From Statistics to Beliefs*

Fahiem Bacchus
Computer Science Dept.
University of Waterloo
Waterloo, Ontario
Canada, N2L 3G1
fbacchus@logos.waterloo.edu

Adam Grove
Computer Science Dept.
Stanford University
Stanford, CA 943005
grove@cs.stanford.edu

Joseph Y. Halpern
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120–6099
halpern@almaden.ibm.com

Daphne Koller
Computer Science Dept.
Stanford University
Stanford, CA 943005
daphne@cs.stanford.edu

Abstract

An intelligent agent uses known facts, including statistical knowledge, to assign *degrees of belief* to assertions it is uncertain about. We investigate three principled techniques for doing this. All three are applications of the *principle of indifference*, because they assign equal degree of belief to all basic "situations" consistent with the knowledge base. They differ because there are competing intuitions about what the basic situations are. Various natural patterns of reasoning, such as the preference for the most specific statistical data available, turn out to follow from some or all of the techniques. This is an improvement over earlier theories, such as work on direct inference and reference classes, which arbitrarily postulate these patterns without offering any deeper explanations or guarantees of consistency.

The three methods we investigate have surprising characterizations: there are connections to the principle of maximum entropy, a principle of maximal independence, and a "center of mass" principle. There are also unexpected connections between the three, that help us understand why the specific language chosen (for the knowledge base) is much more critical in inductive reasoning of the sort we consider than it is in traditional deductive reasoning.

1 Introduction

An intelligent agent must be able to use its accumulated knowledge to help it reason about the situation it is currently facing. Consider a doctor who has a knowledge base consisting of statistical and first-order information regarding symptoms and diseases, and some specific information regarding a particular patient. She wants to make an inference regarding the likelihood that the patient has cancer. The inference of such a likelihood, or *degree of belief,* is an essential step in decision making. We present here a general and principled mechanism for computing degrees of belief. This mechanism has a number of particular realizations which differ in the inferences they support. Through an analysis of these differences and of the principles which underlie the general mechanism, we are able to offer a number of important new insights into this form of reasoning.

To illustrate some of the subtle issues that arise when trying to compute degrees of belief, suppose that the domain consists of American males, and that the agent is interested in assigning a degree of belief to the proposition "Eric (an American male) is (over six feet) tall," given some subset of the following database:

A 20% of American males are tall.

B 25% of Californian males are tall.

C Eric is a Californian male.

A traditional approach to assigning a degree of belief to *Tall(Eric)* is to find an appropriate class—called the *reference class*—which includes Eric and for which we have statistical information, and use the statistics for that class to compute an appropriate degree of belief for Eric. Thus, if the agent's database consists solely of item **A**, then this approach would attach a quite reasonable degree of belief of 0.2 to *Tall(Eric)* using the reference class of American males.

This general approach to computing degrees of belief goes under the name *direct inference*, and dates back to Reichenbach (Rei49), who used the idea in an attempt to reconcile his frequency interpretation of probability with the common practice of attaching probabilities to particular cases. He expounded a principle for direct inference, but did not develop a complete mechanism. Subsequently, a great deal of work has been done on formalizing and mechanizing direct inference (Bac90; Kyb61; Kyb74; Lev80; Pol90; Sal71).

If the database consists only of **A**, there is only one reference class to which Eric is known to belong, so applying direct inference is easy. In general, however, the particular individual or collection of individuals we are reasoning about will belong to many different classes. We might possess conflicting statistics for some of these classes, and for others we might not possess any statistical information at all. The difficulty with direct

*The work of Fahiem Bacchus was supported by NSERC under their operating grants program and by IRIS. The work of Adam Grove, Joseph Halpern, and Daphne Koller was sponsored in part by the Air Force Office of Scientific Research (AFSC), under Contract F49620-91-C-0080. Adam Grove's work was also supported by an IBM Graduate Fellowship. The United States Government is authorized to reproduce and distribute reprints for governmental purposes. This paper appears in *Proceedings of AAAI-92*, 1992, pp. 602–608.

inference, then, is how to choose an appropriate reference class. There are a number of issues that arise in such a choice, but we focus here on three particular problems.

Specific Information: Suppose the knowledge base consists of all three items **A**–**C**. Now Eric is a member of two reference classes: Americans and Californians. Intuition suggests that in this case we should choose the more specific class, Californians. And indeed, all of the systems cited above embody a preference for more specific information, yielding 0.25 as the degree of belief in *Tall(Eric)* in this case.

