}]) \\
&= \mathrm{V}[c_n] + \mathrm{V}[c^{n-1}] \;+ \\
&\quad 2 \times (\mathrm{E}[c_n]\times\mathrm{E}[c^{n-1}] - \mathrm{E}[c_n]\times\mathrm{E}[c^{n-1}] - \mathrm{E}[c_n]\times\mathrm{E}[c^{n-1}] + \mathrm{E}[c_n]\times\mathrm{E}[c^{n-1}]) \\
&= \mathrm{V}[c_n] + \mathrm{V}[c^{n-1}] \\
&= \ldots \\
&= \sum_{k=1}^{n} \mathrm{VQI}[c_k \mid h_i/h_j \mid b].
\end{aligned}$$

By averaging the values of $\mathrm{VQI}[c^n \mid h_i/h_j \mid b]$ over the number of observations n we obtain a measure of the *average variance in the quality of the information* due to c^n. We represent this average by overlining ‘VQI’.

Definition: The Average Variance in the Quality of Information
$\overline{\mathrm{VQI}}[c^n \mid h_i/h_j \mid b] = \mathrm{VQI}[c^n \mid h_i/h_j \mid b] \div n$.

We are now in a position to state a very general version of the second part of the *Likelihood Ratio Convergence Theorem*. It applies to all evidence streams not containing *possibly falsifying outcomes* for h_j. That is, it applies to all evidence streams for which h_j is *fully outcome-compatible* with h_i on each $c_k$in the evidence stream. This theorem is essentially a specialized version of Chebyshev's Theorem, which is a *Weak Law of Large Numbers*.

Likelihood Ratio Convergence Theorem 2*—The Non-Falsifying Refutation Theorem.
Suppose the evidence stream c^n contains only experiments or observations on which h_j is *fully outcome-compatible* with h_i — i.e. suppose that for each condition c_k in sequence c^n, for each of its possible outcomes possible outcomes o_{ku}, either $P[o_{ku} \mid h_i \cdot b \cdot c_k] = 0$ or $P[o_{ku} \mid h_j \cdot b \cdot c_k] > 0$. And suppose that the *Independent Evidence Conditions* hold for evidence stream c^n with respect to each of these hypotheses. Now, choose any positive $\varepsilon < 1$, as small as you like, but large enough (for the number of observations n being contemplated) that the value of $\overline{\mathrm{EQI}}[c^n \mid h_i/h_j \mid b] > -(\log \varepsilon)/n$. Then:

$$P[\vee\{e^n : P[e^n \mid h_j \cdot b \cdot c^n] \,/\, P[e^n \mid h_i \cdot b \cdot c^n] < \varepsilon\} \;\;|\;\; h_i \cdot b \cdot c^n]$$

$$> \; 1 \; - \; \frac{1}{n} \times \frac{\overline{\mathrm{VQI}}[c^n \mid h_i/h_j \mid b]}{(\overline{\mathrm{EQI}}[c^n \mid h_i/h_j \mid b] + (\log \varepsilon)/n\,)^2}$$

Thus, provided that the average expected quality of the information, $\overline{\mathrm{EQI}}[c^n \mid h_i/h_j \mid b]$, for the stream of experiments and observations c^n doesn't get too small (as n increases), and provided that the average variance, $\overline{\mathrm{VQI}}[c^n \mid h_i/h_j \mid b]$, doesn't blow up (e.g. it is bounded above), hypothesis h_i (together with $b \cdot c^n$) *says* it is highly likely that outcomes of c^n will be such as to make the likelihood ratio against h_j as compared to h_i as small as you like, as n increases.

Proof: Let

$$\mathrm{V} \;=\; \mathrm{VQI}[c^n \mid h_i/h_j \mid b]$$

$$\mathrm{E} \;=\; \mathrm{EQI}[c^n \mid h_i/h_j \mid b]$$

$$\mathrm{Q}[e^n] \;=\; \mathrm{QI}[e^n \mid h_i/h_j \mid b \cdot c^n] = \log(P[e^n \mid h_i \cdot b \cdot c^n]/P[e^n \mid h_j \cdot b \cdot c^n])$$

Choose any small $\varepsilon > 0$, and suppose (for n large enough) that $\mathrm{E} > -(\log \varepsilon)/n$. Then we have

$$\begin{aligned}
\mathrm{V} &= \textstyle\sum_{\{e^n:\, P[e^n \mid h_j \cdot b \cdot c^n] > 0\}} (\mathrm{E} - \mathrm{Q})^2 \times P[e^n \mid h_i \cdot b \cdot c^n] \\
&\geq \textstyle\sum_{\{e^n:\, P[e^n \mid h_j \cdot b \cdot c^n] > 0 \;\&\; \mathrm{Q}[e^n] \leq -(\log \varepsilon)\}} (\mathrm{E} - \mathrm{Q})^2 \times P[e^n \mid h_i \cdot b \cdot c^n] \\
&\geq (\mathrm{E} + (\log \varepsilon))^2 \times \textstyle\sum_{\{e^n:\, P[e^n h_j \cdot b \cdot c^n] > 0 \;\&\; \mathrm{Q}[e^n] \leq -(\log \varepsilon)\}} P[e^n \mid h_i \cdot b \cdot c^n] \\
&= (\mathrm{E} + (\log \varepsilon))^2 \times P[\vee\{e^n:\, P[e^n \mid h_j \cdot b \cdot c^n] > 0 \;\&\; \mathrm{Q}[e^n] \leq \log(1/\varepsilon)\} \mid h_i \cdot b \cdot c^n] \\
&= (\mathrm{E} + (\log \varepsilon))^2 \times P[\vee\{e^n:\, P[e^n \mid h_j \cdot b \cdot c^n]/P[e^n \mid h_i \cdot b \cdot c^n] \geq \varepsilon\} \mid h_i \cdot b \cdot c^n]
\end{aligned}$$

So,

$$\begin{aligned}
\frac{\overline{\mathrm{V}}}{n \times (\overline{\mathrm{E}} + (\log \varepsilon)/n)^2} &= \mathrm{V}/(\mathrm{E} + (\log \varepsilon))^2 \\
&\geq P[\vee\{e^n:\, P[e^n \mid h_j \cdot b \cdot c^n]/P[e^n \mid h_i \cdot b \cdot c^n] \geq \varepsilon\} \mid h_i \cdot b \cdot c^n] \\
&= 1 - P[\vee\{e^n:\, P[e^n \mid h_j \cdot b \cdot c^n]/P[e^n \mid h_i \cdot b \cdot c^n] < \varepsilon\} \mid h_i \cdot b \cdot c^n]
\end{aligned}$$

Thus, for any small $\varepsilon > 0$,

$$P[\vee\{e^n:\, P[e^n \mid h_j \cdot b \cdot c^n]/P[e^n \mid h_i \cdot b \cdot c^n] < \varepsilon\} \mid h_i \cdot b \cdot c^n] \geq 1 - \frac{\overline{\mathrm{V}}}{n \times (\overline{\mathrm{E}} + (\log \varepsilon)/n)^2}$$

(End of Proof)

This theorem shows that when $\overline{\mathrm{VQI}}$ is bounded above and $\overline{\mathrm{EQI}}$ has a positive lower bound, a sufficiently long stream of evidence will very likely result in the refutation of false competitors of a true hypothesis. We can show that $\overline{\mathrm{VQI}}$ will indeed be bounded above when a very simple condition is satisfied. This gives us the version of the theorem stated in the main text.

