Reasoning Defeasibly About Probabilities[1]

John L. Pollock
Department of Philosophy
University of Arizona
Tucson, Arizona 85721
pollock@arizona.edu
http://www.u.arizona.edu/~pollock

Abstract

In concrete applications of probability, statistical investigation gives us knowledge of some probabilities, but we generally want to know many others that are not directly revealed by our data. For instance, we may know prob(P/Q) (the probability of P given Q) and prob(P/R), but what we really want is prob($P/Q\&R$), and we may not have the data required to assess that directly. The probability calculus is of no help here. Given prob(P/Q) and prob(P/R), it is consistent with the probability calculus for prob($P/Q\&R$) to have any value between 0 and 1. Is there any way to make a reasonable estimate of the value of prob($P/Q\&R$)?

A related problem occurs when probability practitioners adopt undefended assumptions of statistical independence simply on the basis of not seeing any connection between two propositions. This is common practice, but its justification has eluded probability theorists, and researchers are typically apologetic about making such assumptions. Is there any way to defend the practice?

This paper shows that on a certain conception of probability — nomic probability — there are principles of "probable probabilities" that license inferences of the above sort. These are principles telling us that although certain inferences from probabilities to probabilities are not deductively valid, nevertheless the second-order probability of their yielding correct results is 1. This makes it defeasibly reasonable to make the inferences. Thus I argue that it is defeasibly reasonable to assume statistical independence when we have no information to the contrary. And I show that there is a function $Y(r,s,a)$ such that if prob(P/Q) = r, prob(P/R) = s, and prob(P/U) = a (where U is our background knowledge) then it is defeasibly reasonable to expect that prob($P/Q\&R$) = $Y(r,s,a)$. Numerous other defeasible inferences are licensed by similar principles of probable probabilities. This has the potential to greatly enhance the usefulness of probabilities in practical application.

1. The Problem of Sparse Probability Knowledge

The use of probabilities is ubiquitous in philosophy, science, engineering, artificial intelligence, economics, and many other disciplines. It is generally supposed that the logical and mathematical structure of probabilities is well understood, and completely characterized by the probability calculus. The probability calculus is typically identified with some form of Kolmogoroff's axioms, often supplemented with an axiom of countable additivity. Mathematical probability theory is a mature subdiscipline of mathematics based upon these axioms, and forms the mathematical basis for most applications of probabilities in the sciences.

There is, however, a problem with the supposition that this is all there is to the logical and mathematical structure of probabilities. The uninitiated often suppose that if we know a few basic probabilities, we can compute the values of many others just by applying the probability calculus. Thus it might be supposed that familiar sorts of statistical inference provide us with our basic knowledge of probabilities, and then appeal to the probability calculus enables us to compute other previously unknown probabilities. The picture is of a kind of foundations theory of the epistemology of probability, with the probability calculus providing the inference engine that

[1] This work was supported by NSF grant no. IIS-0412791.

enables us to get beyond whatever probabilities are discovered by direct statistical investigation.

Regrettably, this simple image of the epistemology of probability cannot be correct. The difficulty is that the probability calculus is not nearly so powerful as the uninitiated suppose. If we know the probabilities of some basic propositions P, Q, R, S, … , it is rare that we will be able to compute, just by appeal to the probability calculus, a unique value for the probability of some logical compound like $((P \,\&\, Q) \vee (R \,\&\, S))$. To illustrate, suppose we know that $\text{PROB}(P) = .7$ and $\text{PROB}(Q) = .6$. What can we conclude about $\text{PROB}(P \,\&\, Q)$? All the probability calculus enables us to infer is that $.3 \leq \text{PROB}(P \,\&\, Q) \leq .6$. That does not tell us much. Similarly, all we can conclude about $\text{PROB}(P \vee Q)$ is that $.7 \leq \text{PROB}(P \vee Q) \leq 1.0$. In general, the probability calculus imposes constraints on the probabilities of logical compounds, but it falls far short of enabling us to compute unique values.

Unless we come to a problem already knowing a great deal about the relevant probabilities, the probability calculus will not enable us to compute the values of unknown probabilities that subsequently become of interest to us. Suppose a problem is described by logical compounds of a set of simple propositions $P_1,\ldots,P_n$. Then to be able to compute the probabilities of all logical compounds of these simple propositions, what we must generally know is the probabilities of every conjunction of the form $\text{PROB}((\sim)P_1\&\ldots\&(\sim)P_n)$. The tildes enclosed in parentheses can be either present or absent. These n-fold conjunctions are called *Boolean conjunctions*, and jointly they constitute a "partition". Given fewer than all but one of them, the only constraint the probability calculus imposes on the probabilities of the remaining Boolean conjunctions is that the sum of all of them must be 1. Together, the probabilities of all the Boolean conjunctions determine a complete "probability distribution" — an assignment of unique probabilities to every logical compound of the simple propositions.

In theoretical accounts of the use of probabilities in any discipline, it is generally assumed that we come to a problem equipped with a complete probability distribution. However, in real life this assumption is totally unrealistic. In general, given n simple propositions, there will be 2^n logically independent probabilities of Boolean conjunctions. As Gilbert Harman (1986) observed years ago, for a rather small number of simple propositions, there is a completely intractable number of logically independent probabilities. For example, given just 300 simple propositions, a grossly inadequate number for describing many real-life problems, there will be 2^{300} logically independent probabilities of Boolean conjunctions. 2^{300} is approximately equal to 10^{90}. To illustrate what an immense number this is, recent estimates of the number of elementary particles in the universe put it at $10^{80} - 10^{85}$. Thus to know the probabilities of all the Boolean conjunctions, we would have to know 5 – 10 orders of magnitude more logically independent probabilities than the number of elementary particles in the universe.

Lest one think this is an unrealistic problem, consider a simple example. Pollock (2006) describes a challenge problem for AI planners. This problem generalizes Kushmerick, Hanks and Weld's (1995) "slippery gripper" problem. We are presented with a table on which there are 300 numbered blocks, and a panel of correspondingly numbered buttons. Pushing a button activates a robot arm which attempts to pick up the corresponding block and remove it from the table. We get 100 dollars for each block that is removed. Pushing a button costs two dollars. The hitch is that half of the blocks are greasy. If a block is not greasy, pushing the button will result in its being removed from the table with probability 1.0, but if it is greasy the probability is only 0.01. We are given exactly 300 opportunities to either push a button or do nothing. Between button pushes, we are given the opportunity to look at the table, which costs one dollar. Looking will reveal what blocks are still on the table, but will not reveal directly whether a block is greasy. What should we do? Humans find this problem terribly easy. An informal survey reveals that most people quickly produce the optimal plan: push each button once, and don't bother to look at the table. But when Pollock (2006) surveyed existing AI planners, most could not even encode this problem, much less solve it. The difficulty is that there are too many logically independent probabilities. For every subset K of the 300 blocks, let $p_{K,i}$ be the probability that, when K is the set of blocks on the table, block i is still on the table after the button corresponding to block i is pushed. There are 2^{300} choices of K, so there are more than 2^{300} probabilities $p_{K,i}$ such that $i \in K$. Furthermore, none of them can be derived from any of the others. Thus they must each be encoded separately in describing a complete probability distribution for the problem. It seems to be impossible for a real cognitive agent to encode such a probability distribution.

Although we humans cannot encode a complete probability distribution for the preceding problem, we can deal with problems like the slippery blocks problem. How do we do that? It is, apparently, computationally impossible for the the requisite probabilities to be stored in us from

the start, so they must be produced one at a time as we need them. If they are produced as we need them, there must be some kind of inference mechanism that has the credentials to produce rationally acceptable estimates. We have seen that, unless we begin with more information than it is computationally possible for us to store, we cannot derive the new probability estimates from previously accepted probabilities by way of the probability calculus. So there must be some other rational inference procedures that enable us to generate new probability estimates that do not follow logically, via the probability calculus, from prior probability estimates. What might these rational inference procedures be?

I will call this *the problem of sparse probability knowledge*. It is computationally impossible for us to store explicit knowledge of a complete probability distribution. At any given time, our knowledge of probabilities is worse than just incomplete. The set of probabilities we know is many orders of magnitude smaller than the set of all true probabilities. How then can we be as successful as we are in applying probability to real-world problems?

It is noteworthy that in applying probabilities to concrete problems, probability practitioners commonly adopt undefended assumptions of statistical independence. The probabilities PROB(P) and PROB(Q) are *statistically independent* iff PROB(P&Q) = PROB(P)·PROB(Q). An equivalent definition is that PROB(P/Q) = PROB(P). In the practical use of probabilities it is almost universally assumed, often apologetically, that probabilities are independent unless we have some reason for thinking otherwise. In most real-world applications of probabilities, if we did not make such assumptions about independence we would not be able to compute any of the complex probabilities that interest us. Imagine a case in which we know that the probability is .3 of a Xian (a fictional Chinese car) having a defective door lock if it has power door locks and was manufactured in a certain plant, whereas the probability of its having a defective door lock otherwise is only .01. We also know that the probability of a Xian being manufactured in that plant is .33, and the probability of a Xian having power door locks is .85. If we know nothing else of relevance, we will normally assume that whether the car has power door locks is statistically independent of whether it was manufactured in that plant, and so compute

prob(power-locks & plant) = .33×.85=.28.

Then we can compute the general probability of a Xian having defective door locks:

prob(defect) = prob(defect/power-locks & plant)·prob(power-locks & plant)
+ prob(defect/~(power-locks & plant))·(1–prob(power-locks & plant))
= .3×.28 + .01×(1–.28) = .09.

We could not perform this, or similar computations, without the assumption of independence.

The independence assumption is a defeasible assumption, because obviously we can discover that conditions we thought were independent are unexpectedly correlated. The probability calculus can give us only necessary truths about probabilities, so the justification of such a defeasible assumption must have some other source.

If we have a problem in which we can assume that most propositions are statistically independent of one another, there are compact techniques for storing complete probability distributions using what are called "Bayesian nets" (Pearl 1988). The use of Bayesian nets allow us to explicitly store just that subset of probabilities that cannot be derived from each other by assuming statistical independence, and provides an efficient inference mechanism for recovering derivable probabilities from them. However, this is not the entire solution to the problem of sparse probability knowledge, because in the slippery blocks problem, none of the probabilities $p_{K,i}$ can be derived from others, so they would all have to be encoded separately in a Bayesian net, and that would make the Bayesian net impossibly large.

I will argue that a defeasible assumption of statistical independence is just the tip of the iceberg. There are multitudes of defeasible inferences that we can make about probabilities, and a very rich mathematical theory grounding them. It is these defeasible inferences that enable us to make practical use of probabilities without being able to deduce everything we need via the probability calculus. I will argue that, on a certain conception of probability, there are mathematically derivable second-order probabilities to the effect that various inferences about first-order probabilities, although not deductively valid, will nonetheless produce correct conclusions with probability 1, and this makes it reasonable to accept these inferences defeasibly. The second-order principles are

principles of *probable probabilities*.