However, we must be careful in applying such a preference. For one thing, we must deal with the problem of *disjunctive reference classes*. Consider the disjunctive class D consisting of Eric and all non-tall Californian males. Being a subset of Californian males this is clearly a more specific reference class. If there are many Californians (and thus many non-tall Californians, since 75% of Californians are not tall), using D as the reference class gives a degree of belief for *Tall(Eric)* that is very close to 0. The answer 0.25 seems far more reasonable, showing that we must be careful about how we interpret the principle of preference for more specific information. We remark that two of the most well-developed systems of direct inference, Kyburg's (Kyb74) and Pollock's (Pol90), address this issue by the *ad hoc* device of simply outlawing disjunctive reference classes.

Irrelevant Information: Suppose that the knowledge base consists only of items **A** and **C**. In this case Eric again belongs to two reference classes, but now we do not have any statistics for the more specific class, Californians. The standard, and plausible, approach is to assign a degree of belief of 0.2 to *Tall(Eric)*. That is, we use the statistics from the superclass, American males; this amounts to assuming that the extra information, that Eric is also Californian, is irrelevant. In the face of no knowledge to the contrary, we assume that the subclass has the same statistics as the superclass.

Sampling: Finally, suppose the knowledge base consists only of **B**. In this case we have statistical information about Californians, but all we know about Eric is that he is an American. We could assume that Californians are a representative sample of Americans when it comes to male tallness and adopt the statistics we have for the class of Californians, generating a degree of belief of 0.25 in *Tall(Eric)*.

The process of finding the "right" reference class, and then assigning degrees of belief using the further assumption that the individual in question is randomly chosen from this class, is one way of going from statistical and first-order information to a degree of belief. But, as we have seen, choosing the right reference class is a complex issue. It is typically accomplished by positing some collection of rules for choosing among competing reference classes, e.g., (Kyb83). However, such rules are not easy to formulate. More importantly, they do not provide any general principles which can help elucidate our intuitions about how statistics should influence degrees of belief. Indeed, the whole idea of reference class seems artificial; it does not occur at all in the statement of the problem we are trying to solve. We present a different approach to computing degrees of belief here, one that does not involve finding appropriate reference classes at all. We believe it is a more general, high-level approach, that deals well with the three problems discussed above and, as we show in the full paper, many others besides.

The essential idea is quite straightforward. We view the information in the database as determining a set of situations that the agent considers possible. In order to capture the intuition that the information in our knowledge base is "all we know," we assign each of these possible situations equal probability. After all, our information does not give us any reason to give any of them greater probability than any other. Roughly speaking, the agent's degree of belief in a sentence such as *Tall(Eric)* is then the fraction of the set of situations where *Tall(Eric)* is true. The basic idea here is an old one, going back to Laplace (Lap20), and is essentially what has been called the *principle of insufficient reason* (Kri86) or the *principle of indifference* (Key21).

Our general method, then, revolves around applying indifference to some collection of possible situations. The method has a number of different realizations, as there are competing intuitions involved in defining a "possible situation." We focus on three particular mechanisms for defining situations, each of which leads to a different method of computing degrees of belief. The differences between the three methods reflect different intuitions about how degrees of belief should be generated from statistics. They also point out the important role of language in this process.

Although the approaches are different, they share some reasonable properties. For example, as we show in Section 3, they all generalize deductive inference, they all agree with direct inference in noncontroversial cases, and they all capture a preference for more specific information. Furthermore, since our general method does not depend on finding reference classes, the problem of disjunctive classes completely disappears.

Despite these similarities, the methods differ in a number of significant ways (see Section 3). So which method is "best"? Since all the methods are defined in terms of assigning equal probability to all possible situations, the question comes down to which notion of "situation" is most appropriate. As we show, that depends on what intuitions we are trying to capture. Our framework allows us to obtain an understanding of when each method is most appropriate. In addition, it gives us the tools to consider other methods, and hybrids of these methods. Because there is no unique "best" answer, this is a matter of major importance.