Likelihood Ratio Convergence Theorem 2—The Non-Falsifying Refutation Theorem.
Suppose the evidence stream c^n contains only experiments or observations on which h_j is *fully outcome-compatible* with h_i — i.e. suppose that for each condition c_k in sequence c^n, for each of its possible outcomes possible outcomes o_{ku}, either $P[o_{ku} \mid h_i{\cdot}b{\cdot}c_k] = 0$ or $P[o_{ku} \mid h_j{\cdot}b{\cdot}c_k] > 0$. In addition (as a slight strengthening of the previous supposition), for some $\gamma > 0$ a number smaller than $1/e^2$ (≈ .135; where 'e' is the base of the natural logarithm), suppose that for each possible outcome o_{ku} of each observation condition c_k in c^n, either $P[o_{ku} \mid h_i{\cdot}b{\cdot}c_k] = 0$ or $P[o_{ku} \mid h_j{\cdot}b{\cdot}c_k] / P[o_{ku} \mid h_i{\cdot}b{\cdot}c_k] \geq \gamma$. And suppose that the *Independent Evidence Conditions* hold for evidence stream c^n with respect to each of these hypotheses. Now, choose any positive $\varepsilon < 1$, as small as you like, but large enough (for the number of observations n being contemplated) that the value of $\overline{\mathrm{EQI}}[c^n \mid h_i/h_j \mid b] > -(\log \varepsilon)/n$. Then:

$$P[\vee\{e^n : P[e^n \mid h_j{\cdot}b{\cdot}c^n] / P[e^n \mid h_i{\cdot}b{\cdot}c^n] < \varepsilon\} \quad | \quad h_i{\cdot}b{\cdot}c^n]$$

$$> 1 \;-\; \frac{1}{n} \times \frac{(\log \gamma)^2}{(\overline{\mathrm{EQI}}[c^n \mid h_i/h_j \mid b] + (\log \varepsilon)/n\,)^2}$$

Proof: This follows from Theorem 2* together with the following observation:

If for each c_k in c^n, for each of its possible outcomes o_{ku}, either $P[o_{ku} \mid h_j{\cdot}b{\cdot}c_k] = 0$ or $P[o_{ku} \mid h_j{\cdot}b{\cdot}c_k]/P[o_{ku} \mid h_i{\cdot}b{\cdot}c_k] \geq \gamma > 0$, for some lower bound $\gamma < 1/e^2$ (≈ .135; where 'e' is the base of the natural logarithm), then $\overline{\mathrm{V}} = \overline{\mathrm{VQI}}[c^n \mid h_i/h_j \mid b] \leq (\log \gamma)^2$.

To see that this observation holds, assume its antecedent.

1. First notice that when $0 < P[e_k \mid h_j{\cdot}b{\cdot}c_k] < P[e_k \mid h_i{\cdot}b{\cdot}c_k]$ we have

   $$(\log[P[e_k \mid h_i{\cdot}b{\cdot}c_k]/P[e_k \mid h_j{\cdot}b{\cdot}c_k]])^2 \times P[e_k \mid h_i{\cdot}b{\cdot}c_k]$$
   $$\leq\; (\log \gamma)^2 \times P[e_k \mid h_i{\cdot}b{\cdot}c_k].$$

   So we only need establish that when $P[e_k \mid h_j{\cdot}b{\cdot}c_k] > P[e_k \mid h_i{\cdot}b{\cdot}c_k] > 0$, we will also have this relationship — i.e., we will also have

   $$(\log[P[e_k \mid h_i{\cdot}b{\cdot}c_k]/P[e_k \mid h_j{\cdot}b{\cdot}c_k]])^2 \times P[e_k \mid h_i{\cdot}b{\cdot}c_k]$$
   $$\leq\; (\log \gamma)^2 \times P[e_k \mid h_i{\cdot}b{\cdot}c_k].$$

   (Then it will follow easily that $\overline{\mathrm{VQI}}[c^n \mid h_i/h_j \mid b] \leq (\log \gamma)^2$, and we'll be done.)

2. To establish the needed relationship, suppose that $P[e_k \mid h_j{\cdot}b{\cdot}c_k] > P[e_k \mid h_i{\cdot}b{\cdot}c_k] > 0$. Notice that for all $p \leq q$, p and q between 0 and 1, the function $g(p) = (\log(p/q))^2 \times p$ has a minimum at $p = q$, where $g(p) = 0$, and (for $p < q$) has a maximum value at $p = q/e^2$ — i.e., at $p/q = 1/e^2$. (To get this, take the derivative of $g(p)$ with respect to p and set it equal to 0; this gives a maximum for $g(p)$ at $p = q/e^2$.)

   So, for $0 < P[e_k \mid h_i{\cdot}b{\cdot}c_k] < P[e_k \mid h_j{\cdot}b{\cdot}c_k]$ we have

   $$(\log(P[e_k \mid h_i{\cdot}b{\cdot}c_k]/P[e_k \mid h_j{\cdot}b{\cdot}c_k]))^2 \times P[e_k \mid h_i{\cdot}b{\cdot}c_k]$$
   $$\leq\; (\log(1/e^2))^2 \times P[e_k \mid h_j{\cdot}b{\cdot}c_k] \;\leq\; (\log \gamma)^2 \times P[e_k \mid h_j{\cdot}b{\cdot}c_k]$$

   (since, for $\gamma \leq 1/e^2$ we have $\log \gamma \leq \log(1/e^2) < 0$; so $(\log \gamma)^2 \geq (\log(1/e^2))^2 > 0$).

3. Now (assuming the antecedent of the theorem), for each c_k,

   $$\mathrm{VQI}[c_k \mid h_i/h_j \mid b]$$
   $$= \sum_{\{o_{ku}:\, P[o_{ku} \mid h_j{\cdot}b{\cdot}c_k] > 0\}} (\mathrm{EQI}[c_k] - \mathrm{QI}[c_k])^2 \times P[o_{ku} \mid h_i{\cdot}b{\cdot}c_k]$$
   $$= \sum_{\{o_{ku}:\, P[o_{ku} \mid h_j{\cdot}b{\cdot}c_k] > 0\}} (\mathrm{EQI}[c_k]^2 - 2{\times}\mathrm{QI}[c_k]{\times}\mathrm{EQI}[c_k] + \mathrm{QI}[c_k]^2) \times P[o_{ku} \mid h_i{\cdot}b{\cdot}c_k]$$
   $$= \sum_{\{o_{ku}:\, P[o_{ku} \mid h_j{\cdot}b{\cdot}c_k] > 0\}} \mathrm{EQI}[c_k]^2 \times P[o_{ku} \mid h_i{\cdot}b{\cdot}c_k] - 2{\times}\mathrm{EQI}[c_k] \times \sum_{\{o_{ku}:\, P[o_{ku} \mid h_j{\cdot}b{\cdot}c_k] > 0\}} \mathrm{QI}[c_k] \times P[o_{ku} \mid h_i{\cdot}b{\cdot}c_k] + \sum_{\{o_{ku}:\, P[o_{ku} \mid h_j{\cdot}b{\cdot}c_k] > 0\}} \mathrm{QI}[c_k]^2 \times P[o_{ku} \mid h_i{\cdot}b{\cdot}c_k]$$
   $$= \sum_{\{o_{ku}:\, P[o_{ku} \mid h_j{\cdot}b{\cdot}c_k] > 0\}} \mathrm{QI}[c_k]^2 \times P[o_{ku} \mid h_i{\cdot}b{\cdot}c_k] - \mathrm{EQI}[c_k]^2$$
   $$\leq$$

$$\sum\{o_{ku}: P[o_{ku} \mid h_j{\cdot}b{\cdot}c_k] > 0\} \mathrm{QI}[c_k]^2 \times P[o_{ku} \mid h_i{\cdot}b{\cdot}c_k]$$
$$\leq \sum\{o_{ku}: P[o_{ku} \mid h_j{\cdot}b{\cdot}c_k] > 0\} (\log \gamma)^2 \times P[o_{ku} \mid h_i{\cdot}b{\cdot}c_k]$$
$$\leq (\log \gamma)^2.$$

So,

$$\overline{\mathrm{VQI}}[c_k \mid h_i/h_j \mid b] = (1/n) \times \sum_{k=1}^{n} \mathrm{VQI}[c_k|h_i/h_j \mid b] \leq (\log \gamma)^2.$$

[Back to Text]

Author & Citation Info | Friends PDF Preview | InPho Search | PhilPapers Bibliography

Notes to Inductive Logic

1. Although enumerative inductive arguments may seem similar to what classical statisticians call *estimation*, they are not really the same thing. As classical statisticians are quick to point out, *estimation* does not use the sample to *inductively support* a conclusion about the whole population. *Estimation* is not supposed to be a kind of inductive inference. Rather, *estimation* is a decision strategy. The sample frequency will be within two standard deviations of the population frequency in about 95% of all samples. So, if one adopts the strategy of *accepting as true* the claim that the population frequency is within two standard deviations of the sample frequency, and if one uses this strategy repeatedly for various samples, one should be right about 95% of the time. I will discuss enumerative induction in much more detail later in the article.