2. Two Kinds of Probability

No doubt the currently most popular theory of the foundations of probability is the subjectivist theory due originally to Ramsey and Savage, and developed at length by many more recent scholars. However, my solution to the problem of sparse probability knowledge requires that we start with objective probabilities. Historically, there have been two general approaches to probability theory. What I will call *generic probabilities*[2] are general probabilities, relating properties or relations. The generic probability of *an A* being a *B* is not about any particular *A,* but rather about the *property* of being an *A*. In this respect, its logical form is the same as that of relative frequencies. I write generic probabilities using lower case "prob" and free variables: prob(Bx/Ax). For example, we can talk about the probability of an adult male of Slavic descent being lactose intolerant. This is not about any particular person — it expresses a relationship between the property of being an adult male of Slavic descent and the property of being lactose intolerant. Most forms of statistical inference or statistical induction are most naturally viewed as giving us information about generic probabilities. On the other hand, for many purposes we are more interested in propositions that are about particular persons, or more generally, about specific matters of fact. For example, in deciding how to treat Herman, an adult male of Slavic descent, his doctor may want to know the probability that Herman is lactose intolerant. This illustrates the need for a kind of probability that attaches to propositions rather than relating properties and relations. These are sometimes called "single case probabilities", although that terminology is not very good because such probabilities can attach to propositions of any logical form. For example, we can ask how probable it is that there are no human beings over the age of 130. In the past, I called these "definite probabilities", but now I will refer to them as *singular probabilities*.

The distinction between singular and generic probabilities is commonly overlooked by contemporary probability theorists, perhaps because of the popularity of subjective probability (which has no way to make sense of generic probabilities). But most objective approaches to probability tie probabilities to relative frequencies in some essential way, and the resulting probabilities have the same logical form as the relative frequencies. That is, they are generic probabilities. The simplest theories identify generic probabilities with relative frequencies (Russell 1948; Braithwaite 1953; Kyburg 1961, 1974; Sklar 1970, 1973).[3] The simplest objection to such "finite frequency theories" is that we often make probability judgments that diverge from relative frequencies. For example, we can talk about a coin being fair (and so the generic probability of a flip landing heads is 0.5) even when it is flipped only once and then destroyed (in which case the relative frequency is either 1 or 0). For understanding such generic probabilities, we need a notion of probability that talks about *possible* instances of properties as well as actual instances. Theories of this sort are sometimes called "hypothetical frequency theories". C. S. Peirce was perhaps the first to make a suggestion of this sort. Similarly, the statistician R. A. Fisher, regarded by many as "the father of modern statistics", identified probabilities with ratios in a "hypothetical infinite population, of which the actual data is regarded as constituting a random sample" (1922, p. 311). Karl Popper (1956, 1957, and 1959) endorsed a theory along these lines and called the resulting probabilities *propensities*. Henry Kyburg (1974a) was the first to construct a precise version of this theory (although he did not endorse the theory), and it is to him that we owe the name "hypothetical frequency theories". Kyburg (1974a) also insisted that von Mises should also be considered a hypothetical frequentist. There are obvious difficulties for spelling out the details of a hypothetical frequency theory. More recent attempts to formulate precise versions of what might be regarded as hypothetical frequency theories are van Fraassen (1981), Bacchus (1990), Halpern (1990), Pollock (1990), Bacchus et al (1996). I will take my jumping-off point to be the theory of Pollock (1990), which I will sketch briefly in section three.

After brief thought, most philosophers find the distinction between singular and generic probabilities intuitively clear. However, this is a distinction that sometimes puzzles probability theorists many of whom have been raised on an exclusive diet of singular probabilities. They are sometimes tempted to confuse generic probabilities with probability distributions over random

[2] In the past, I followed Jackson and Pargetter 1973 in calling these "indefinite probabilities", but I never liked that terminology.

[3] William Kneale (1949) traces the frequency theory to R. L. Ellis, writing in the 1840's, and John Venn (1888) and C. S. Peirce in the 1880's and 1890's.

variables. Although historically most theories of objective probability were theories of generic probability, mathematical probability theory tends to focus exclusively on singular probabilities. When mathematicians talk about variables in connection with probability, they usually mean "random variables", which are not variables at all but functions assigning values to the different members of a population. Generic probabilities have single numbers as their values. Probability distributions over random variables are just what their name implies — distributions of singular probabilities rather than single numbers.

It has always been acknowledged that for practical decision-making we need singular probabilities rather than generic probabilities. For example, in deciding whether to trust the door locks on my Xian, I want to know the probability of *its* having defective locks, not the probability of Xians in general having defective locks. So theories that take generic probabilities as basic need a way of deriving singular probabilities from them. Theories of how to do this are theories of *direct inference*. Theories of objective generic probability propose that statistical inference gives us knowledge of generic probabilities, and then direct inference gives us knowledge of singular probabilities. Reichenbach (1949) pioneered the theory of direct inference. The basic idea is that if we want to know the singular probability PROB(Fa), we look for the narrowest reference class (or reference property) G such that we know the generic probability prob(Fx/Gx) and we know Ga, and then we identify PROB(Fa) with prob(Fx/Gx). For example, actuarial reasoning aimed at setting insurance rates proceeds in roughly this fashion. Kyburg (1974) was the first to attempt to provide firm logical foundations for direct inference. Pollock (1990) took that as its starting point and constructed a modified theory with a more epistemological orientation. The present paper builds upon some of the basic ideas of the latter.

The appeal to generic probabilities and direct inference has seemed promising for avoiding the computational difficulties attendant on the need for a complete probability distribution. Instead of assuming that we come to a problem with an antecedently given complete probability distribution, one can assume more realistically that we come to the problem with some limited knowledge of generic probabilities and then infer singular probabilities from the latter as we need them. For example, I had no difficulty giving a description of the probabilities involved in the slippery blocks problem, but I did that by giving an informal description of the generic probabilities rather than the singular probabilities. We described it by reporting that the generic probability prob(Gx/Bx) of a block being greasy is .5, and the generic probability prob($\sim Tx(s{+}1)/Txs$ & Pxs & Gx) of a block being successfully removed from the table at step s if it is greasy is .01, but prob($\sim Tx(s{+}1)/Txs$ & Pxs & $\sim Gx$) = 1.0. We implicitly assumed that prob($\sim Tx(s{+}1)/\sim Txs$) = 1. These probabilities completely describe the problem. For solving the decision-theoretic planning problem, we need singular probabilities rather than generic probabilities, but one might hope that these can be recovered by direct inference from this small set of generic probabilities as they are needed.

Unfortunately, I do not think that this hope will be realized. The appeal to generic probabilities and direct inference helps a bit with the problem of sparse probability knowledge, but it falls short of constituting a complete solution. The difficulty is that the problem recurs at the level of generic probabilities. Direct statistical investigation will apprise us of the values of some generic probabilities, and then others can be derived by appeal to the probability calculus. But just as for singular probabilities, the probability calculus is a weak crutch. We will rarely be able to derive more than rather broad constraints on unknown probabilities. A simple illustration of this difficulty arises when we know that prob(Ax/Bx) = r and prob(Ax/Cx) = s, where $r \neq s$, and we know both that Ba and Ca. What should we conclude about the value of PROB(Aa)? Direct inference gives us defeasible reasons for drawing the conflicting conclusions that PROB(Aa) = r and PROB(Aa) = s, and standard theories of direct inference give us no way to resolve the conflict, so they end up telling us that there is no conclusion we can justifiably draw about the value of PROB(Aa). Is this reasonable? Suppose we have two unrelated diagnostic tests for some rare disease, and Bernard tests positive on both tests. Intuitively, it seems this should make it more probable that Bernard has the disease than if we only have the results of one of the tests. This suggests that, given the values of prob(Ax/Bx) and prob(Ax/Cx), there ought to be something useful we can say about the value of prob(Ax/Bx&Cx), and then we can apply direct inference to the latter to compute the singular probability that Bernard has the disease. Existing theories give us no way to do this, and the probability calculus imposes no constraint at all on the value of prob(Ax/Bx&Cx).

I believe that standard theories of direct inference are much too weak to solve the problem of sparse probability knowledge. What I will argue in this paper is that new mathematical results, coupled with ideas from the theory of nomic probability introduced in Pollock (1990), provide the justification for a wide range of new principles supporting defeasible inferences about the

expectable values of unknown probabilities. These principles include familiar-looking principles of direct inference, but they include many new principles as well. For example, among them is a principle enabling us to defeasibly estimate the probability of Bernard having the disease when he tests positive on both tests. I believe that this broad collection of new defeasible inference schemes provides the solution to the problem of sparse probability knowledge and explains how probabilities can be truly useful even when we are massively ignorant about most of them.

3. Nomic Probability

Pollock (1990) developed a possible worlds semantics for objective generic probabilities,[4] and I will take that as my starting point for the present theory of probable probabilities. The proposal was that we can identify the *nomic probability* prob(Fx/Gx) with the proportion of physically possible G's that are F's. A *physically possible* G is defined to be an ordered pair $\langle w,x\rangle$ such that w is a physically possible world (one compatible with all of the physical laws) and x has the property G at w. Let us define the *subproperty relation* as follows:

$F \preccurlyeq G$ iff it is physically necessary (follows from true physical laws) that $(\forall x)(Fx \rightarrow Gx)$.

$F \cong G$ iff it is physically necessary (follows from true physical laws) that $(\forall x)(Fx \leftrightarrow Gx)$.

We can think of the subproperty relation as a kind of nomic entailment relation (holding between properties rather than propositions). More generally, F and G can have any number of free variables (not necessarily the same number), in which case $F \preccurlyeq G$ iff the universal closure of ($F \rightarrow G$) is physically necessary.

Given a suitable proportion function ρ, we could stipulate that, where $\mathfrak{F}$ and $\mathfrak{G}$ are the sets of physically possible F's and G's respectively:

$$\text{prob}_x(Fx/Gx) = \rho(\mathfrak{F},\mathfrak{G}).^5$$

However, it is unlikely that we can pick out the right proportion function without appealing to prob itself, so the postulate is simply that *there is* some proportion function related to prob as above. This is merely taken to tell us something about the formal properties of prob. Rather than axiomatizing prob directly, it turns out to be more convenient to adopt axioms for the proportion function. Proportion functions are a generalization of measure functions, studied in mathematics in measure theory. Pollock (1990) showed that, given the assumptions adopted there, ρ and prob are interdefinable, so the same empirical considerations that enable us to evaluate prob inductively also determine ρ.