There has been a great deal of work that can be viewed as attempts to generate degrees of belief given a database. Besides the work on reference classes mentioned above, much of Jaynes's work on maximum entropy (Jay78) can be viewed in this light. Perhaps the work closest in spirit to ours is that of Carnap (Car52), Johnson (Joh32), and the more recent work of Paris and Vencovska (PV89; PV92) and Goodwin (Goo92). We compare our work with these others in some detail in the full paper; here we can give only brief hints of the relationship.

2 The Three Methods

We assume that the knowledge base consists of sentences written in a formal language that allows us to express both statistical information and first-order information. In particular, we use a simplified version of a probability logic developed in (Bac90) and (Hal90), which we describe very briefly here.

To represent the statistical information, we augment first-order logic by allowing proportion formulas of the form $\|\psi(x)\|_x$, which denotes the proportion of individuals in the domain satisfying ψ when instantiated for x. Notice that this proportion is well defined in any first-order model (over an appropriate vocabulary) if the model's domain is finite; in the following, this will always be the case. For example, $\|\mathit{Californian}(x)\|_x = .1$ says that 10% of the domain elements are Californians, while $\|\mathit{Tall}(x)|\mathit{Californian}(x)\|_x = .25$ says that 25% of Californians are tall, via the standard abbreviation for conditional probabilities (and thus represents assertion **B** from the introduction).[1]

We want to use the information in the knowledge base to compute a degree of belief. Note that there is an important distinction between statistical information such as "25% of Californian males are tall" and a degree of belief such as "the likelihood that Eric is tall is .25". The former represents real-world data, while the latter is attached by the agent, hopefully using a principled method, to assertions about the world that are, in fact, either true or false.[2] Following (Hal90), we give semantics to degrees of belief in terms of a set of first-order models or *possible worlds*, together with a probability distribution over this set. The degree of belief in a sentence φ is just the probability of the set of worlds where φ is true. For our method the set of possible worlds is easily described: given a vocabulary Φ and domain size N, it is the collection of all first-order structures over the vocabulary Φ with domain $\{1, \ldots, N\}$.

The probability distribution is generated by applying the principle of indifference to equivalence classes of worlds ("situations"). We assign equal probability to every equivalence class, and then, applying the principle of indifference again, we divide up the probability assigned to each class equally among the individual worlds in that class.

Alternate realizations of our method arise from different intuitions as to how to group the worlds into equivalence classes. We consider three natural groupings, which lead to the three methods mentioned in the introduction. (Of course, other methods are possible, but we focus on these three for now, deferring further discussion to the full paper.) Once we have the probability distribution on the worlds, we compute the degree of belief in φ given a database KB by using straightforward conditional probability: it is simply the probability of the set of worlds where $\varphi \wedge KB$ is true divided by the probability of the set of worlds where KB is true.

In this paper we restrict attention to vocabularies having only constants and unary predicates. Our methods make perfect sense when applied to richer vocabularies (see the full paper), but the characterizations of these methods given in Section 3 hold only in the unary case.

In the first approach, which we call the *random-worlds* approach, we identify situations and worlds. Hence, by the principle of indifference, each world is assigned equal probability.

In the second approach, which we call the *random-structures* approach, we group into a single equivalence class worlds that are *isomorphic* with respect to the predicates in the vocabulary.[3] The intuition underlying this approach is that individuals with exactly the same properties are in some sense indistinguishable, so worlds where they are simply renamed should be treated as being equivalent.

Suppose, for example, that our vocabulary consists of a unary predicate P and a constant c, and that the domain size is N. Since P can denote any subset, and c any member, of the domain, there will be $N2^N$ possible worlds, each with a distinct interpretation of the vocabulary. In the random-worlds approach each world is an equally likely situation with equal probability. In the random-structures approach, on the other hand, worlds in which the cardinality of P's denotation is the same are isomorphic, and thus will all be grouped into a single situation. Hence, there are only $N + 1$

[1] As pointed out in (PV89; GHK94), we actually want to use "approximately equals" rather than true equality when describing statistical information. If we use true equality, the statement $\|\mathit{Californian}\ (x)\|_x = .1$ would be false if the domain size were not a multiple of 10. In this paper, we ignore the subtleties involved with approximate equality, since they are not relevant to our discussion.

[2] This distinction between statistical information (frequencies) and degrees of belief has long been noted in studies of the foundations of probability. See, for instance, Carnap's discussion in (Car50).