2. Another way of understanding axiom (5) is to view it as a generalization of the *deduction theorem and its converse*. The *deduction theorem and converse* says this: $C \vDash (B \supset A)$ if and only if $(C{\cdot}B) \vDash A$. Given axioms (1-4), axiom (5) is equivalent to the following:

5*. $(1 - P_\alpha[(B \supset A) \mid C]) = (1 - P_\alpha[A \mid (B{\cdot}C)]) \times P_\alpha[B \mid C]$.

The conditional probability $P_\alpha[A \mid (B{\cdot}C)]$ completely discounts the possibility that B is false, whereas the probability of the conditional $P_\alpha[(B \supset A) \mid C]$ depends significantly on how probable B is (given C), and must approach 1 if $P_\alpha[B \mid C]$ is near 0. Rule (5*) captures how this difference between the conditional probability and the probability of a conditional works. It says that the distance below 1 of the support-strength of C for $(B \supset A)$ equals *the product of* the distance below 1 of the support strength of $(B{\cdot}C)$ for A *and* the support strength of C for B. This makes good sense: the support of C for $(B \supset A)$ (i.e., for $({\sim}B \vee A)$) is closer to 1 than the support of $(B{\cdot}C)$ for A by the multiplicative factor $P_\alpha[B \mid C]$, which reflects the degree to which C supports ${\sim}B$. According to Rule (5*), then, for any fixed value of $P_\alpha[A \mid (B{\cdot}C)] < 1$, as $P_\alpha[B \mid C]$ approaches 0, $P_\alpha[(B \supset A) \mid C]$ must approach 1.

3. This is not what is commonly referred to as *countable additivity*. Countable additivity requires a language in which infinitely long disjunctions are defined. It would then

specify that $P_\alpha[((B_1 \vee B_2) \vee \ldots) \mid C] = \sum_i P_\alpha[B_i \mid C]$. The present result may be derived (without appealing to *countable additivity*) as follows. For each distinct i and j, let $C \vDash \sim(B_i \cdot B_j)$; and suppose that $P_\alpha[D \mid C] < 1$ for at least one sentence D. First notice that we have, for each i greater than 1 and less than n, $C \vDash (\sim(B_1 \cdot B_{i+1}) \cdot \ldots \cdot \sim(B_i \cdot B_{i+1}))$; so $C \vDash \sim(((B_1 \vee B_2) \vee \ \ldots \vee B_i) \cdot B_{i+1})$. Then, for any finite list of the first n of the B_i (for each value of n),

$$\begin{aligned} & P_\alpha[(((B_1 \vee B_2) \vee \ \ldots \vee B_{n-1}) \vee B_n) \mid C] \\ = \quad & P_\alpha[((B_1 \vee B_2) \vee \ldots \ \vee B_{n-1}) \mid C] + P_\alpha[B_\mathrm{n} \mid C] \\ = \quad & \ldots \\ = \quad & \sum_{i=1}^{n} P_\alpha[B_i \mid C]. \end{aligned}$$

By definition,

$$\sum_{i=1}^{\infty} P_\alpha[B_i \mid C] = \lim_n \sum_{i=1}^{n} P_\alpha[B_i \mid C].$$

So, $\lim_n P_\alpha[((B_1 \vee B_2) \vee \ldots \vee B_n) \mid C] = \sum_{i=1}^{\infty} P_\alpha[B_i \mid C]$

4. Here are the usual axioms when *unconditional probability* is taken as basic:

> P_α is a function from statements to real numbers between 0 and 1 that satisfies the following rules:
>
> 1. if $\vDash$A (i.e. if A is a logical truth), then $P_\alpha[A] = 1$;
> 2. if $\vDash\sim(A \cdot B)$ (i.e. if A and B are logically incompatible), then $P_\alpha[(A \vee B)] = P_\alpha[A] + P_\alpha[B]$;
>
> Definition: if $P_\alpha[B] > 0$, then $P_\alpha[A \mid B] = P_\alpha[(A \cdot B)] / P_\alpha[B]$.

5. Bayesians often refer to the probability of an evidence statement on a hypothesis, $P[e \mid h \cdot b \cdot c]$, as the *likelihood of the hypothesis*. This can be a somewhat confusing convention since it is clearly the evidence that is made likely to whatever degree by the hypothesis. So, I will disregard the usual convention here. Also, presentations of probabilistic inductive logic often suppress c and b, and simply write '$P[e \mid h]$'. But c and b are important parts of the logic of the likelihoods. So I will continue to make them explicit.

6. These attempts have not been wholly satisfactory thus far, but research continues. For an illuminating discussion of the logic of direct inference and the difficulties involved in providing a formal account, see the series of papers (Levi, 1977), (Kyburg, 1978) and (Levi, 1978). Levi (1980) develops a very sophisticated approach.

Kyburg has developed a logic of statistical inference based solely on logical direct inference probabilities (Kyburg, 1974). Kyburg's logical probabilities do not satisfy the usual axioms of probability theory. The series of papers cited above compares Kyburg's approach to a kind of Bayesian inductive logic championed by Levi (e.g., in Levi, 1967).

7. This idea should not be confused with *positivism*. A version of *positivism* applied to likelihoods would hold that if two theories assign the same likelihood values to all possible evidence claims, then they are essentially the same theory, though they may be couched in different words. In short: *same likelihoods* implies *same theory*. The view suggested here, however, is not *positivism*, but its inverse, which should be much less controversial: *different likelihoods* implies *different theories*. That is, given that all of the relevant background and auxiliaries are made explicit (represented in 'b'), if two scientists disagree significantly about the likelihoods of important evidence claims on a given hypothesis, they must understand the empirical content of that hypothesis quite differently. To that extent, though they may employ the same syntactic expressions, they use them to express empirically distinct hypotheses.

8. Call an object *grue* at a given time *just in case* either the time is earlier than the the first second of the year 2030 and the object is green or the time is not earlier than the first second of 2030 and the object is blue. Now the statement 'All emeralds are green (at all times)' has the same syntactic structure as 'All emeralds are grue (at all times)'. So, if syntactic structure determines priors, then these two hypotheses should have the same prior probabilities. Indeed, both should have prior probabilities approaching 0. For, there

are an infinite number of competitors of these two hypotheses, each sharing the same syntactic structure: consider the hypotheses 'All emeralds are grue$_n$ (at all times)', where an object is grue$_n$ at a given time *just in case* either the time is earlier than the first second of the n^{th} day after January 1, 2030, and the object is green *or* the time is not earlier than the first second of the n^{th} day after January 1, 2030, and the object is blue. A purely syntactic specification of the priors should assign all of these hypotheses the same prior probability. But these are mutually exclusive hypotheses; so their prior probabilities must sum to a value no greater than 1. The only way this can happen is for 'All emeralds are green' and each of its grue$_n$ competitors to have prior probability values either equal to 0 or infinitesimally close to it.

9. This assumption may be substantially relaxed without affecting the analysis below; we might instead only suppose that the ratios $P_\alpha[c^n \mid h_j{\cdot}b]/P_\alpha[c^n \mid h_i{\cdot}b]$ are bounded so as not to get exceptionally far from 1. If *that* supposition were to fail, then the mere occurrence of the experimental conditions would count as very strong evidence for or against hypotheses — a highly implausible effect. Our analysis could include such bounded condition-ratios, but this would only add inessential complexity to our treatment.