Note that prob_x is a variable-binding operator, binding the variable x. When there is no danger of confusion, I will omit the subscript "x", but sometimes we will want to quantify into probability contexts, in which case it will be important to distinguish between the variables bound by "prob" and those that are left free. To simplify expressions, I will often omit the variables, writing "prob(F/G)" for "prob(Fx/Gx)" when no confusion will result.

It is often convenient to write proportions in the same logical form as probabilities, so where φ and θ are open formulas with free variable x, let $\rho_x(\varphi/\theta) = \rho(\{x \mid \varphi \,\&\, \theta\},\{x \mid \theta\})$. Note that ρ_x is a variable-binding operator, binding the variable x. Again, when there is no danger of confusion, I will typically omit the subscript "x".

I will make three classes of assumptions about the proportion function. Let $\#X$ be the cardinality of a set X. If Y is finite, I assume:

$$\rho(X,Y) = \frac{\#X \cap Y}{\#Y}.$$

[4] Somewhat similar semantics were proposed by Halpern (1990) and Bacchus et al (1996).

[5] Probabilities relating n-place relations are treated similarly. I will generally just write the one-variable versions of various principles, but they generalize to n-variable versions in the obvious way.

However, for present purposes the proportion function is most useful in talking about proportions among infinite sets. The sets $\mathfrak{F}$ and $\mathfrak{G}$ will invariably be infinite, if for no other reason than that there are infinitely many physically possible worlds in which there are *F*'*s* and *G*'*s*.

My second set of assumptions is that the standard axioms for conditional probabilities hold for proportions. These axioms automatically hold for relative frequencies among finite sets, so the assumption is just that they also hold for proportions among infinite sets.

That further assumptions are needed derives from the fact that the standard probability calculus is a calculus of singular probabilities rather than generic probabilities. A calculus of generic probabilities is related to the calculus of singular probabilities in a manner roughly analogous to the relationship between the predicate calculus and the propositional calculus. Thus we get some principles pertaining specifically to relations that hold for generic probabilities but cannot even be formulated in the standard probability calculus. For instance, Pollock (1990) endorsed the following two principles:

Individuals:
prob(*Fxy*/*Gxy* & *y* = *a*) = prob(*Fxa*/*Gxa*)

PPROB:
prob(*Fx*/*Gx* & prob(*Fx*/*Gx*) = *r*) = *r*.

I will not assume either of these principles in this paper, but I mention them just to illustrate that there are reasonable-seeming principles governing generic probabilities that are not even well formed in the standard probability calculus.

What I do need in the present paper is three assumptions about proportions that go beyond merely imposing the standard axioms for the probability calculus. The three assumptions I will make are:

Finite Set Principle:
For any set *B*, $N > 0$, and open formula Φ,

$$\rho_X\big(\Phi(X) \,/\, X \subseteq B \;\&\; \#X = N\big) =$$
$$\rho_{x_1,\ldots,x_N}\big(\Phi(\{x_1,\ldots,x_N\}) \,/\, x_1,\ldots,x_N \text{ are pairwise distinct } \&\; x_1,\ldots,x_N \in B\big).$$

Projection Principle:
If $0 \le p,q \le 1$ and $(\forall y)(Gy \rightarrow \rho_x(Fx/\, Rxy) \in [p,q])$, then $\rho_{x,y}(Fx/\, Rxy \;\&\; Gy) \in [p,q]$.[6]

Crossproduct Principle:
If *C* and *D* are nonempty, $\rho\big(A \times B, C \times D\big) = \rho(A,C) \cdot \rho(B,D)$.

Note that these three principles are all theorems of elementary set theory when the sets in question are finite. For instance, to illustrate the finite case of the projection principle, let *F* be "*x* is an even non-negative integer", let *Rxy* be "*x* and *y* are non-negative integers and $x \le y$", and let *Gy* be "$y \in \{5,6,7\}$. Then $\rho_x(Fx/Rx5) = \rho_x(Fx/Rx7) = ½$ and $\rho_x(Fx/Rx5) = 4/7$. Thus $(\forall y)(Gy \rightarrow \rho_x(Fx/\, Rxy) \in [4/7,1/2])$. And $\rho_{x,y}(Fx/\, Rxy \;\&\; Gy) = 11/21 \in [4/7,1/2]$.

The crossproduct principle holds for finite sets because $\#(A \times B) = (\#A) \cdot (\#B)$, and hence

$$\rho\big(A \times B, C \times D\big) = \frac{\#((A \times B) \cap (C \times D))}{\#(C \times D)} = \frac{\#((A \cap C) \times (B \cap D))}{\#(C \times D)}$$
$$= \frac{\#(A \cap C) \cdot \#(B \cap D)}{\#C \cdot \#D} = \frac{\#(A \cap C)}{\#C} \cdot \frac{\#(B \cap D)}{\#D} = \rho(A,C) \cdot \rho(B,D).$$

My assumption is simply that ρ continues to have these algebraic properties even when applied to infinite sets. I take it that this is a fairly conservative set of assumptions.

I often hear the objection that in affirming the Crossproduct Principle, I must be making a

[6] Note that this is a different (and more conservative) principle than the one called "Projection" in Pollock (1990).

hidden assumption of statistical independence. However, that is to confuse proportions with probabilities. The Crossproduct Principle is about proportions — not probabilities. For finite sets, proportions are computed by simply counting members and computing ratios of cardinalities. It makes no sense to talk about statistical independence in this context. For infinite sets we cannot just count members any more, but the algebra is the same. It is because the algebra of proportions is simpler than the algebra of probabilities that it is useful to axiomatize nomic probabilities indirectly by adopting axioms for proportions.

The preceding amounts to a "realistic possible worlds semantics" for nomic probability. A realistic possible world semantics takes possible worlds, objects in possible world, properties, relations, and propositions as basic. There are many different approaches to how these concepts are to be understood, but for the most part it makes no different to the present paper what approach is taken. All that my mathematics requires is that propositions, properties, and relations are closed under various operations that everyone grants them to be closed under. As long as the proportion function satisfies my postulates, the mathematical results follow.

To be contrasted with realistic possible world semantics are model theoretic semantics (e.g., Halpern 1990, Bacchus et al 1996). A model-theoretic approach constructs set-theoretic models and interprets formal languages in terms of them. It it mathematically precise, but it is only as good as the model theory. You can construct model theories that validate almost anything. If your objective is to use model theory to illuminate pre-analytic concepts, it is important to justify the model theory. Model theoretic approaches to modalities rely upon formal analogues to possible worlds, but it has become apparent that the formal analogues are not precise. The simplest analogue generates Carnap's modal logic, which no one thinks is right. To get even S5 one must make basically ad hoc moves regarding the accessibility relation. This is a topic I discussed at great length in my (1984a). What I argued was that to get the model theory right, you have to start with a realistic possible worlds semantics and justify it. The appeal to model theory cannot replace the appeal to a realistic possible world semantics.

Pollock (1990) derived the entire epistemological theory of nomic probability from a single epistemological principle coupled with a mathematical theory that amounts to a calculus of nomic probabilities. The single epistemological principle that underlies probabilistic reasoning is the *statistical syllogism*, which can be formulated as follows:

Statistical Syllogism:

If F is projectible with respect to G and $r > 0.5$, then $\ulcorner Gc$ & $\text{prob}(F/G) \geq r \urcorner$ is a defeasible reason for $\ulcorner Fc \urcorner$, the strength of the reason being a monotonic increasing function of r.

I take it that the statistical syllogism is a very intuitive principle, and it is clear that we employ it constantly in our everyday reasoning. For example, suppose you read in the newspaper that George Bush is visiting Guatemala, and you believe what you read. What justifies your belief? No one believes that everything printed in the newspaper is true. What you believe is that certain kinds of reports published in certain kinds of newspapers tend to be true, and this report is of that kind. It is the statistical syllogism that justifies your belief.

The projectibility constraint in the statistical syllogism is the familiar projectibility constraint on inductive reasoning, first noted by Goodman (1955). One might wonder what it is doing in the statistical syllogism. But it was argued in (Pollock 1990), on the strength of what were taken to be intuitively compelling examples, that the statistical syllogism must be so constrained. Furthermore, it was shown that without a projectibility constraint, the statistical syllogism is self-defeating, because for any intuitively correct application of the statistical syllogism it is possible to construct a conflicting (but unintuitive) application to a contrary conclusion. This is the same problem that Goodman first noted in connection with induction. Pollock (1990) then went on to argue that the projectibility constraint on induction derives from that on the statistical syllogism.

The projectibility constraint is important, but also problematic because no one has a good analysis of it. I will not discuss it further here. I will just assume, without argument, that the second-order probabilities employed below in the theory of probable probabilities satisfy the projectibility constraint, and hence can be used in the statistical syllogism.

The statistical syllogism is a defeasible inference scheme, so it is subject to defeat. I believe that the only primitive (underived) principle of defeat required for the statistical syllogism is that of subproperty defeat:

Subproperty Defeat for the Statistical Syllogism:

If H is projectible with respect to G, then $\ulcorner Hc$ & $\text{prob}(F/G\&H) < \text{prob}(F/G)\urcorner$ is an undercutting defeater for the inference by the statistical syllogism from $\ulcorner Gc$ & $\text{prob}(F/G) \geq r\urcorner$ to $\ulcorner Fc\urcorner$.[7]

In other words, information about c that lowers the probability of its being F constitutes a defeater. Note that if $\text{prob}(Fx/G\&H)$ is high, one may still be able to make a weaker inference to the conclusion that Fc, but from the distinct premise $\ulcorner Gc$ & $\text{prob}(F/G\&H) = s\urcorner$.

Pollock (1990) argued that we need additional defeaters for the statistical syllogism besides subproperty defeaters, formulated several candidates for such defeaters. But one of the conclusions of the research described in this paper is that the additional defeaters can all be viewed as derived defeaters, with subproperty defeaters being the only primitive defeaters for the statistical syllogism.

4. Indifference

Principles of probable probabilities are derived from combinatorial theorems about proportions in finite sets. I will begin with a very simple principle that is in fact not very useful, but will serve as a template for the discussion of more useful principles.

Suppose we have a set of 10,000,000 objects. I announce that I am going to select a subset, and ask you how many members it will have. Most people will protest that there is no way to answer this question. It could have any number of members from 0 to 10,000,000. However, if you answer, "Approximately 5,000,000," you will almost certainly be right. This is because, although there are subsets of all sizes from 0 to 10,000,000, there are many more subsets whose sizes are approximately 5,000,000 than there are of any other size. In fact, 99% of the subsets have cardinalities differing from 5,000,000 by less than .08%. If we let "$x \approx_{\delta} y$" mean "the difference between x and y is less than or equal to δ", the general theorem is:

Finite Indifference Principle:

For every $\varepsilon,\delta > 0$ there is an N such that if U is finite and $\#U > N$ then

$$\rho_X\left(\rho(X,U) \approx_{\delta} 0.5 \;/\; X \subseteq U\right) \geq 1-\varepsilon.$$

In other words, the proportion of subsets of U which are such that $\rho(X,U)$ is approximately equal to .5, to any given degree of approximation, goes to 1 as the size of U goes to infinity. To see why this is true, suppose $\#U = n$. If $r \leq n$, the number of r-membered subsets of U is $C(n,r) = \dfrac{n!}{r!(n-r)!}$. It is illuminating to plot $C(n,r)$ for variable r and various fixed values of n.[8] See figure 1. This illustrates that the sizes of subsets of U will cluster around $\dfrac{n}{2}$, and they cluster more tightly as n increases. This is precisely what the Indifference Principle tells us.