[3] Note that we only consider the predicate denotations when deciding on a world's equivalence class, and ignore the denotations of constants. This is consistent with Carnap's approach (Car52), and is crucial for our results. See the full paper for further discussion of this point.

equally likely situations, one for each possible size r of P's denotation. Each such situation is assigned probability $1/(N+1)$ and that probability is divided equally among the $N\binom{N}{r}$ worlds in that situation. So, according to random-worlds, it is much more likely that the number of individuals satisfying P is $\lfloor N/2 \rfloor$ than that it is 1, whereas for random-structures these two possibilities are equally likely.

More generally, suppose the vocabulary consists of the unary predicate symbols $P_1, \ldots, P_k$ and the constants $c_1, \ldots, c_\ell$. We can consider the 2^k *atoms* that can be formed from the predicate symbols, namely, the formulas of the form $Q_1 \wedge \ldots \wedge Q_k$, where each Q_i is either P_i or $\neg P_i$. If we have a domain of size N, there will be $N^\ell (2^N)^k$ possible worlds, corresponding to all choices for the denotations of the ℓ constants and k predicates. Given two possible worlds w_1 and w_2, it is easy to see that they are isomorphic with respect to the predicates if and only if for every atom the number of individuals satisfying that atom in w_1 is the same as in w_2. This means that a random-structures situation is completely described by a tuple $(d_1, \ldots, d_{2^k})$ with $d_1 + \cdots + d_{2^k} = N$, specifying how many domain elements satisfy each atom. Using standard combinatorics, it can be shown that there are exactly $\binom{N+2^k-1}{2^k-1}$ such situations.

The third method we consider, which we call the *random-propensities* approach, attempts to measure the *propensity* of an individual to satisfy each of the predicates. If our vocabulary contains the unary predicates $P_1, \ldots, P_k$ and the domain has size N, then a situation in this approach is specified by a tuple $(e_1, \ldots, e_k)$; the worlds contained in this situation are all those where e_i of the domain elements satisfy P_i, for all i.[4] Intuitively, e_i/N is a measure of the propensity of an individual to have property P_i. It is not difficult to see that there are $(N+1)^k$ distinct situations.

Suppose, for example, that the vocabulary consists of the unary predicates P and Q and that the domain consists of three elements $\{a, b, c\}$. There are $(2^3)^2 = 64$ distinct possible worlds, one for each choice of denotation for P and Q. In the random-worlds approach each of these worlds will be assigned probability 1/64. In the random-structures approach there are $\binom{3+2^2-1}{2^2-1} = \binom{6}{3} = 20$ distinct situations. Each will be given probability 1/20 and that probability will then be divided equally among the worlds in the situation. For example, the world w that assigns P the denotation $\{a\}$ and Q the denotation $\{a, c\}$ belongs to the situation in which the atom $P \wedge \neg Q$ has size 0 and all other atoms have size 1. There are 6 worlds in this situation, so w will be assigned probability $\frac{1}{6\times 20}$. In the random-propensities approach there are $(3+1)^2 = 16$ distinct situations. Each will be given probability 1/16 to be divided equally among the worlds in the situation. For example, one of these situations is specified by the tuple $(1, 2)$ consisting of all those worlds where one element satisfies P and two satisfy Q. This situation contains 9 worlds, including the world w specified above. Hence, under random-propensities w is assigned probability $\frac{1}{9\times 16}$.

We remark that two of our three methods—the random-worlds method and the random-structures method—are not original to us. They essentially date back to Laplace (Lap20), and were investigated in some detail by Johnson (Joh32) and Carnap (Car50; Car52). (These two methods correspond to Carnap's state-description and structure-description techniques, respectively.) We believe that the random-propensities method is new; as we shall show, it has some quite attractive properties.

If KB is a formula describing the knowledge base, φ is a first-order sentence, and N is the domain size, we denote by $\Pr_N^w(\varphi|KB)$, $\Pr_N^s(\varphi|KB)$, and $\Pr_N^p(\varphi|KB)$ the degree of belief in φ given knowledge base KB according to the random-worlds, random-structures, and random-propensities methods, respectively. We write $\Pr_N^*(\varphi|KB)$ in those cases where the degree of belief is independent of the approach.