10. For example, when a new disease is discovered, a new hypothesis h_{u+1} about that disease being a possible cause of patients' symptoms is made explicit. The old catch-all was, "the symptoms are caused by some unknown disease — some disease other than h_1, …, h_u". So the new catch-all hypothesis must now state that "the symptoms are caused by one of the remaining unknown diseases — some disease other than $h_1,\dots, h_u, h_{u+1}$". And, clearly, $P_\alpha[h_K \mid b] = P_\alpha[{\sim}h_1{\cdot}\dots{\cdot}{\sim}h_u \mid b] = P_\alpha[{\sim}h_1{\cdot}\dots{\cdot}{\sim}h_u{\cdot}\ (h_{u+1} \vee {\sim}h_{u+1}) \mid b] = P_\alpha[{\sim}h_1{\cdot}\dots{\cdot}{\sim}h_u{\cdot}{\sim}h_{u+1} \mid b] + P_\alpha[h_{u+1} \mid b] = P_\alpha[h_{K*} \mid b] + P_\alpha[h_{u+1} \mid b]$. Thus, the new hypothesis h_{u+1} is "peeled off" of the old catch-all hypothesis h_K, leaving a new catch-all hypothesis h_{K*} with a prior probability value equal to that of the old catch-all minus the prior of the new hypothesis.

11. This claim depends, of course, on h_i being evidentially distinct from each alternative h_j. I.e., there must be conditions c_k with possible outcomes o_{ku} on which the likelihoods differ: $P[o_{ku} \mid h_i{\cdot}b{\cdot}c_k] \neq P[o_{ku} \mid h_j{\cdot}b{\cdot}c_k]$. Otherwise h_i and h_j are empirically equivalent, and no amount of evidence can support one over the other. (Did you think a confirmation theory could possibly do better? — could somehow employ evidence to confirm the true hypothesis over *evidentially equivalent* rivals?) If the true hypothesis has evidentially equivalent rivals, then convergence result just implies that the odds against *the disjunction* of the true hypothesis with these rivals very probably goes to 0, so the posterior probability of this *disjunction* goes to 1. Among evidentially equivalent hypotheses the ratio of their posterior probabilities equals the ratio of their priors: $P_\alpha[h_j \mid b{\cdot}c^n{\cdot}e^n] / P_\alpha[h_i \mid b{\cdot}c^n{\cdot}e^n] = P_\alpha[h_j \mid b] / P_\alpha[h_i \mid b]$. So the true hypothesis will have a posterior probability near 1 (after evidence drives the posteriors of evidentially distinguishable rivals near to 0) *just in case* plausibility arguments and considerations (expressed in b) make each evidentially indistinguishible rival so much less plausible by comparison that the sum of each of their comparative plausibilities (as compared to the true hypothesis) remains very small.

One more comment about this. It is tempting to identify *evidential distinguishability* (via the *evidential likelihoods*) with *empirical distinguishability*. But many plausibility arguments in the sciences, such as *thought experiments*, draw on broadly empirical considerations, on what we know or strongly suspect about how the world works based on our experience of the world. Although this kind of "evidence" may not be representable via *evidential likelihoods* (because the hypotheses it bears on don't deductively or probabilistically imply it), it often plays an important role in scientific assessments of hypotheses — in assessments of whether a hypothesis is so extraordinary that only really extraordinary likelihood evidence could rescue it. It is (arguably) a distinct virtue of the Bayesian logic of evidential support that it permits such considerations to be figured into the net evaluation of support for hypotheses.

12. This is a good place to describe one reason for thinking that *inductive support functions* must be distinct from subjectivist or personalist *degree-of-belief functions*. Although likelihoods have a high degree of objectivity in many scientific contexts, it is difficult for *belief functions* to properly represent objective likelihoods. This is an aspect of the *problem of old evidence*.

Belief functions are supposed to provide an idealized model of belief strengths for agents. They extend the notion of ideally consistent belief to a probabilistic notion of ideally coherent belief strengths. There is no harm in this kind of idealization. It is supposed to supply a normative guide for real decision making. An agent is supposed to make decisions based on her belief-strengths about the state of the world, her belief strengths about possible consequences of actions, and her assessment of the desirability (or *utility*) of these consequences. But the very role that *belief functions* are supposed to play in decision making makes them ill-suited to inductive inferences where the *likelihoods* are often supposed to be objective, or at least possess inter-subjectively agreed values that represent the empirical import of hypotheses. For the purposes of decision making, degree-of-belief functions *should* represent the agent's belief strengths *based on everything she presently knows*. So, degree-of-belief likelihoods must represent how strongly the agent would believe the evidence if the hypothesis were added to *everything else she presently knows*. However, support-function likelihoods are supposed to represent what the hypothesis (together with explicit background and experimental conditions) *says* or *implies* about the evidence. As a result, *degree-of-belief* likelihoods are saddled with a version of the *problem of old evidence* – a problem not shared by support function likelihoods. Furthermore, it turns out that the old evidence problem for likelihoods is much worse than is usually recognized.

Here is the problem. If the agent is already certain of an evidence statement e, then her *belief-function* likelihoods for that statement must be 1 on every hypothesis. I.e., if Q_γ is her *belief function* and $Q_\gamma[e] = 1$, then it follows from the axioms of probability theory that $Q_\gamma[e \mid h_i{\cdot}b{\cdot}c] = 1$, regardless of what h_i says — even if h_i implies that e is quite unlikely (given $b{\cdot}c$). But the problem goes even deeper. It not only applies to evidence that the agent *knows with certainty*. It turns out that almost anything the agent learns that can change how strongly she believes e will also influence the value of her *belief-function* likelihood for e, because $Q_\gamma[e \mid h_i{\cdot}b{\cdot}c]$ represents the agent's belief strength given *everything she knows*.

To see the difficulty with less-than-certain evidence, consider the following example. Let e be any statement that is statistically implied to degree r by a hypothesis h together with experimental conditions c (e.g. e says "the coin lands *heads* on the next toss" and $h{\cdot}c$ says "the coin is fair and is tossed in the usual way on the next toss"). Then the correct objective likelihood value is just $P[e \mid h{\cdot}c] = r$ (e.g. for $r = 1/2$). Let d be a statement that is intuitively not relevant in any way to how likely e should be on $h{\cdot}c$ (e.g. let d say "Jim will be really pleased with the outcome of that next toss"). Suppose some rational agent has a degree-of-belief function Q for which the likelihood for e due to $h{\cdot}c$ agrees with the objective value: $Q[e \mid h{\cdot}c] = r$ (e.g. with $r = 1/2$).

Our analysis will show that this agent's belief-strength for d given $\sim e{\cdot}h{\cdot}c$ will be a relevant factor; so suppose that her degree-of-belief in that regard has any value s other than 1: $Q[d \mid \sim e{\cdot}h{\cdot}c] = s < 1$ (e.g., suppose $s = 1/2$). This is a very weak supposition. It only says that adding $\sim e{\cdot}h{\cdot}c$ to everything else the agent currently knows leaves her less than certain that d is true.

Now, suppose this agent learns the following bit of new information in a completely convincing way (e.g. I seriously tell her so, and she believes me completely): $(d \vee e)$ (i.e., Jim will be really pleased with the outcome of the next toss unless it comes up *heads*).