[7] There are two kinds of defeaters. Rebutting defeaters attack the conclusion of an inference, and undercutting defeaters attack the inference itself without attacking the conclusion. Here I assume some form of the OSCAR theory of defeasible reasoning (Pollock 1995). For a sketch of that theory see Pollock (2006a).

[8] Note that throughout this paper I employ the definition of $n!$ in terms of the Euler gamma function. Specifically, $n! = \int_0^{\infty} t^n e^{-t} dt$. This has the result that $n!$ is defined for any positive real number n, not just for integers, but for the integers the definition agrees with the ordinary recursive definition. This makes the mathematics more convenient.

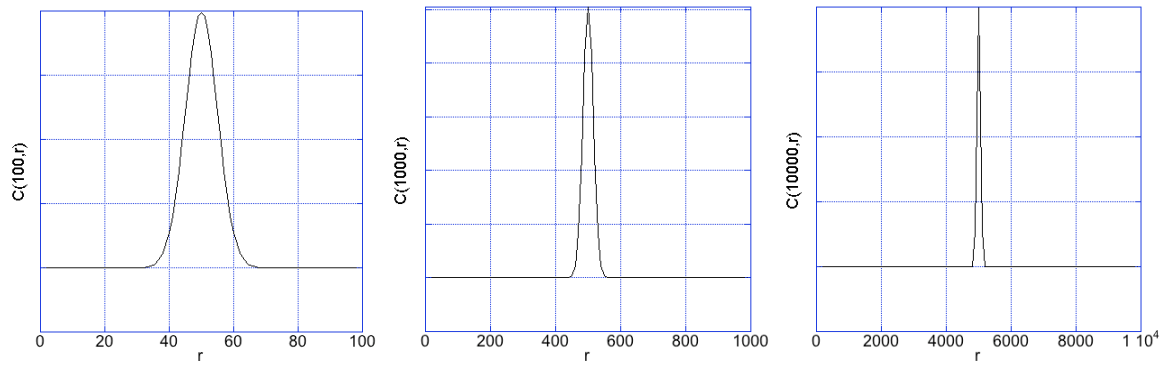

Figure 1. $C(n,r)$ for $n = 100$, $n = 1000$, and $n = 10000$.

The reason the Indifference Principle holds is that $C(n,r)$ becomes "needle-like" in the limit. As we proceed, I will state a number of similar combinatorial theorems, and in each case they have similar intuitive explanations. The cardinalities of relevant sets are products of terms of the form $C(n,r)$, and their distribution becomes needle-like in the limit. In this paper, I will omit the proofs of theorems. They will be presented elsewhere in detail, and can be found on my website in a much longer version of this paper (http://oscarhome.soc-sci.arizona.edu/ftp/PAPERS/Probable-Probabilities-long.pdf

The finite indifference principle is a mathematical theorem about finite sets. It tells us that for fixed $\varepsilon,\delta > 0$, there is an N such that if U is finite but contains at least N members, then the proportion of subsets X of a set U which are such that $\rho(X,U) \underset{\delta}{\approx} 0.5$ is greater than $1-\varepsilon$. This suggests that the proportion is also is greater than $1-\varepsilon$ when U *is* infinite. But if the proportion is is greater than $1-\varepsilon$ for every $\varepsilon > 0$, it follows that the proportion is 1. In other words:

If U is infinite then for every $\delta > 0$, $\rho_X\left(\rho(X,U) \underset{\delta}{\approx} 0.5 \;/\; X \subseteq U\right) = 1$.

Given the rather simple assumptions I made about ρ in section three, we can derive this infinitary principle from the finite principle. First, we can use familiar looking mathematics to prove:

Law of Large Numbers for Proportions:

If B is infinite and $\rho(A/B) = p$ then for every $\varepsilon,\delta > 0$, there is an N such that

$\rho_X\left(\rho(A/X) \underset{\delta}{\approx} p \;/\; X \subseteq B \;\&\; X \text{ is finite } \&\; \#X \geq N\right) \geq 1-\varepsilon$.

Note that unlike Laws of Large Numbers for probabilities, the Law of Large Numbers for Proportions does not require an assumption of statistical independence. This is because it is derived from the crossproduct principle, and as remarked in section three, no such assumption is required (or even intelligible) for the crossproduct principle.

Given the law of large numbers, the finite indifference principle can be shown to entail:

Infinitary Indifference Principle:

If U is infinite then for every $\delta > 0$, $\rho_X\left(\rho(X,U) \underset{\delta}{\approx} 0.5 \;/\; X \subseteq U\right) = 1$.

Nomic probabilities are proportions among physically possible objects. For any property F that is not extraordinarily contrived, the set $\Im$ of physically possible F's will be infinite.[9] Thus the

[9] The following principles apply only to properties for which there are infinitely many physically possible instances, but I will not explicitly include the qualification "non-contrived" in the principles.

infinitary indifference principle for proportions implies an analogous principle for nomic probabilities:

Probabilistic Indifference Principle:

For any property G and for every $\delta > 0$,

$$\text{prob}_X\Big(\text{prob}(X/G) \underset{\delta}{\approx} 0.5 \ / \ X \preccurlyeq G\Big) = 1.$$ [10]

Next note that we can apply the statistical syllogism to the probability formulated in the probabilistic indifference principle. For every $\delta > 0$, this gives us a defeasible reason for expecting that if $F \preccurlyeq G$, then $\text{prob}(F/G) \underset{\delta}{\approx} 0.5$, and these conclusions jointly entail that prob(F/G) = 0.5. For any property F, (F&G) $\preccurlyeq$ G, and prob(F/G) = prob(F&G/G). Thus we are led to a defeasible inference scheme:

Indifference Principle:

For any properties F and G, it is defeasibly reasonable to assume that prob(F/G) = 0.5.

The indifference principle is my first example of a principle of probable probabilities. We have a quadruple of principles that go together: (1) the finite indifference principle, which is a theorem of combinatorial mathematics; (2) the infinitary indifference principle, which follows from the finite principle given the law of large numbers for proportions; (3) the probabilistic indifference principle, which is a theorem derived from (2); and (4) the Indifference Principle, which is a principle of defeasible reasoning that follows from (3) with the help of the statistical syllogism. All of the principles of probable probabilities that I will discuss have analogous quadruples of principles associated with them. Rather than tediously listing all four principles in each case, I will encapsulate the four principles in the simple form:

Expectable Indifference Principle:

For any properties F and G, the expectable value of prob(F/G) = 0.5.

So in talking about expectable values, I am talking about this entire quadruple of principles.

I have chosen the indifference principle as my first example of a principle of probable probabilities because the argument for it is simple and easy to follow. However, as I indicated at the start, this principle is only occasionally useful. If we were choosing the properties F in some random way, it would be reasonable to expect that prob(F/G) = 0.5. However, pairs of properties F and G which are such that prob(F/G) = 0.5 are not very useful to us from a cognitive perspective, because knowing that something is a G then carries no information about whether it is an F. As a result, we usually only enquire about the value of prob(F/G) when we have reason to believe there is a connection between F and G such that prob(F/G) ≠ 0.5. Hence in actual practice, application of the indifference principle to cases that really interest us will almost invariably be defeated. This does not mean, however, that the indifference principle is never useful. For instance, if I give Jones the opportunity to pick either of two essentially identical balls, in the absence of information to the contrary it seems reasonable to take the probability of either choice to be .5. This can be justified as an application of either the indifference principle or the generalized indifference principle.

That applications of the indifference principle are often defeated illustrates an important point about nomic probability and principles of probable probabilities. The fact that a nomic probability is 1 does not mean that there are no counter-instances. In fact, there may be infinitely many counter-instances. Consider the probability of a real number being irrational. Plausibly, this probability is 1, because the cardinality of the set of irrationals is infinitely greater than the cardinality of the set of rationals. But there are still infinitely many rationals. The set of rationals is infinite, but it has

[10] If we could assume countable additivity for nomic probability, the Indifference Principle would imply that $\text{prob}_X\Big(\text{prob}(X,G) = 0.5 \ / \ X \preccurlyeq G\Big) = 1$. Countable additivity is generally assumed in mathematical probability theory, but most of the important writers in the foundations of probability theory, including de Finetti (1974), Reichenbach (1949), Jeffrey (1983), Skyrms (1980), Savage (1954), and Kyburg (1974), have either questioned it or rejected it outright. Pollock (2006) gives what I consider to be a compelling counter-example to countable additivity. So I will have to remain content with the more complex formulation of the Indifference Principle.

measure 0 relative to the set of real numbers.

A second point is that in classical probability theory (which is about singular probabilities), conditional probabilities are defined as ratios of unconditional probabilities:

$$\textsc{prob}(P/Q) = \frac{\textsc{prob}(P\,\&\,Q)}{\textsc{prob}(Q)}.$$

However, for generic probabilities, there are no unconditional probabilities, so conditional probabilities must be taken as primitive. These are sometimes called "Popper functions". The first people to investigate them were Karl Popper (1938, 1959) and the mathematician Alfred Renyi (1955). If conditional probabilities are defined as above, $\textsc{prob}(P/Q)$ is undefined when $\textsc{prob}(Q) = 0$. However, for nomic probabilities, prob(F/G&H) can be perfectly well-defined even when prob(G/H) = 0. One consequence of this is that, unlike in the standard probability calculus, if prob(F/G) = 1, it does not follow that prob(F/G&H) = 1. Specifically, this can fail when prob(H/G) = 0. Thus, for example,

prob(2x is irrational / x is a real number) = 1

but

prob(2x is irrational / x is a real number & x is rational) = 0.

In the course of developing the theory of probable probabilities, we will find numerous examples of this phenomenon, and they will generate defeaters for the defeasible inferences licensed by our principles of probable probabilities.

5. Independence

Now let us turn to a truly useful principle of probable probabilities. It was remarked above that probability practitioners commonly assume statistical independence when they have no reason to think otherwise, and so compute that prob(A&B/C) = prob(A/C)·prob(B/C). In other words, they assume that *A and B are statistically independent relative to C*. This assumption is ubiquitous in almost every application of probability to real-world problems. However, the justification for such an assumption has heretofore eluded probability theorists, and when they make such assumptions they tend to do so apologetically. We are now in a position to provide a justification for a general assumption of statistical independence.