We often do not know the precise domain size N, but do know that it is large. This leads us to consider the asymptotic behavior of the degree of belief as N gets large. We define $\Pr_\infty^w(\varphi|KB) = \lim_{N\to\infty} \Pr_N^w(\varphi|KB)$; $\Pr_\infty^s$, $\Pr_\infty^p$ and $\Pr_\infty^*$ are similarly defined.[5]

Our methods can also be viewed as placing different priors on the set of first-order structures. Viewed in this way, they are instances of Bayesian inference, since we compute degrees of belief by conditioning on this prior distribution, given our database. But the deepest problem when applying Bayesian inference is always finding the prior distribution, or, even more fundamentally, finding the appropriate space of possibilities. This is precisely the problem we address here.

3 Understanding the Methods

As a first step to understanding the three techniques, we look for general properties characterizing their behavior. Then we examine some specific properties which tell us how the techniques behave in various paradigmatic reasoning situations.

3.1 Characterizing the Methods

Recall that when the vocabulary consists of k unary predicates, these predicates define 2^k mutually exclusive and exhaustive *atoms*, $A_1, \ldots, A_{2^k}$. Every possible world defines a tuple $(p_1, \ldots, p_{2^k})$ where p_i is the proportion of domain individuals that satisfy the atom A_i.

[4] Note that again we consider only the predicate denotations when deciding on a world's equivalence class.

[5] There is no guarantee that these limits exist; in complex cases, they may not. As our examples suggest, in typical cases they do (see (GHK94)).

Given a database KB we can form the set of tuples defined by the set of worlds which satisfy KB; this set can be viewed as the set of proportions consistent with KB. Let $S(KB)$ denote the closure of this set.

We can often find a single point in $S(KB)$ that will characterize the degrees of belief generated by our different methods. In the random-worlds method this is the *maximum entropy* point of $S(KB)$ (see (GHK94; PV89)). In the random-structures method, the characteristic point is the *center of mass* of $S(KB)$. Finally, in the random-propensities method, the characteristic point maximizes the statistical independence of the predicates in the vocabulary. We formalize these latter two characterizations and describe the conditions under which they hold in the full paper.[6] When applicable, the characteristic point determines the degree of belief in φ given KB; we construct a particular probability structure (described also in (GHK94)) whose proportions are exactly those defined by the characteristic point. The probability of φ given KB is exactly the probability of φ given KB in this particular structure.

Suppose that the vocabulary consists only of $\{P, c\}$, and the database KB is simply $\|P(x)\|_x \in [\alpha, \beta]$. What does the above tell us about the degree of belief in $P(c)$ under the three methods? In this case, there are only two atoms, P and $\neg P$, and $S(KB)$ consists of all pairs (p_1, p_2) such that $p_1 \in [\alpha, \beta]$. Since the random-worlds method tries to maximize entropy, it focuses on the pair (p_1, p_2) where p_1 is as close as possible to $1/2$. The random-structures method considers the center of mass of the region of consistent proportions, which is clearly attained when $p_1 = (\alpha+\beta)/2$. Since there is only one predicate in the vocabulary, the "maximum independence" characterization of the random-propensities method gives no useful information here. However, it can be shown that for this vocabulary, the random-propensities method and the random-structures method give the same answer. Thus, we get $\Pr^w_\infty(P(c)|KB) = \gamma$, where $\gamma \in [\alpha, \beta]$ minimizes $|\gamma - \frac{1}{2}|$, and $\Pr^s_\infty(P(c)|KB) = \Pr^p_\infty(P(c)|KB) = \frac{\alpha+\beta}{2}$.[7]

Notice also that we were careful to say that the vocabulary is $\{P, c\}$ here. Suppose the vocabulary were larger, say $\{P, Q, c, d\}$. This change has no impact on the random-worlds and the random-propensities method; we still get the same answers as for the smaller vocabulary. In general, the degree of belief in φ given KB does not depend on the vocabulary for these two methods. As shown in (?), this is not true in the case of the random-structures method. We return to this point in the next section.

[6] The conditions required vary. Roughly speaking, the maximum-entropy characterization of random-worlds almost always works in practice; the center-of-mass technique finds degrees of belief for a smaller class of formulas, although there are few restrictions on KB; maximum-independence works for most formulas, but is not sufficient to handle the fairly common case where $S(KB)$ contains several points that maximize independence equally.

[7] All of our methods give point-valued degrees of belief. In examples like this is may be desirable to allow interval-valued degrees of belief; we defer discussion to the full paper.

3.2 Properties of the Methods

As we mentioned in the introduction, all of our methods share some reasonable properties.