Thus, on the usual Bayesian degree-of-belief account the agent is supposed to update her belief function Q to arrive at a new belief function Q_{new} by the updating rule:

$$Q_{new}[S] = Q[S \mid (d \vee e)], \text{ for each statement } S.$$

However, this update of the the agent's belief function *has to screw up* the objectivity of her new belief-function likelihood for e on $h{\cdot}c$, because she now should have:

$$\begin{aligned} Q_{\mathrm{new}}[e \mid h{\cdot}c] &= Q_{\mathrm{new}}[e{\cdot}h{\cdot}c] \,/\, Q_{\mathrm{new}}[h{\cdot}c] = Q[e{\cdot}h{\cdot}c \mid (d \vee e)] \,/\, Q[h{\cdot}c \mid (d \vee e)] \\ &= Q[(d \vee e){\cdot}(e{\cdot}h{\cdot}c)] \,/\, Q[(d \vee e){\cdot}(h{\cdot}c)] = Q[(d \vee e){\cdot}e \mid h{\cdot}c] \,/\, Q[(d \vee e) \mid h{\cdot}c] = \\ &Q[e \mid h{\cdot}c] \,/\, Q[((d{\cdot}\sim e) \vee e) \mid h{\cdot}c] = Q[e \mid h{\cdot}c] \,/\, [Q[e \mid h{\cdot}c] + Q[d{\cdot}\sim e \mid h{\cdot}c]] = \\ &Q[e \mid h{\cdot}c] \,/\, [Q[e \mid h{\cdot}c] + Q[d \mid \sim e \cdot h{\cdot}c] \times Q[\sim e \mid h{\cdot}c]] = r \,/\, [r + s{\times}(1-r)] = 1 \,/ \\ &[1 + s{\times}(1-r)/r]. \end{aligned}$$

Thus, the updated belief function likelihood must have value $Q_{\mathrm{new}}[e \mid h{\cdot}c] = 1 \,/\, [1 + s{\times}(1-r)/r]$.> This factor can be equal to the correct likelihood value r just in case $s = 1$. For example, for $r = 1/2$ and $s = 1/2$ we get $Q_{\mathrm{new}}[e \mid (h{\cdot}c] = 2/3$.

The point is that even the most trivial knowledge of disjunctive claims involving e may completely upset the value of the likelihood for an agent's belief function. And an agent will almost always have some such trivial knowledge. Updating on such conditionals can force the agent's *belief functions* to deviate widely from the evidentially relevant objective values of likelihoods on which scientific hypotheses should be tested.

More generally, it can be shown that the incorporation into a belief function Q of almost any kind of evidence for or against the truth of a prospective evidence claim e — even uncertain evidence for e, as may come through Jeffrey updating — completely undermines the objective or inter-subjectively agreed likelihoods that a belief function might have expressed before updating. This should be no surprise. The agent's belief function likelihoods reflect her *total degree-of-belief* in e, based on a hypothesis h together with *everything else she knows* about e. So the agent's present belief function may capture appropriate public likelihoods for *e* only if *e* is completely isolated from the agents other beliefs. And this will rarely be the case.

One Bayesian subjectivist response to this kind of problem is that the *belief functions* employed in scientific inductive inferences should often be “counterfactual” belief functions, which represent what the agent *would believe* if e were subtracted (in some suitable way) from everything else she knows (see, e.g., Howson & Urbach, 1993). However, our examples show that merely subtracting *e* won't do. One must also subtract any disjunctive statements containing *e*. And it can be shown that one must subtract any uncertain evidence for or against *e* as well. So the counterfactual belief function idea needs a lot of working out if it is to rescue the idea that *subjectivist Bayesian belief functions* can provide a viable account of the likelihoods employed by the sciences in inductive inferences.

13. To see the point more clearly, consider an example. To keep things simple, let's suppose our background *b* says that the chances of *heads* for tosses of this coin is some whole percentage between 0% and 100%. Let *c* say that the coin is tossed in the usual random way; let *e* say that the coin comes up heads; and for each *r* that is a whole fraction of 100 between 0 and 1, let $h_{[r]}$ be the *simple statistical hypothesis* asserting that the chance of heads on each random toss of this coin is *r*. Now consider the *composite statistical hypothesis* $h_{[>.65]}$, which asserts that the chance of heads on each random (independent) toss is greater than .65. From the axioms of probability we derive the following relationship: $P_\alpha[e \mid h_{[>.65]}\cdot c\cdot b] = P[e \mid h_{[.66]}\cdot c\cdot b] \times P_\alpha[h_{[.66]} \mid h_{[>.65]}\cdot c\cdot b] + P[e \mid h_{[.67]}\cdot c\cdot b] \times P_\alpha[h_{[.67]} \mid h_{[>.65]}\cdot c\cdot b] + \ldots + P[e \mid h_{[1]}\cdot c\cdot b] \times P_\alpha[h_{[1]} \mid h_{[>.65]}\cdot c\cdot b]$. The issue for the *likelihoodist* is that the values of the terms of form $P_\alpha[h_{[r]} \mid h_{[>.65]}\cdot c\cdot b]$ are not objectively specified by the composite hypothesis $h_{[>.65]}$ (together with $c\cdot b$), but the value of the likelihood $P_\alpha[e \mid h_{[>.65]}\cdot c\cdot b]$ depends essentially these non-objective factors. So, likelihoods based on composite statistical hypotheses fail to possess the kind of objectivity that *likelihoodists* require.

14. The **Law of Likelihood** and the **Likelihood Principle** have been formulated in slightly different ways by various logicians and statisticians. The **Law of Likelihood** was first identified by that name in Hacking (1965), and has been invoked more recently by the *likelihoodist* statisticians A.F.W. Edwards (1972) and R. Royall (1997). R.A. Fisher (1922) argued for the **Likelihood Principle** early in the 20th century, though he didn't call it that. One of the first places it is discussed under that name is (Savage, et al., 1962). It is also advocated by Edwards (1972) and Royall (1997).

15. What it means for a sample to be *randomly selected* from a population is philosophically controversial. Various analyses of the concept have been proposed, and disputed. For our purposes an account of the following sort will suffice. To say

> *S* is a random sample of population *B* with respect to attribute *A*

means that

> the selection set *S* is generated by a process that has an objective chance (or propensity) *r* of choosing individual objects that have attribute *A* from among the objects in population *B*, where on each selection the chance value *r* agrees with the value *r* of the frequency of *A*s among the *B*s, *F*[*A*,*B*].

Defined this way, randomness implies probabilistic independence among the outcomes of selections with regard to whether they exhibit attribute *A*, on any given hypothesis about the true value of the frequency *r* of *A*s among the *B*s.

The tricky part of generating a randomly selected set from the popualtion is to find a selection process for which the chance of selecting an *A* each time matches the true frequency without already knowing what the true frequency value is — i.e. without already knowing what the value of *r* is. However, there clearly are ways to do this. Here is one way:

> the sample *S* is generated by a process that on each selection gives each member of *B* an equal chance of being selected into *S* (like drawing balls from a well-shaken urn).

Here, schematically, is another way:

> find a subclass of *B*, call it *C*, from which *S* can be generated by a process that gives every member of *C* an equal chance of being selected into *S*, where *C* is *representative* of *B* with respect to *A* in the sense that the frequency of *A* in *C* is almost precisely the same as the frequency of *A* in *B*.

Polsters use a process of this kind. Ideally a poll of registered voters, population *B*, should select a sample *S* in a way that gives every registered voter the same chance of getting selected into S. But that may be impractical. However, it suffices if the sample is selected from a representative subpopulation C of B — e.g., from registered voters who answered the telephone between the hours of 7 PM and 9 PM in the middle of the week. Of course, the claim that a given subpopulation C is *representative* is itself a hypothesis that is open to inductive support by evidence. Professional polling organizations do a lot of research to calibrate their sampling technique, to find out what sort of subpopulations C they may draw on as highly representative. For example, one way to see if registered voters who answer the phone during the evening, mid-week, are likely to constitute a representative sample is to conduct a large poll of such voters immediately after an election, when the result is known, to see how representative of the actual vote count the count from of the subpopulation turns out to be.

Notice that although the selection set S is *selected from* B, S cannot be a subset of B, not if S can be generated by *sampling with replacement*. For, a specific member of B may be randomly selected into S more than once. If S were a subset of B, any specific member of B could only occur once in S. That is, consider the case where S consists of n selections from B, but where the process happens to select the same member b of B twice. Then, were S a subset of B, although b is selected into S twice, S can only possess b as a member once, so S has at most $n-1$ members after all (even fewer if other members of B are selected more than once). So, rather than being members of B, the members of S must be *representations of members of* B, like names, where the same member of B may be represented by different names. However, the representations (or names) in S technically may not be the sorts of things that can possess attribute A. So, technically, on this way of handling the problem, when we say that a member of S *exhibits* A, this is shorthand for *the referent of S in B possesses attribute A*.