Although it is harder to prove than the finite indifference principle, the following combinatorial principle holds in general:

Finite Independence Principle:

For $0 \le r,s \le 1$ and for every $\varepsilon,\delta > 0$ there is an N such that if U is finite and $\#U > N$, then

$$\rho_{X,Y,Z}\left(\rho(X \cap Y, Z) \underset{\delta}{\approx} r \cdot s \;/\; X,Y,Z \subseteq U \;\&\; \rho(X,Z) = r \;\&\; \rho(Y,Z) = s\right) \ge 1 - \varepsilon.$$

In other words, for a large finite set U, subsets X,Y and Z of U tend to be such that $\rho(X\cap Y,Z)$ is approximately equal to $\rho(X,Z)\cdot\rho(Y,Z)$, and for any fixed degree of approximation, the proportion of subsets of U satisfying this approximation goes to 1 as the size of U goes to infinity.

Given the law of large numbers for proportions, the finite independence principle entails:

Infinitary Independence Principle:

For $0 \le ,r,s \le 1$, if U is infinite then for every $\delta > 0$:

$$\rho_{X,Y,Z}\left(\rho(X \cap Y, Z) \underset{\delta}{\approx} r \cdot s \;/\; X,Y,Z \subseteq U \;\&\; \rho(X,Z) = r \;\&\; \rho(Y,Z) = s\right) = 1.$$

As before, this entails:

Probabilistic Independence Principle:

For $0 \leq r,s \leq 1$ and for any property U, for every $\delta > 0$:

$$\text{prob}_{X,Y,Z}\left(\text{prob}(X \,\&\, Y / Z) \underset{\delta}{\approx} r \cdot s \;\; / \;\; X,Y,Z \preccurlyeq U \,\&\, \text{prob}(X / Z) = r \,\&\, \text{prob}(Y / Z) = s\right) = 1.$$

Again, applying the statistical syllogism to the second-order probability in the probabilistic independence principle, we get:

Principle of Statistical Independence:

⌜prob(A/C) = r & prob(B/C) = s⌝ is a defeasible reason for ⌜prob(A&B / C) = r·s⌝.

Again, we can encapsulate these four principles in a single principle of expectable values:

Principle of Expectable Statistical Independence:

If prob(A/C) = r and prob(B/C) = s, the expectable value of prob(A&B / C) = r·s.

So a provable combinatorial principle regarding finite sets ultimately makes it reasonable to expect, in the absence of contrary information, that properties will be statistically independent of one another. This is the reason why, when we see no connection between properties that would force them to be statistically dependent, we can reasonably expect them to be statistically independent.

The assumption of statistical independence sometimes fails. Clearly, this can happen when there are causal connections between properties. But it can also happen for purely logical reasons. For example, if $A = B$, A and B cannot be independent unless $r = 1$. More general defeaters for the principle of statistical independence will emerge below.

6. The Probable Probabilities Theorem

Principles like that of Statistical Independence are supported by a general combinatorial theorem, which underlies the entire theory of probable probabilities. Given a list of variables $X_1,\ldots,X_n$ ranging over subsets of a set U, Boolean compounds of these sets are compounds formed by union, intersection, and set-complement. So, for example $(X \cup Y)–Z$ is a Boolean compound of X, Y, and Z. *Linear constraints* on the Boolean compounds either state the values of certain proportions, e.g., stipulating that $\rho(X,Y) = r$, or they relate proportions using linear equations. For example, if we know that $X = Y \cup Z$, that generates the linear constraint

$$\rho(X,U) = \rho(Y,U) + \rho(Z,U) - \rho(X \cap Z,U).$$

Our general theorem is:

Probable Proportions Theorem:

Let $U,X_1,\ldots,X_n$ be a set of variables ranging over sets, and consider a finite set LC of linear constraints on proportions between Boolean compounds of those variables. If LC is consistent with the probability calculus, then for any pair of Boolean compounds P,Q of $U,X_1,\ldots,X_n$ there is a real number r between 0 and 1 such that for every $\varepsilon,\delta > 0$, there is an N such that if U is finite and $\#U > N$, then

$$\rho_{X_1,\ldots,X_n}\left(\rho(P,Q) \underset{\delta}{\approx} r \;/\; LC \;\&\; X_1,\ldots,X_n \subseteq U\right) \geq 1 - \varepsilon.$$

This is the theorem that underlies all of the principles developed in this paper. Given the law of large numbers for proportions, we can prove:

Limit Principle for Proportions:

Consider a finite set *LC* of linear constraints on proportions between Boolean compounds of a list of variables $U,X_1,\ldots,X_n$. For any real number *r* between 0 and 1, if for every $\varepsilon,\delta > 0$, if there is an *N* such that for any finite set *U* such that $\#U > N$,

$$\rho_{X_1,\ldots,X_n}\left(\rho(P,Q) \underset{\delta}{\approx} r \;/\; LC \;\&\; X_1,\ldots,X_n \subseteq U\right) \geq 1-\varepsilon,$$

then for any infinite set *U*, for every $\delta > 0$:

$$\rho_{X_1,\ldots,X_n}\left(\rho(P,Q) \underset{\delta}{\approx} r \;/\; LC \;\&\; X_1,\ldots,X_n \subseteq U\right) = 1.$$

Given the limit principle for proportions, the Probable Proportions Theorem entails:

Probable Probabilities Theorem:

Let $U,X_1,\ldots,X_n$ be a set of variables ranging over properties and relations, and consider a finite set *LC* of linear constraints on probabilities between truth-functional compounds of those variables. If *LC* is consistent with the probability calculus, then for any pair of truth-functional compounds *P*,*Q* of $U,X_1,\ldots,X_n$ there is a real number *r* between 0 and 1 such that for every $\delta > 0$,

$$\text{prob}_{X_1,\ldots,X_n}\left(\text{prob}(P/Q) \underset{\delta}{\approx} r \;/\; LC \;\&\; X_1,\ldots,X_n \preccurlyeq U\right) = 1.$$

In other words, given the constraints *LC*, the expectable value of prob(*P*/*Q*) = *r*.

This establishes the existence of expectable values for probabilities under very general circumstances. The theorem can probably be generalized further, e.g., to linear inequalities, or even to nonlinear constraints, but this is what I have established so far.

The Probable Probabilities Theorem tells us that there are expectable values. It turns out that there is a general strategy for finding and proving theorems describing these expectable values, and I have written a computer program (in Common LISP) that will often do this automatically, both finding the theorems and producing human readable proofs. It can be downloaded from http://oscarhome.soc-sci.arizona.edu/ftp/OSCAR-web-page/CODE/Code for probable probabilities.zip.

I will go on to illustrate these general results with several interesting theorems about probable probabilities.

7. Nonclassical Direct Inference

Pollock (1984) noted (a restricted form of) the following limit principle:

Finite Principle of Agreement:

For $0 \leq a,b,c,r \leq 1$ and for every $\varepsilon,\delta > 0$, there is an *N* such that if *U* is finite and $\#U > N$, then:

$$\rho_{X,Y}\left(\begin{array}{l}\rho(X,Y \cap Z) \underset{\delta}{\approx} r \;/\; X,Y,Z \subseteq U \;\&\; \rho(X,Y) = r \\ \qquad\qquad\qquad\qquad \&\; \rho(X,U) = a \;\&\; \rho(Y,U) = b \;\&\; \rho(Z,U) = c\end{array}\right) \geq 1-\varepsilon.$$

In the theory of nomic probability (Pollock 1984,1990), this used this to ground a theory of direct inference. We can now improve upon that theory. As above, the Finite Principle of Agreement yields a principle of expectable values:

Nonclassical Direct Inference:

If prob(*A*/*B*) = *r*, the expectable value of prob(*A*/*B*&*C*) = *r*.

This is a kind of "principle of insufficient reason". It tells us that if we have no reason for thinking otherwise, we should expect that strengthening the reference property in a nomic probability

leaves the value of the probability unchanged. This is called "nonclassical direct inference" because, although it only licenses inferences from generic probabilities to other generic probabilities, it turns out to have strong formal similarities to classical direct inference (which licenses inferences from generic probabilities to singular probabilities), and as we will see in section seven, principles of classical direct inference can be derived from it.

It is important to realize that the principle of agreement, and the corresponding principle of nonclassical direct inference, are equivalent (with one slight qualification) to the probabilistic product principle and the defeasible principle of statistical independence. This turns upon the following simple theorem of the probability calculus:

Independence and Agreement Theorem:

If $\text{prob}(C/B) > 0$ then $\text{prob}(A/B\&C) = \text{prob}(A/B)$ iff A and C are independent relative to B.

As a result, anyone who shares the commonly held intuition that we should be able to assume statistical independence in the absence of information to the contrary is also committed to endorsing nonclassical direct inference. This is important, because I have found that many people do have the former intuition but balk at the latter.

There is a variant of the principle of agreement that is equivalent to the first version but often more useful:

Finite Principle of Agreement II:

For $0 \leq r \leq 1$ and for every $\varepsilon,\delta > 0$, there is an N such that if U is finite and $\#U > N$, then:

$$\rho_{X,Y}\left(\rho(X,Z) \underset{\delta}{\approx} r \;/\; X,Y \subseteq U \;\&\; Z \subseteq Y \;\&\; \rho(X,Y) = r\right) \geq 1-\varepsilon.$$

This yields an equivalent variant of the principle of nonclassical direct inference:

Nonclassical Direct Inference II:

If $C \preccurlyeq B$ and $\text{prob}(A/B) = r$, the expectable value of $\text{prob}(A/C) = r$.

The principle of nonclassical direct inference supports many defeasible inferences that seem intuitively reasonable but are not licensed by the probability calculus. For example, suppose we know that the probability of a twenty year old male driver in Maryland having an auto accident over the course of a year is .07. If we add that his girlfriend's name is "Martha", we do not expect this to alter the probability. There is no way to justify this assumption within a traditional probability framework, but it is justified by nonclassical direct inference.

Nonclassical direct inference is a principle of defeasible inference, so it is subject to defeat. The simplest and most important kind of defeater is a *subproperty defeater*. Suppose $C \preccurlyeq D \preccurlyeq B$ and we know that $\text{prob}(A/B) = r$, but $\text{prob}(A/D) = s$, where $s \neq r$. This gives us defeasible reasons for drawing two incompatible conclusions, viz., that $\text{prob}(A/C) = r$ and $\text{prob}(A/D) = s$. The *principle of subproperty defeat* tells us that because $D \preccurlyeq B$, the latter inference takes precedence and defeats the inference to the conclusion that $\text{prob}(A/C) = r$:

Subproperty Defeat for Nonclassical Direct Inference:

$\ulcorner C \preccurlyeq D \preccurlyeq B$ and $\text{prob}(A/D) = s \neq r\urcorner$ is an undercutting defeater for the inference by nonclassical direct inference from $\ulcorner C \preccurlyeq B$ and $\text{prob}(A/B) = r\urcorner$ to $\ulcorner \text{prob}(A/C) = r\urcorner$.