1) Deductive inference: All three methods generalize deductive inference; any fact that follows from the database is given degree of belief 1.

Proposition 1: *If* $\models KB \Rightarrow \varphi$ *then* $\Pr^*_\infty(\varphi|KB) = 1$.

2) Direct inference: All three methods agree with direct inference in noncontroversial cases. To be precise, say the reference class $\mathcal{C}$ is specified by some formula $\psi(x)$; we have statistical information about the proportion of $\mathcal{C}$'s that satisfy some property φ, e.g., the information $\|\varphi(x)|\psi(x)\|_x \in [\alpha, \beta]$. All we know about a constant c is that it belongs to the class $\mathcal{C}$, i.e., we know only $\psi(c)$. In this case we have only one reference class, and direct inference would use the statistics from this class to generate a degree of belief in $\varphi(c)$. In such cases, all three of our methods also reflect the statistics we have for $\mathcal{C}$.

Proposition 2: *Let c be a constant, and let $\varphi(x), \psi(x)$ be formulas that do not mention c. Then* $\Pr^*_\infty(\varphi(c)|\, \|\varphi(x)|\psi(x)\|_x \in [\alpha, \beta] \wedge \psi(c)) \in [\alpha, \beta]$.

Therefore, in the example from the introduction, if the database consists only of **A**, then we will obtain a degree of belief of 0.2 from all three methods.

3) Specific Information: Suppose we have statistics for φ relative to classes $\mathcal{C}_1$ and $\mathcal{C}_2$. If $\mathcal{C}_1$ is more specific, then we generally prefer to use its statistics.

Proposition 3: *Suppose KB has the form* $\|\varphi(x)|\psi_1(x)\|_x \in [\alpha_1, \beta_1] \wedge \|\varphi(x)|\psi_2(x)\|_x \in [\alpha_2, \beta_2] \wedge \psi_1(c) \wedge \forall x(\psi_1(x) \Rightarrow \psi_2(x))$, *where φ, ψ_1, and ψ_2 do not mention c. Then* $\Pr^*_\infty(\varphi(c)|KB) \in [\alpha_1, \beta_1]$.

This result demonstrates that if the knowledge base consists of items **A**–**C** from the introduction, then all three methods generate a degree of belief of 0.25 in *Tall*(*Eric*), preferring the information about the more specific class, Californians.

4) Irrelevant information: Often, databases contain information that appears to be irrelevant to the problem at hand. We usually want the computed degree of belief to be unaffected by this extra information. This turns out to be the case for the random-worlds and the random-propensities methods, but not for the random-structures method. The proposition below formalizes one special case of this phenomenon.

Proposition 4: *Let $\psi(x)$ be a formula not mentioning c or P, let KB be $(\|P(x)|\psi(x)\|_x \in [\alpha, \beta]) \wedge \psi(c)$, and let θ be a formula not mentioning P. Then $\Pr_\infty^w(P(c)|KB) = \Pr_\infty^w(P(c)|KB \wedge \varphi) = \Pr_\infty^p(P(c)|KB) = \Pr_\infty^p(P(c)|KB \wedge \varphi) \in [\alpha, \beta]$.*

This result demonstrates that if our knowledge base consists of items **A** and **C** from the introduction, then we obtain a degree of belief of 0.2 in *Tall*(*Eric*) using either random-worlds or random-propensities; these methods allow us to inherit statistics from superclasses, thus treating subclasses for which we have no special statistical information as irrelevant. In contrast, the random-structures method assigns a degree of belief of 0.5 to *Tall*(*Eric*) in this example. This can be quite reasonable in certain situations, since if the subclass is worthy of a name, it might be special in some way, and our statistics for the superclass might not apply.

5) Sampling: Suppose $\|Q(x)\|_x = \beta$ and $\|P(x)|Q(x)\|_x = \alpha$. Intuitively, here we want to think of β as being small, so that Q defines a small sample of the total domain. We know the proportion of P's in this small sample is α. Can we use this information when the appropriate reference class is the entire domain? In a sense, this is a situation which is dual to the previous one, since the reference class we are interested in is larger than that for which we have statistics (Q). One plausible choice in this case is to use the statistics from the smaller class; i.e., treat it as sample data from which we can induce information relevant to the superset. This is what is done by the random-propensities method. The random-worlds method and the random-structures method enforce a different intuition; since we have no information whatsoever as to the overall proportion of P's satisfying $\neg Q$, we assume by default that it is 1/2. Thus, on a fraction β of the domain, the proportion of P's is α, on the remaining fraction $(1-\beta)$ of the domain, the proportion of P's is 1/2. This says that the proportion of P's is $\alpha\beta + (1-\beta)/2$. Formally:

Proposition 5: *Let KB be $\|P(x)|Q(x)\|_x = \alpha \wedge \|(Q(x)\|_x = \beta$. Then $\Pr_\infty^p(P(c)|KB) = \alpha$ and $\Pr_\infty^w(P(c)|KB) = \Pr_\infty^s(P(c)|KB) = \alpha\beta + (1-\beta)/2$.*

There are reasonable intuitions behind both answers here. The first, as we have already said, corresponds to sampling. For the second, we could argue that since the class Q is sufficiently distinguished to merit a name in our language, it might be dangerous to treat it as a random sample.

These propositions are just a small sample of the patterns of reasoning encountered in practice. But they demonstrate that the issues we raised in the introduction are handled well by our approach. Furthermore, in those cases where the methods differ, they serve to highlight competing intuitions about what the "reasonable inference" is. The fact that our techniques automatically give reasonable answers for these basic problems leads us to believe that our approach is a useful way to attack the problem.

4 Understanding the Alternatives

How do we decide which, if any, of our three techniques is appropriate in a particular situation? We do not have a universal criterion. Nevertheless, as we now show, the different methods make implicit assumptions about language and the structure of the domain. By examining these assumptions we can offer some suggestions as to when one method might be preferable to another.

Recall that random-structures groups isomorphic worlds together, in effect treating the domain elements as indistinguishable. If the elements are distinguishable, random-worlds may be a more appropriate model. We remark that this issue of distinguishability is of crucial importance in statistical physics and quantum mechanics. However, there are situations where it is not as critical. In particular, we show in (?) and in the full paper that, as long as there are "enough" predicates in the vocabulary, the random-worlds method and the random-structures method are essentially equivalent. "Enough" here means "sufficient" to distinguish the elements in the domain; in a domain of size N, it turns out that $3 \log N$ unary predicates suffice. Hence, the difference between distinguishability and indistinguishability can often be explained in terms of the richness of our vocabulary.

The random-propensities method gives the language an even more central role. It assumes that there is information implicit in the choice of predicates. To illustrate this phenomenon, consider the well-known "grue/bleen" paradox (Goo55). A person who has seen many emeralds, all of which were green, might place a high degree of belief in "all emeralds are green." Now suppose that, as well as the concepts green and blue, we also consider "grue"—green before the year 2000, and blue after—and "bleen" (blue before 2000, and green after). All the evidence for emeralds being green that anyone has seen is also evidence for emeralds being grue, but no one believes that "all emeralds are grue." Inferring "grueness" seems unintuitive. This suggests that inductive reasoning must go beyond logical expressiveness to use judgements about which predicates are most "natural."

This intuition is captured by the random-propensities approach. Consider the following simplified version of the "grue/bleen" paradox. Let the vocabulary Φ consist of two unary predicates, G (for "green") and B (for "before the year 2000"), and a constant c. We identify "blue" with "not green" and so take "Grue" to be $(G \wedge B) \vee (\neg G \wedge \neg B)$.[8] The domain elements are observations of emeralds.

[8] This does not capture the full complexity of the paradox, since the true definition of "grue" requires the emerald to change color over time.

If our database KB is $||G(x)|B(x)||_x = 1$, then it can be shown than $\Pr^p_\infty(G(c)|KB \land \neg B(c)) = 1$ and $\Pr^p_\infty(\mathit{Grue}(c)|KB \land \neg B(c)) = 0$. That is, the method "learns" natural concepts such as "greenness" and not unnatural ones such as "grueness". By way of contrast, $\Pr^w_\infty(G(c)|KB\land\neg B(c)) = \Pr^w_\infty(\mathit{Grue}(c)|KB\land\neg B(c)) = \Pr^s_\infty(G(c)|KB\land\neg B(c)) = \Pr^s_\infty(\mathit{Grue}(c)|KB\land\neg B(c)) = 0.5$. To understand this phenomenon, recall that the random-worlds and random-structures methods treat "grue" and "green" symmetrically; they are both the union of two atoms. The random-propensities method, on the other hand, gives "green" special status as a predicate in the vocabulary.