16. This is closely analogous to the Stable-Estimation Theorem of (Edwards, Lindman, Savage, 1993). Here is a proof of Case 1, i.e. where the number of members of the reference class B is finite and where for some integer u at least as large as the size of B there is a specific (perhaps very large) integer K such that the prior probability of a hypothesis stating a frequency outside region R is never more than K times as large as a hypothesis stating a frequency within region R. (The proof is Case 2 is almost exactly the same, but draws on integrals wherever the present proof draws on sums using the '∑' expression.)

A few observations before proceeding to the main derivation:

1. The hypotheses under consideration consist of all expressions of form $F[A,B] = k/u$, where u is as described above and k is a non-negative integer between 0 and u.
2. R is some set of fractions of form k/u for a contiguous sequence of non-negative integers k that includes the sample frequency m/n.
3. In the following derivation all sums over values r in R are abbreviations for sums over integers k such that k/u is in R; similarly, all sums over values s not in R are abbreviations for sums over integers k such that k/u is not in R. The sum over $\{s \mid s=k/u\}$ represents the sum over all integers k from 0 through u.
4. Define L to be the smallest value of a prior probability $P_\alpha[F[A,B]=r \mid b]$ for r a fraction in R. Notice that $L > 0$ because, by supposition, finite $K \geq P_\alpha[F[A,B]=s \mid b] / P_\alpha[F[A,B]=r \mid b]$ for the largest value of $P_\alpha[F[A,B]=s \mid b]$ for which s is outside of R and the smallest value of $P_\alpha[F[A,B]=r \mid b]$ for which r is outside of region R.
5. Thus, from the definition of L and of K, it follows that: $K \geq P_\alpha[F[A,B]=s \mid b] / L$ for each value of $P_\alpha[F[A,B]=s \mid b]$ for which s is outside of R; and $1 \leq P_\alpha[F[A,B]=r \mid b] / L$ for each value of $P_\alpha[F[A,B]=r \mid b]$ for which r is inside of R.
6. It follows that:

$$\sum_{s \notin R} s^m \times (1-s)^{n-m} \times (P_\alpha[F[A,B]=s \mid b] / L)$$
$$\leq \sum_{s \notin R} s^m \times (1-s)^{n-m} \times P_\alpha[F[A,B]=s \mid b] \times K$$

and

$$\sum_{s \notin R} r^m \times (1-r)^{n-m} \times (P_\alpha[F[A,B]=r \mid b] / L)$$
$$\geq \sum_{r \in R} r^m \times (1-r)^{n-m} \times P_\alpha[F[A,B]=r \mid b].$$

7. For $\beta[R, m+1, n-m+1]$ defined as $\int_R r^m (1-r)^{n-m}\, dr / \int_0^1 r^m (1-r)^{n-m}\, dr$, when u is large, its an established mathematical fact that

$$\sum_{r \in R} r^m \times (1-r)^{n-m} \;\Big/\; \sum_{s \in \{s \mid s=k/u\}} s^m \times (1-s)^{n-m}$$

is extremely close to the value of β[R, m+1, n−m+1].

We now proceed to the main part of the derivation.

From the Odds Form of Bayes' Theorem (Equation 10) we have,

$$\Omega_\alpha[F[A,B] \notin R \mid F[A,S]{=}m/n \cdot \mathrm{Rnd}[S,B,A] \cdot \mathrm{Size}[S]{=}n \cdot b]$$

$$= \frac{\sum_{s \notin R} P_\alpha[F[A,B]{=}s \mid F[A,S]{=}m/n \cdot \mathrm{Rnd}[S,B,A] \cdot \mathrm{Size}[S]{=}n \cdot b]}{\sum_{r \in R} P_\alpha[F[A,B]{=}r \mid F[A,S]{=}m/n \cdot \mathrm{Rnd}[S,B,A] \cdot \mathrm{Size}[S]{=}n \cdot b]}$$

$$= \frac{\sum_{s \notin R} P[F[A,S]{=}m/n \mid F[A,B]{=}s \cdot \mathrm{Rnd}[S,B,A] \cdot \mathrm{Size}[S]{=}n \cdot b] \times P_\alpha[F[A,B]{=}s \mid b]}{\sum_{r \in R} P[F[A,S]{=}m/n \mid F[A,B]{=}r \cdot \mathrm{Rnd}[S,B,A] \cdot \mathrm{Size}[S]{=}n \cdot b] \times P_\alpha[F[A,B]{=}r \mid b]}$$

$$= \frac{\sum_{s \notin R} s^m \times (1-s)^{n-m} \times P_\alpha[F[A,B]{=}s \mid b]}{\sum_{r \in R} r^m \times (1-r)^{n-m} \times P_\alpha[F[A,B]{=}r \mid b]}$$

$$= \frac{\sum_{s \notin R} s^m \times (1-s)^{n-m} \times (P_\alpha[F[A,B]{=}s \mid b] / \mathrm{L})}{\sum_{r \in R} r^m \times (1-r)^{n-m} \times (P_\alpha[F[A,B]{=}r \mid b] / \mathrm{L})}$$

$$\leq \frac{\sum_{s \notin R} s^m \times (1-s)^{n-m} \times K}{\sum_{r \in R} r^m \times (1-r)^{n-m}}$$

$$= K \times \frac{\sum_{s \notin R} s^m \times (1-s)^{n-m} - \sum_{r \in R} r^m \times (1-r)^{n-m}}{\sum_{r \in R} r^m \times (1-r)^{n-m}}$$

$$= K \times \left[\frac{\sum_{s \notin R} s^m \times (1-s)^{n-m}}{\sum_{r \in R} r^m \times (1-r)^{n-m}} - 1 \right]$$

$$\approx K \times [(1/\beta[R, m{+}1, n{-}m{+}1]) - 1].$$

Thus,

$\Omega_\alpha[F[A,B] \notin R \mid F[A,S]{=}m/n \cdot \mathrm{Rnd}[S,B,A] \cdot \mathrm{Size}[S]{=}n \cdot b]$
$\quad \leq K \times [(1/\beta[R, m{+}1, n{-}m{+}1]) - 1].$

Then by equation (11), which expresses the relationship between *posterior probability* and *posterior odds against*,

$P_\alpha[F[A,B] \in R \mid F[A,S]{=}m/n \cdot \mathrm{Rnd}[S,B,A] \cdot \mathrm{Size}[S]{=}n \cdot b]$

$\quad = 1 / (1 + \Omega_\alpha[F[A,B] \notin R \mid F[A,S]{=}m/n \cdot \mathrm{Rnd}[S,B,A] \cdot \mathrm{Size}[S]{=}n \cdot b]$

$\quad \geq 1 / (1 + K \times [(1/\beta[R, m{+}1, n{-}m{+}1]) - 1]).$

17. To get a better idea of the import of this theorem, let's consider some specific values. First notice that the factor $r \times (1-r)$ can never be larger than (1/2)×(1/2) = 1/4; and the

closer r is to 1 or 0, the smaller $r\times(1-r)$ becomes. So, whatever the value of r, the factor $q/((r\times(1-r)/n)^{1/2} \le 2\times q\times n^{1/2}$. Thus, for any chosen value of q,

$$P[r-q < F[A,S] < r+q \mid F[A,B] = r\cdot \mathrm{Rnd}[S,B,A]\cdot \mathrm{Size}[S] = n] \\ \ge 1 - 2\times\Phi[-2\times q\times n^{1/2}].$$

For example, if $q = .05$ and $n = 400$, then we have (for any value of r),

$$P[r-.05 < F[A,S] < r+.05 \mid F[A,B] = r\cdot \mathrm{Rnd}[S,B,A]\cdot \mathrm{Size}[S] = 400] \ge .95.$$

For $n = 900$ (and margin $q = .05$) this lower bound raises to .997:

$$P[r-.05 < F[A,S] < r+.05 \mid F[A,B] = r\cdot \mathrm{Rnd}[S,B,A]\cdot \mathrm{Size}[S] = 900] \ge .997.$$

If we are interested in a smaller margin of error q, we can keep the same sample size and find the value of the lower bound for that value of q. For example,

$$P[r-.03 < F[A,S] < r+.03 \mid F[A,B] = r\cdot \mathrm{Rnd}[S,B,A]\cdot \mathrm{Size}[S] = 900] \ge .928.$$

By increasing the sample size the bound on the likelihood can be made as close to 1 as we want, for any margin q we choose. For example:

$$P[r-.01<F[A,S] <r+.01 \mid F[A,B] = r\cdot \mathrm{Rnd}[S,B,A]\cdot \mathrm{Size}[S] = 38000] \ge .9999.$$

As the sample size n becomes larger, it becomes extremely likely that the sample frequency will come to within any specified region close to the true frequency r, as close as you wish.