We obtain this defeater by noting that the principle of nonclassical direct inference is licensed by an application of the statistical syllogism to the probability

$$(1)\quad \text{prob}_{A,B,C}\left(\text{prob}(A/C) \underset{\delta}{\approx} r \;/\; A,B,C \preccurlyeq U \text{ and } C \preccurlyeq B \text{ and } \text{prob}(A/B) = r\right) = 1.$$

We can easily establish the following principle, which appeals to a more comprehensive set of assumptions:

$$(2)\ \text{prob}_{A,B,C}\left(\begin{array}{c}\text{prob}(A/C)\approx_{\delta} s\ /\ A,B,C,D \preccurlyeq U \text{ and } C \preccurlyeq D \text{ and } D \preccurlyeq B \text{ and}\\ \text{prob}(A/B)=r \text{ and } \text{prob}(A/D)=s\end{array}\right)=1.$$

If $r \neq s$ then (2) entails:

$$(3)\ \text{prob}_{A,B,C}\left(\begin{array}{c}\text{prob}(A/C)\approx_{\delta} r\ /\ A,B,C,D \preccurlyeq U \text{ and } C \preccurlyeq D \text{ and } D \preccurlyeq B \text{ and}\\ \text{prob}(A/B)=r \text{ and } \text{prob}(A/D)=s\end{array}\right)=0.$$

The reference property in (3) is more specific than that in (1), so (3) gives us a subproperty defeater for the application of the statistical syllogism to (1).

A simpler way of putting all of this is that corresponding to (2) we have the following principle of expectable values:

Subproperty Defeat for Nonclassical Direct Inference:

If $C \preccurlyeq D \preccurlyeq B$, $\text{prob}(A/D) = s$, $\text{prob}(A/B) = r$, $\text{prob}(A/U) = a$, $\text{prob}(B/U) = b$, $\text{prob}(C/U) = c$, $\text{prob}(D/U) = d$, then the expectable value of $\text{prob}(A/C) = s$ (rather than r).

As above, principles of expectable values that appeal to more information take precedence over (i.e., defeat the inferences from) principles that appeal to a subset of that information.

Because the principles of nonclassical direct inference and statistical independence are equivalent, subproperty defeaters for nonclassical direct inference generate analogous defeaters for the principle of statistical independence:

Subproperty Defeat for Statistical Independence:

$\ulcorner (B\&C) \preccurlyeq D \preccurlyeq C$ and $\text{prob}(A/D) = p \neq r\urcorner$ is an undercutting defeater for the inference by the principle of statistical independence from $\ulcorner \text{prob}(A/C) = r$ & $\text{prob}(B/C) = s\urcorner$ to $\ulcorner \text{prob}(A\&B/C) = r{\cdot}s\urcorner$.

This is because $\text{prob}(A\&B/C) = r{\cdot}s$ only if $\text{prob}(A/B\&C) = \text{prob}(A/C)$, and this defeater makes it unreasonable to believe the former.

8. Classical Direct Inference

Direct inference is normally understood as being a form of inference from generic probabilities to singular probabilities rather than from generic probabilities to other generic probabilities. However, I showed in my (1990) that these inferences are derivable from nonclassical direct inference if we identify singular probabilities with a special class of generic probabilities. The present treatment is a generalization of that given in my (1984 and 1990).[11] Let $\mathbf{K}$ be the conjunction of all the propositions the agent knows to be true, and let $\mathfrak{R}$ be the set of all physically possible worlds at which $\mathbf{K}$ is true ("$\mathbf{K}$-worlds"). I propose that we define the singular probability PROB(P) to be the proportion of $\mathbf{K}$-worlds at which P is true. Where $\mathcal{P}$ is the set of all P-worlds:

$$\text{PROB}(P) = \rho(\mathcal{P},\mathfrak{R}).$$

More generally, where $\mathcal{Q}$ is the set of all Q-worlds, we can define:

$$\text{PROB}(P/Q) = \rho(\mathcal{P},\ \mathcal{Q}\cap\mathfrak{R}).$$

Formally, this is analogous to Carnap's (1950,1952) logical probability, with the important difference that Carnap took ρ to be logically specified, whereas I take the identity of ρ to be a

[11] Bacchus (1990) gave a somewhat similar account of direct inference, drawing on my 1983 and 1984.

contingent fact. ρ is determined by the values of contingently true nomic probabilities, and their values are discovered by various kinds of statistical induction.

It turns out that singular probabilities, so defined, can be identified with a special class of nomic probabilities:

Representation Theorem for Singular Probabilities:

(1) $\textsc{prob}(Fa) = \text{prob}(Fx/x = a \mathbin{\&} \mathbf{K})$;

(2) If it is physically necessary that $[\mathbf{K} \rightarrow (Q \leftrightarrow Sa_1\ldots a_n)]$ and that $[(Q\&\mathbf{K}) \rightarrow (P \leftrightarrow Ra_1\ldots a_n)]$, and Q is consistent with $\mathbf{K}$, then $\textsc{prob}(P/Q) = \text{prob}(Rx_1\ldots x_n/ Sx_1\ldots x_n \mathbin{\&} x_1 = a_1 \mathbin{\&} \ldots \mathbin{\&} x_n = a_n \mathbin{\&} \mathbf{K})$.

(3) $\textsc{prob}(P) = \text{prob}(P \mathbin{\&} x{=}x/x = x \mathbin{\&} \mathbf{K})$.

$\textsc{prob}(P)$ is a kind of "mixed physical/epistemic probability", because it combines background knowledge in the form of $\mathbf{K}$ with generic probabilities.[12]

The probability $\text{prob}(Fx/x = a \mathbin{\&} \mathbf{K})$ is a peculiar–looking nomic probability. It is an generic probability, because "x" is a free variable, but the probability is only about one object. As such it cannot be evaluated by statistical induction or other familiar forms of statistical reasoning. However, it can be evaluated using nonclassical direct inference. If $\mathbf{K}$ entails Ga, nonclassical direct inference gives us a defeasible reason for expecting that $\textsc{prob}(Fa) = \text{prob}(Fx/x = a \mathbin{\&} \mathbf{K}) = \text{prob}(Fx/Gx)$. This is a familiar form of "classical" direct inference — that is, direct inference from nomic probabilities to singular probabilities. More generally, we can derive:

Classical Direct Inference:

$\ulcorner Sa_1\ldots a_n$ is known and $\text{prob}(Rx_1\ldots x_n/\ Sx_1\ldots x_n \mathbin{\&} Tx_1\ldots x_n) = r\urcorner$ is a defeasible reason for

$\ulcorner \textsc{prob}(Ra_1\ldots a_n\ /\ Ta_1\ldots a_n) = r\urcorner$.

Similarly, we get subproperty defeaters:

Subproperty Defeat for Classical Direct Inference:

$\ulcorner V \preccurlyeq S$, $Va_1\ldots a_n$ is known, and $\text{prob}(Rx_1\ldots x_n/\ Vx_1\ldots x_n \mathbin{\&} Tx_1\ldots x_n) \neq r\urcorner$ is an undercutting defeater for the inference by classical direct inference from $\ulcorner Sa_1\ldots a_n$ is known and $\text{prob}(Rx_1\ldots x_n/\ Sx_1\ldots x_n \mathbin{\&} Tx_1\ldots x_n) = r\urcorner$ to $\ulcorner \textsc{prob}(Ra_1\ldots a_n\ /\ Ta_1\ldots a_n) = r\urcorner$.

Because singular probabilities are generic probabilities in disguise, we can also use nonclassical direct inference to infer singular probabilities from singular probabilities. Thus $\ulcorner \textsc{prob}(P/Q) = r\urcorner$ gives us a defeasible reason for expecting that $\textsc{prob}(P/Q\&R) = r$. We can employ principles of statistical independence similarly. For example, $\ulcorner \textsc{prob}(P/R) = r \mathbin{\&} \textsc{prob}(Q/R) = s\urcorner$ gives us a defeasible reason for expecting that $\textsc{prob}(P\&Q/R) = r{\cdot}s$.

9. Computational Inheritance

Suppose we have two seemingly unrelated diagnostic tests for a disease, and Bernard tests positive on both tests. We know that the probability of his having the disease if he tests positive on the first test is .8, and the probability if he tests positive on the second test is .75. But what should we conclude about the probability of his having the disease if he tests positive on both tests? The probability calculus gives us no guidance here. Nor does direct inference. Direct inference gives us one reason for thinking the probability of Bernard having the disease is .8, and it gives us a different reason for drawing the conflicting conclusion that the probability is .75. It gives us no way to combine the information. Intuitively, it seems that the probability of his having the disease should be higher if he tests positive on both tests. But how can we justify this?

This is a general problem for theories of direct inference. When we have some conjunction $\ulcorner G_1$

[12] See chapter six of my (2006) for further discussion of these mixed physical/epistemic probabilities.

&…& G_n⌝ of properties and we want to know the value of prob(F/G_1 &…& G_n), if we know that prob(F/G_1) = r *and we don't know anything else of relevance*, we can infer defeasibly that prob(F/G_1 &…& G_n) = r. Similarly, if we know that an object a has the properties G_1,…,G_n and we know that prob(F/G_1) = r *and we don't know anything else of relevance*, we can infer defeasibly that PROB(Fa) = r. The difficulty is that we usually know more. We typically know the value of prob(F/G_i) for some $i \neq 1$. If prob(F/G_i) = $s \neq r$, we have defeasible reasons for both ⌜prob(F/G_1 &…&G_n) = r⌝ and ⌜prob(F/G_1 &…&G_n) = s⌝, and also for both ⌜PROB(Fa) = r⌝ and ⌜PROB(Fa) = s⌝. As these conclusions are incompatible they all undergo collective defeat. Thus the standard theory of direct inference leaves us without a conclusion to draw. The upshot is that the earlier suggestion that direct inference can solve the computational problem of dealing with singular probabilities without having to have a complete probability distribution was premature. Direct inference will rarely give us the probabilities we need.

Knowledge of generic probabilities would be vastly more useful in real application if there were a function Y(r,s) such that, in a case like the above, when prob(F/G) = r and prob(F/H) = s, we could defeasibly expect that prob(F/G&H) = Y(r,s), and hence (by nonclassical direct inference) that PROB(Fa) = Y(r,s). I call this *computational inheritance*, because it computes a new value for PROB(Fa) from previously known generic probabilities. Direct inference, by contrast, is a kind of "noncomputational inheritance". It is *direct* in that PROB(Fa) simply inherits a value from a known generic probability. I call the function used in computational inheritance "the Y-function" because its behavior would be as diagrammed in figure 2.