The importance of the choice of predicates in the random-propensities approach can be partially explained in terms of an important connection between it and the random-worlds approach. Suppose we are interested in the predicate *Tall*. A standard approach to defining the semantics of *Tall* is to order individuals according to height, and choose a cutoff point such that an individual is considered "tall" exactly if he is taller than the cutoff. It turns out that if we add this implicit information about the meaning of *Tall* to the knowledge base, and use the random-worlds approach, we obtain the random-propensities approach. Intuitively, the location of the cutoff point reflects the propensity of a random individual to be tall. Many predicates can be interpreted in a similar fashion, and random-propensities might be an appropriate method in these cases. However, many problems will include different kinds of predicates, requiring different treatment. Therefore, in most practical situations, a combination of the methods would almost certainly be used.

In conclusion, we believe that we have offered a new approach to the problem of computing degrees of belief from statistics. Our approach relies on notions that seem to be much more fundamental than the traditional notion of "choosing the right reference class." As should be clear from our examples, none of the three methods discussed here is universally applicable. Instead, they seem to represent genuine alternative intuitions applicable to different situations. We feel that the elucidation of these alternative intuitions is in itself a useful contribution.

References

F. Bacchus. *Representing and Reasoning with Probabilistic Knowledge.* MIT Press, Cambridge, Mass., 1990.

R. Carnap. *Logical Foundations of Probability.* University of Chicago Press, Chicago, 1950.

R. Carnap. *The Continuum of Inductive Methods.* University of Chicago Press, Chicago, 1952.

A. J. Grove, J. Y. Halpern, and D. Koller. Asymptotic conditional probabilities: the non-unary case. Research Report RJ 9564, IBM, 1993. To appear in *Journal of Symbolic Logic.*

A. J. Grove, J. Y. Halpern, and D. Koller. Random worlds and maximum entropy. *Journal of A.I. Research*, 2:33–88, 1994.

N. Goodman. *Fact, fiction, and forecast*, chapter III. Harvard University Press, 1955.

S. D. Goodwin. Second order direct inference: A reference class selection policy. *International Journal of Expert Systems: Research and Applications*, 5(3):185–210, 1992.

J. Y. Halpern. An analysis of first-order logics of probability. *Artificial Intelligence*, 46:311–350, 1990.

E. T. Jaynes. Where do we stand on maximum entropy? In R. D. Levine and M. Tribus, editors, *The Maximum Entropy Formalism*, pages 15–118. MIT Press, Cambridge, Mass., 1978.

W. E. Johnson. Probability: The deductive and inductive problems. *Mind*, 41(164):409–423, 1932.

J. M. Keynes. *A Treatise on Probability.* Macmillan, London, 1921.

J. von Kries. *Die Principien der Wahrscheinlichkeitsrechnung und Rational Expectation.* Freiburg, 1886.

H. E Kyburg, Jr. *Probability and the Logic of Rational Belief.* Wesleyan University Press, Middletown, Connecticut, 1961.

H. E. Kyburg, Jr. *The Logical Foundations of Statistical Inference.* Reidel, Dordrecht, Netherlands, 1974.

H. E. Kyburg, Jr. The reference class. *Philosophy of Science*, 50(3):374–397, 1983.

P. S. de Laplace. *Essai Philosophique sur les Probabilités.* 1820. English translation is *Philosophical Essay on Probabilities*, Dover Publications, New York, 1951.

I. Levi. *The Enterprise of Knowledge.* MIT Press, Cambridge, MA, 1980.

J. L. Pollock. *Nomic Probabilities and the Foundations of Induction.* Oxford University Press, Oxford, U.K., 1990.

J. B. Paris and A. Vencovska. On the applicability of maximum entropy to inexact reasoning. *International Journal of Approximate Reasoning*, 3:1–34, 1989.

J. B. Paris and A. Vencovska. A method for updating justifying minimum cross entropy. *International Journal of Approximate Reasoning*, 7:1–18, 1992.

H. Reichenbach. *The Theory of Probability.* University of California Press, Berkeley, 1949. This is a translation and revision of the German edition, published as *Wahrscheinlichkeitslehre*, in 1935.

W. Salmon. *Statistical Explanation and Statistical Relevance.* University of Pittsburgh Press, Pittsburgh, 1971.