18. That is, for each inductive support function P_α, the posterior $P_\alpha[h_j \mid b\cdot c^n\cdot e^n]$ must go to 0 as the ratio $P_\alpha[h_j \mid b\cdot c^n\cdot e^n] / P_\alpha[h_i \mid b\cdot c^n\cdot e^n]$ goes to 0; and that must occur if the likelihood ratios $P[e^n \mid h_j\cdot b\cdot c^n] / P[e^n \mid h_i\cdot b\cdot c^n]$ approach 0, provided that and the prior probability $P_\alpha[h_i \mid b]$ is greater than 0. The Likelihood Ratio Convergence Theorem will show that when $h_i\cdot b$ is true, it is very likely that the evidence will indeed be such as to drive the likelihood ratios as near to 0 as you please, for a long enough (or strong enough) evidence stream. (If the stream is *strong* in that the likelihood ratios of individual bits of evidence are small, then to bring about a very small cumulative likelihood ratio, the evidence stream need not be as long.) As likelihood ratios head towards 0, the only way a Bayesian agent can avoid having her inductive support function(s) yield posterior probabilities for h_j that approach 0 (as n gets large) is to continually change her prior probability assessments. That means either continually finding and adding new plausibility arguments (i.e. adding to or modifying b) that on ballance favor h_j over h_i, or continually reassessing the support strength due to plausibility arguments already available, or both.

Technically, continual reassessments of support strengths that favor h_j over h_i based on already extant arguments (in b) means switching to new support functions (or new *vagueness sets* of them) that assign h_j ever higher prior probabilities as compared to h_i based on the same arguments in b. In any case, such revisions of argument strengths may avoid the convergence towards 0 of the posterior probability of h_j only if it proceeds at a rate that keeps ahead of the rate at which the evidence drives the likelihood ratios towards 0.

For a thorough presentation of the most prominent Bayesian convergence results and a discussion of their weaknesses see (Earman, 1992, Ch. 6). However, Earman does not discuss the convergence theorems under consideration here (due to the fact that the convergence results discussed here first appeared in (Hawthorne, 1993), just after Earman's book came out).

19. In scientific contexts all of the most important kinds of cases where large components of the evidence fail to be *result-independent* of one another are cases where some part of the total evidence helps to tie down the numerical value of a parameter that plays an important role in the likelihood values the hypothesis specifies for other large parts of the total evidence. In cases where this only happens rather *locally*, where the evidence for a parameter value influences the likelihoods of only a very small part of the total evidence that bears on the hypothesis, we can treat the conjunction of *the evidence for the parameter value* with *the evidential outcomes whose likelihood the parameter value influences* as a single *chunk of evidence*, which is then *result-independent* of the rest of the evidence (on each alternative hypothesis). This is the sort of *chuncking of the evidence* into *result-independent parts* suggested in the main text.

However, in cases where the value of a parameter left unspecified by the hypothesis has a wide-ranging influence on many of the likelihood values the hypothesis specifies, another strategy for obtaining *result-independence* among these components of the evidence will do the job. A hypothesis that has an unspecified parameter value is in effect equivalent to a *disjunction of more specific hypotheses*, where each disjunct consists of a more precise version of the original hypothesis, a version in which the value for the parameter has been "filled in". Relative to each of these more precise hypotheses, any evidence for or against the parameter value that hypothesis specifies is evidence for or against that more precise hypothesis itself. Furthermore, the evidence whose likelihood values depend on the parameter value (and because of that, failed to be *result-independent* of the parameter value evidence relative to the original hypothesis) is *result-independent* of the parameter value evidence relative to each of these more precise hypotheses — because each of the precise hypotheses already identifies precisely what (it claims) the value of the parameter is. Thus, wherever the workings of the logic of evidential support is made more perspicuous by treating evidence as composed of *result-independent chunks*, one may treat hypotheses whose unspecified parameter values interfere with *result-independence* as *disjunctively composite hypotheses*, and apply the evidential logic to these more specific disjuncts, and thereby regain *result-independence*.

20. Technically, suppose that O_k can be further "subdivided" into more outcome-descriptions by replacing o_{kv} with two "mutually exclusive parts", $o_{kv}{}^*$ and $o_{kv}{}^\#$, to produce new outcome space $O_k{}^\$ = \{o_{k1},\ldots,o_{kv}{}^*,o_{kv}{}^\#,\ldots,o_{kw}\}$, where $P[o_{kv}{}^* \cdot o_{kv}{}^\# \mid h_i \cdot b \cdot c_k] = 0$ and $P[o_{kv}{}^* \mid h_i \cdot b \cdot c_k] + P[o_{kv}{}^\# \mid h_i \cdot b \cdot c_k] = P[o_{kv} \mid h_i \cdot b \cdot c_k]$; and suppose similar relationships hold for h_j. Then the new EQI* (based on $O_k{}^*$) is greater than or equal to EQI (based on O_k); and EQI* > EQI *just in case* at least one of the new likelihood ratios, e.g., $P[o_{kv}{}^* \mid h_i \cdot b \cdot c_k] / P[o_{kv}{}^* \mid h_j \cdot b \cdot c_k]$, differs in value from the "undivided" outcome's likelihood ratio, $P[o_{kv} \mid h_i \cdot b \cdot c_k] / P[o_{kv} \mid h_i \cdot b \cdot c_k]$. A supplement linked to this article proves this claim.

21. The likely rate of convergence will almost always be much faster than the worst case bound provided by Theorem 2. To see the point more clearly, let's look at a very simple example. Suppose h_i says that a certain bent coin has a propensity for "heads" of 2/3 and h_j says the propensity is 1/3. Let the evidence stream consist of outcomes of tosses. In this case the average EQI equals the EQI of each toss, which is 1/3; and the smallest possible likelihood ratio occurs for "heads", which yields the value $\gamma = ½$. So, the value of the lower bound given by Theorem 2 for the likelihood of getting an outcome sequences with a likelihood ratio below ε (for h_j over h_i) is

$$1 - (1/n)(\log ½)^2/((1/3) + (\log \varepsilon)/n)^2 = 1 - 9/(n\times(1 + 3(\log \varepsilon)/n)^2.$$

Thus, according to the theorem, the likelihood of getting an outcome sequence with a likelihood ratio less than $\varepsilon = 1/16$ (=.06) when h_i is true and the number of tosses is n = 52 is *at least* .70; and for n = 204 tosses the likelihood is *at least* .95.