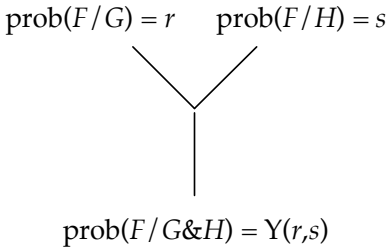

Figure 2. The Y-function

It has generally been assumed that there is no such function as the Y-function (Reichenbach 1949). Certainly, there is no function Y(r,s) such that we can conclude *deductively* that if prob(F/G) = r and prob(F/H) = s then prob(F/G&H) = Y(r,s). For any r and s that are neither 0 nor 1, prob(F/G&H) can take any value between 0 and 1. However, that is equally true for nonclassical direct inference. That is, if prob(F/G) = r we cannot conclude deductively that prob(F/G&H) = r. Nevertheless, that will tend to be the case, and we can defeasibly expect it to be the case. Might something similar be true of the Y-function? That is, could there be a function Y(r,s) such that we can defeasibly expect prob(F/G&H) to be Y(r,s)? It follows from the Probable Probabilities Theorem that the answer is "Yes". It is more useful to begin by looking at a three-place function rather than a two-place function. Let us define:

$$\mathrm{Y}(r,s\,|a) = \frac{rs(1-a)}{a(1-r-s)+rs}$$

I use the non-standard notation "Y(r,s |a)" rather than "Y(r,s,a)" because the first two variables will turn out to work differently than the last variable.

Let us define:

B and *C* are *Y-independent for A relative to U* iff $A,B,C \preccurlyeq U$ and

(a) prob(C/B & A) = prob(C/A)

and

(b) prob(C/B & ~A) = prob(C/U & ~A).

The key theorem underlying computational inheritance is the following theorem of the probability calculus:

Y-Theorem:

Let $r = \text{prob}(A/B)$, $s = \text{prob}(A/C)$, and $a = \text{prob}(A/U)$. If B and C are Y-independent for A relative to U then $\text{prob}(A/B\&C) = \text{Y}(r,s\,|a)$.

In light of the Y-theorem, we can think of Y-independence as formulating an independence condition for C and D which says that they make independent contributions to A — contributions that "add" in accordance with the Y-function, rather than "undermining" each other.

By virtue of the principle of statistical independence, we have a defeasible reason for expecting that the independence conditions (a) and (b) hold. Thus the Y-theorem supports the following principle of defeasible reasoning:

Computational Inheritance:

$\ulcorner B,C \preccurlyeq U$ & $\text{prob}(A/B) = r$ & $\text{prob}(A/C) = s$ & $\text{prob}(A/U) = a\urcorner$ is a defeasible reason for $\ulcorner \text{prob}(A/B \,\&\, C) = \text{Y}(r,s\,|a)\urcorner$.

It should be noted that we can prove analogues of Computational Inheritance for finite sets, infinite sets, and probabilities, in essentially the same way we prove the Y-theorem. This yields the following principle of expectable values:

Y-Principle:

If $B,C \preccurlyeq U$, $\text{prob}(A/B) = r$, $\text{prob}(A/C) = s$, and $\text{prob}(A/U) = a$, then the expectable value of $\text{prob}(A/B \,\&\, C) = \text{Y}(r,s\,|a)$.

In the corresponding quadruple of principles, the Finite Y-Principle can be proven directly, or derived from the Finite Principle of Agreement. Similarly, the Y-Principle is derivable from the Principle of Agreement. Then the Y-Principle for Probabilities is derivable from either the Y-Principle or from the Principle of Agreement for Probabilities.

To get a better feel for what the principle of computational inheritance tells us, it is useful to examine plots of the Y-function. Figure 3 illustrates that $\text{Y}(r,s\,|.5)$ is symmetric around the right-leaning diagonal.

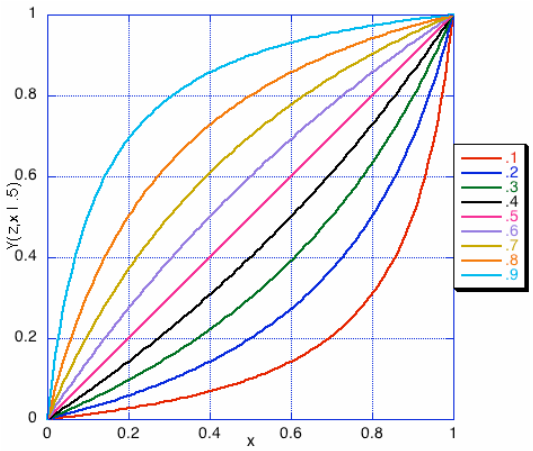

Figure 3. $\text{Y}(z,x\,|.5)$, holding z constant
(for several choices of z as indicated in the key).

Varying a has the effect of warping the Y-function up or down relative to the right-leaning diagonal. This is illustrated in figure 4 for several choices of a.

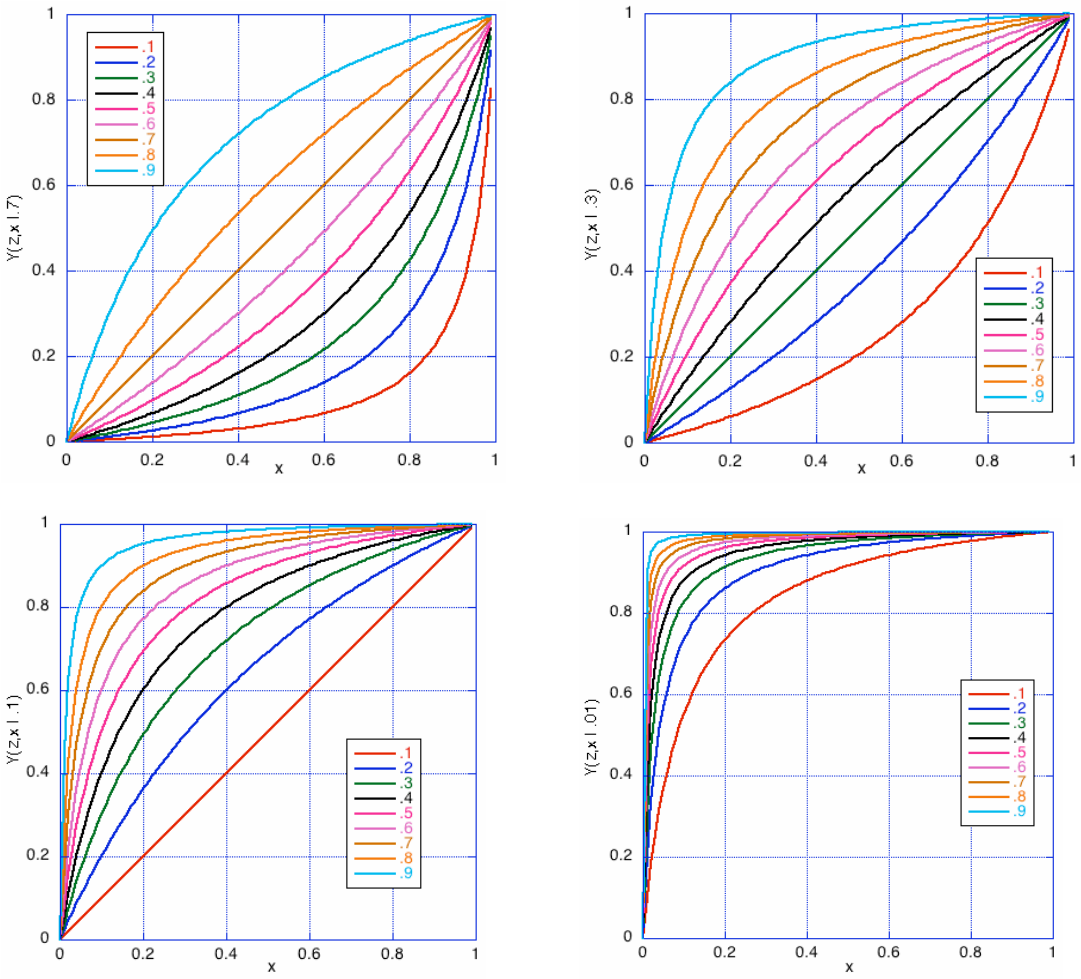

Figure 4. Y(z,x |a) holding z constant (for several choices of z), for a = .7, a = .3, a = .1, and a = .01.

The Y-function has a number of important properties.[13] In particular, it is important that the Y-function is commutative and associative in the first two variables:

Theorem 1: Y(r,s |a) = Y(s,r |a).

Theorem 2: Y(r,Y(s,t |a) |a) = Y(Y(r,s |a),t |a).

Theorems 1 and 2 are very important for the use of the Y-function in computing probabilities. Suppose we know that prob(A/B) = .6, prob(A/C) = .7, and prob(A/D) = .75, where B,C,D $\preccurlyeq$ U and prob(A/U) = .3. In light of theorems 1 and 2 we can combine the first three probabilities in any order and infer defeasibly that prob(A/B&C&D) = Y(.6,Y(.7,.75 |.3) |.3) = Y(Y(.6,.7 |.3),.75 |.3) = .98. This makes it convenient to extend the Y-function recursively so that it can be applied to an arbitrary number of arguments (greater than or equal to 3):

[13] It turns out that the Y-function has been studied for its desirable mathematical properties in the theory of associative compensatory aggregation operators in fuzzy logic (Dombi 1982; Klement, Mesiar, and Pap 1996; Fodor, Yager, and Rybalov 1997). Y(r,s |a) is the function $D_\lambda(r,s)$ for $\lambda = \frac{1-a}{a}$ (Klement, Mesiar, and Pap 1996). The Y-theorem may provide further justification for its use in that connection.

If $n \geq 3$, $\mathrm{Y}(r_1,\ldots,r_n \,|a) = \mathrm{Y}(r_1,\mathrm{Y}(r_2,\ldots,r_n \,|a) \;|a)$.

Then we can then strengthen the Y-Principle as follows:

Generalized Y-Principle:

If $B_1,\ldots,B_n \preccurlyeq U$, $\mathrm{prob}(A/B_1) = r_1,\ldots$, $\mathrm{prob}(A/B_n) = r_n$, and $\mathrm{prob}(A/U) = a$, the expectable value of $\mathrm{prob}(A/\ B_1 \,\&\ldots\&\, B_n \,\&\, C) = \mathrm{Y}(r_1,\ldots,r_n \,|a)$.