To see the amount by which the lower bound provided by the theorem is in fact *overly cautious*, consider what the usual binomial distribution for the coin tosses in this example implies about the likely values of the likelihood ratios. The likelihood ratio for exactly k "heads" in n tosses is $((1/3)^k (2/3)^{n-k}) / ((2/3)^k (1/3)^{n-k}) = 2^{n-2k}$; and we want this likelihood ratio to have a value less than ε. A bit of algebraic manipulation shows that to get this likelihood ratio value to be below ε, the percentage of "heads" needs to be $k/n > ½ - ½(\log \varepsilon)/n$. Using the normal approximation to the binomial distribution (with mean = 2/3 and variance = $(2/3)(1/3)/n$) the actual likelihood of obtaining an outcome sequence having more than $½ - ½(\log \varepsilon)/n$ "heads" (which we just saw corresponds to getting a likelihood ratio less than ε, thus disfavoring the 1/3 propensity hypothesis as compared to the 2/3 propensity hypothesis by that much) when the true propensity for "heads" is 2/3 is given by the formula

$$\Phi[(\text{mean} - (½ - ½(\log \varepsilon)/n))/(\text{variance})^{½}] = \Phi[(1/8)^{½} n^{½}(1 + 3(\log \varepsilon)/n)]$$

(where $\Phi[x]$ gives the value of the standard normal distribution from $-\infty$ to x). Now let ε = 1/16 (= .0625), as before. So the actual likelihood of obtaining a stream of outcomes with likelihood ratio this small when h_i is true and the number of tosses is n = 52 is $\Phi[1.96] > .975$, whereas the lower bound given by Theorem 2 was .70. And if the number of tosses is increased to n = 204, the likelihood of obtaining an outcome sequence with a likelihood ratio this small (i.e., ε = 1/16) is $\Phi[4.75] > .999999$, whereas the lower bound from Theorem 2 for this likelihood is .95. Indeed, to actually get a likelihood of .95 that the evidence stream will produce a likelihood ratio less than ε >.06, the number of tosses needed is only n = 43, rather than the 204 tosses the bound given by the theorem requires in order to get up to the value .95. (Note: These examples employ "identically distributed" trials — repeated tosses of a coin — as an illustration. But Convergence Theorem 2 applies much more generally. It applies to any evidence sequence, no matter how diverse the probability distributions for the various experiments or observations in the sequence.)

22. It should now be clear why the boundedness of EQI above 0 is important. Convergence Theorem 2 applies only when $\overline{\text{EQI}}[c^n \mid h_i/h_j \mid b] > -(\log \varepsilon)/n$. But this requirement is not a strong assumption. For, the **Nonnegativity of EQI Theorem** shows that the empirical distinctness of two hypotheses on a single possible outcome *suffices* to make the average EQI positive for the whole sequence of experiments. So, given any small fraction $\varepsilon > 0$, the value of $-(\log \varepsilon)/n$ (which is always greater than 0 when $\varepsilon < 0$) will eventually become smaller than $\overline{\text{EQI}}$, provided that the degree to which the hypotheses are empirical distinct for the various observations c_k does not on average degrade too much as the length n of the evidence stream increases.

When the possible outcomes for the sequence of observations are independent and identically distributed, Theorems 1 and 2 effectively reduce to L. J. Savage's Bayesian Convergence Theorem [Savage, pg. 52-54], although Savage's theorem doesn't supply explicit lower bounds on the probability that the likelihood ratio will be small. Independent, identically distributed outcomes most commonly result from the repetition of identical statistical experiments (e.g., repeated tosses of a coin, or repeated measurements of quantum systems prepared in identical states). In such experiments a hypothesis will specify the same likelihoods for the same kinds of outcomes from one observation to the next. So $\overline{\text{EQI}}$ will remain constant as the number of experiments, n, increases. However, Theorems 1 and 2 are much more general. They continue to hold when the sequence of observations encompasses completely unrelated experiments that have different distributions on outcomes — experiments that have nothing in common but their connection to the hypotheses they test.

23. In many scientific contexts this is the best we can hope for. But it still provides a very reasonable representation of inductive support. Consider, for example, the hypothesis that the land masses of Africa and South America separated and drifted apart over the eons, the *drift hypothesis*, as opposed to the hypothesis that the continents have fixed positions acquired when the earth first formed and cooled and contracted, the *contraction hypothesis*. One may not be able to determine anything like precise likelihoods, on each hypothesis, for the evidence that: (1) the shape of the east coast of South America matches the shape of the west coast of Africa as closely as it in fact does; (2) the geology of the two coasts match up so closely when they are “fitted together” in the obvious way; (3) the plant and animal species on these distant continents should be as similar as they are, as compared to how similar species are among other distant continents. Although neither the *drift hypothesis* nor the *contraction hypothesis* supplies anything like precise likelihoods for these evidential claims, experts readily agree that each of these observations is *much more likely* on the *drift hypothesis* than on the *contraction hypothesis*. That is, the likelihood ratio for this evidence on the *contraction hypothesis* as compared to the *drift hypothesis* is very small. Thus, jointly these observations constitute very strong evidence for *drift* over *contraction*.

Historically, the case of continental drift is more complicated. Geologists tended to largely dismiss this evidence until the 1960s. This was not because the evidence wasn't strong in its own right. Rather, this evidence was found unconvincing because it was not sufficient to overcome prior plausibility considerations that made the *drift hypothesis* extremely implausible — much less plausible than the *contraction hypothesis*. The problem was that there seemed to be no plausible mechanism by which *drift* might occur. It was argued, quite plausibly, that no known force could push or pull the continents apart, and that the less dense continental material could not push through the denser material that makes up the ocean floor. These plausibility objections were overcome when a plausible mechanism was articulated — i.e. the continental crust floats atop molten material and moves apart as convection currents in the molten material carry it along. The case was pretty well clinched when evidence for this mechanism was found in the form of “spreading zones” containing alternating strips of magnetized material at regular distances from mid-ocean ridges. The magnetic alignments of materials in these strips corresponds closely to the magnetic alignments found in magnetic materials in dateable sedimentary layers at other locations on the earth. These magnetic alignments indicate time periods when the direction of earth's magnetic field has reversed. And this gave geologists a way of measuring the rate at which the sea floor might spread and the continents move apart. Although geologists may not be able to determine anything like precise values for the likelihoods of any of this evidence on each of the alternative hypotheses, the evidence is universally agreed to be *much more likely* on the *drift hypothesis* than on the alternative *contraction hypothesis*. The *likelihood ratio* for this evidence on the *contraction hypothesis* as compared to the *drift hypothesis* is somewhat vague, but extremely small. The vagueness is only in regard how extremely small the likelihood ratio is. Furthermore, with the emergence of a plausible mechanism, the *drift hypothesis* hypothesis is no longer so overwhelmingly implausible *prior* to taking the likelihood evidence into account. Thus, even when precise values for individual likelihoods are not available, the value of *a likelihood ratio range* may be *objective enough* to strongly refute one hypothesis as compared to another. Indeed, the *drift hypothesis* is itself strongly supported by the evidence; for, no alternative hypothesis that has the slightest amount of comparative plausibility can account for the available evidence nearly so well. (That is, no plausible alternative makes the evidence anywhere near so likley.) Given the currently available evidence, the only issues left open (for now) involve comparing various alternative versions of the drift hypothesis (involving differences of detail) against one another.

24. To see the point of the third clause, suppose it were violated. That is, suppose there are possible outcomes for which the likelihood ratio is very near 1 for just one of the two support functions. Then, even a very long sequence of such outcomes might leave the likelihood ratio for one support function almost equal to 1, while the likelihood ratio for the other support function goes to an extreme value. If that can happen for support functions in a class that represent likelihoods for various scientists in the community, then the empirical contents of the hypotheses is either too vague or too much in dispute for meaningful empirical evaluation to occur.

25. If there are a few directionally controversial likelihood ratios, where P_α says the ratio is somewhat greater than 1, while and P_β assigns a value somewhat less than 1, these may not greatly effect the trend of P_α and P_β towards agreement on the refutation and support of hypotheses *provided that* the controversial ratios are not so extreme as to overwhelm the stream of other evidence on which the likelihood ratios do directionally agree. Even so, researches will want to get straight on what the hypothesis *says* or *implies* about such cases. While that remains in dispute, the empirical content of the hypothesis remains unsettling vague.

James Hawthorne <*hawthorne@ou.edu*>