If we know that $\mathrm{prob}(A/B) = r$ and $\mathrm{prob}(A/C) = s$, we can also use nonclassical direct inference to infer defeasibly that $\mathrm{prob}(A/B\&C) = r$. If $s \neq a$, $\mathrm{Y}(r,s \,|a) \neq r$, so this conflicts with the conclusion that $\mathrm{prob}(A/B\&C) = \mathrm{Y}(r,s \,|a)$. However, as above, the inference described by the Y-principle is based upon a probability with a more inclusive reference property than that underlying Nonclassical Direct Inference (that is, it takes account of more information), so it takes precedence and yields an undercutting defeater for Nonclassical Direct Inference:

Computational Defeat for Nonclassical Direct Inference**:**

$\ulcorner A,B,C \preccurlyeq U$ and $\mathrm{prob}(A/C) \neq \mathrm{prob}(A/U)\urcorner$ is an undercutting defeater for the inference from $\ulcorner \mathrm{prob}(A/B) = r\urcorner$ to $\ulcorner \mathrm{prob}(A/B\&C) = r\urcorner$ by Nonclassical Direct Inference.

It follows that follows that we have defeater for the principle of statistical independence:

Computational Defeat for Statistical Independence:

$\ulcorner A,B,C \preccurlyeq U$ and $\mathrm{prob}(A/B) \neq \mathrm{prob}(A/U)\urcorner$ is an undercutting defeater for the inference from $\ulcorner \mathrm{prob}(A/B) = r \,\&\, \mathrm{prob}(A/C) = s\urcorner$ to $\ulcorner \mathrm{prob}(A\&B/C) = r{\cdot}s\urcorner$ by Statistical Independence.

The phenomenon of Computational Inheritance makes knowledge of generic probabilities useful in ways it was never previously useful. It tells us how to combine different probabilities that would lead to conflicting direct inferences and still arrive at a univocal value. Consider Bernard again, who has symptoms suggesting a particular disease, and tests positive on two independent tests for the disease. Suppose the probability of a person with those symptoms having the disease is .6. Suppose the probability of such a person having the disease is they test positive on the first test is .7, and the probability of their having the disease if they test positive on the second test is .75. What is the probability of their having the disease if they test positive on both tests? We can infer defeasibly that it is $\mathrm{Y}(.7,.75 \,|.6) = .875$. We can then apply classical direct inference to conclude that the probability of Bernard's having the disease is .875. This is a result that we could not have gotten from the probability calculus alone. Similar reasoning will have significant practical applications, for example in engineering where we have multiple imperfect sensors sensing some phenomenon and we want to arrive at a joint probability regarding the phenomenon that combines the information from all the sensors.

Again, because singular probabilities are generic probabilities in disguise, we can apply computational inheritance to them as well and infer defeasibly that if $\textsc{prob}(P) = a$, $\textsc{prob}(P/Q) = r$, and $\textsc{prob}(P/R) = s$ then $\textsc{prob}(P/Q\&R) = \mathrm{Y}(r,s \,|a)$.

10. Inverse Probabilities and the Statistical Syllogism

All of the principles of probable probabilities that have been discussed so far are related to defeasible assumptions of statistical independence. As we have seen, Nonclassical Direct Inference is equivalent to a defeasible assumption of statistical independence, and Computational Inheritance follows from a defeasible assumption of Y-independence. This might suggest that all principles of probable probabilities derive ultimately from various defeasible independence assumptions. However, this section turns to a set of principles that do not appear to be related to statistical independence in any way.

Where $A,B \preccurlyeq U$, suppose we know the value of $\mathrm{prob}(A/B)$. If we know the base rates $\mathrm{prob}(A/U)$ and $\mathrm{prob}(B/U)$, the probability calculus enables us to compute the value of the *inverse probability*

prob(~B/~A&U):

Theorem 3: If $A,B \preccurlyeq U$ then

$$\text{prob}(\sim B/\sim A\&U) = \frac{1-\text{prob}(A/U)-\text{prob}(B/U)+\text{prob}(A/B)\cdot\text{prob}(B/U)}{1-\text{prob}(A/U)}.$$

However, if we do not know the base rates then the probability calculus imposes no constraints on the value of the inverse probability. It can nevertheless be shown that there are expectable values for it, and generally, if prob(A/B) is high, so is prob(~B/~A&U).

Inverse Probabilities I:

If $A,B \preccurlyeq U$ and we know that prob(A/B) = r, but we do not know the base rates prob(A/U) and prob(B/U), the following values are expectable:

$$\text{prob}(B/U) = \frac{.5}{r^r(1-r)^{1-r}+.5};$$

$$\text{prob}(A/U) = .5-\frac{.25-.5r}{r^r(1-r)^{1-r}+.5};$$

$$\text{prob}(\sim A/\sim B\&U) = .5;$$

$$\text{prob}(\sim B/\sim A\&U) = \frac{r^r}{(1-r)^r+r^r}.$$

These values are plotted in figure 5. Note that when prob(A/B) > prob(A/U), we can expect prob(~B/~A&U) to be almost as great as prob(A/B).

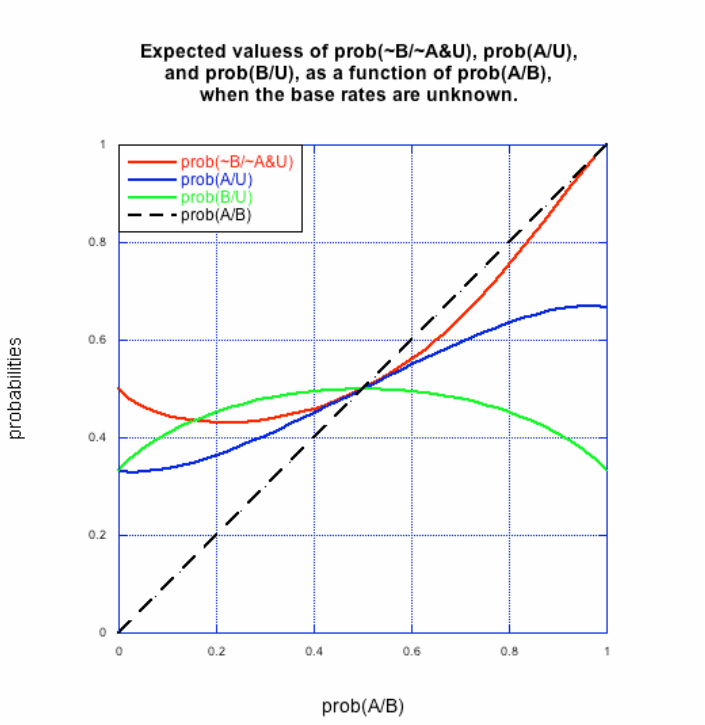

Figure 5. Expectable values of prob(~B/~A&U), prob(A/U), and prob(B/U), as a function of prob(A/B), when the base rates are unknown.

Sometimes we know one of the base rates but not both:

Inverse Probabilities II:

If $A,B \preccurlyeq U$ and we know that prob$(A/B) = r$ prob$(B/U) = b$, but we do not know the base rate prob(A/U) , the following values are expectable:

$$\text{prob}(A/U) = .5(1 - (1 - 2r)b);$$

$$\text{prob}(\sim A/\sim B\&U) = \frac{.5 + b(.5 - r)}{1 + b(1 - r)};$$

$$\text{prob}(\sim B/\sim A\&U)) = \frac{1 - b}{1 + b(1 - 2r)}.$$

Figure 6 plots the expectable values of prob$(\sim B/\sim A\&U)$ (when they are greater than .5) as a function of prob(A/B), for fixed values of prob(B/U). The diagonal dashed line indicates the value of prob(A/B), for comparison. The upshot is that for low values of prob(B/U), prob$(\sim B/\sim A\&U)$ can be expected to be higher than prob(A/B), and for all values of prob(B/U), prob$(\sim B/\sim A\&U)$ will be fairly high if prob(A/B) is high. Furthermore, prob$(\sim B/\sim A\&U) > .5$ iff prob$(B/U) < \frac{1}{3 - 2r}$.

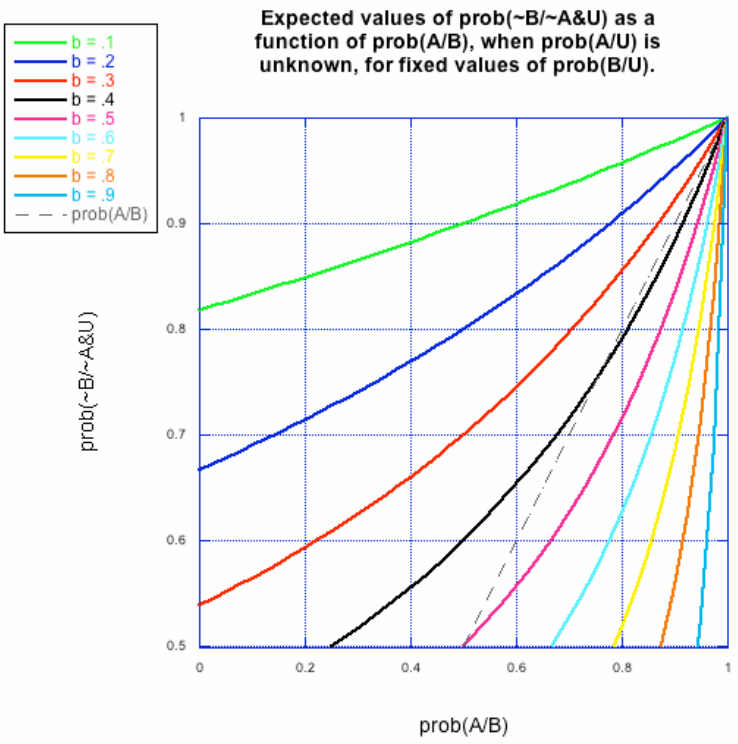

Figure 6. Expectable values of prob$(\sim B/\sim A\&U)$ as a function of prob(A/B), when prob(A/U) is unknown, for fixed values of prob(B/U).

The most complex case occurs when we do know the base-rate prob(A/U) but we do not know the base-rate prob(B/U):

Inverse Probabilities III:

If $A,B \preccurlyeq U$ and we know that prob$(A/B) = r$ and prob$(A/U) = a$, but we do not know the base rate prob(B/U), then:

(a) where b is the expectable value of prob(B/U), $\left(\frac{r \cdot b}{a - r \cdot b}\right)^{r} \cdot \left(\frac{(1-r)b}{1-a-(1-r)b}\right)^{1-r} = 1$;

(b) the expectable value of prob$(\sim B/\sim A \& U) = 1 - \frac{1-r}{1-a} b$.

The equation characterizing the expectable value of prob(B/U) does not have a closed-form solution. However, for specific values of a and r, the solutions are easily computed using hill-climbing algorithms. The results are contained in figure 7. When prob(A/B) = prob(A/U), the expected value for prob$(\sim B/\sim A)$ is .5, and when prob(A/B) > prob(A/U), prob$(\sim B/\sim A \& U)$ > .5. If prob(A/U) < .5, the expected value of prob$(\sim B/\sim A \& U)$ is greater than prob(A/B).

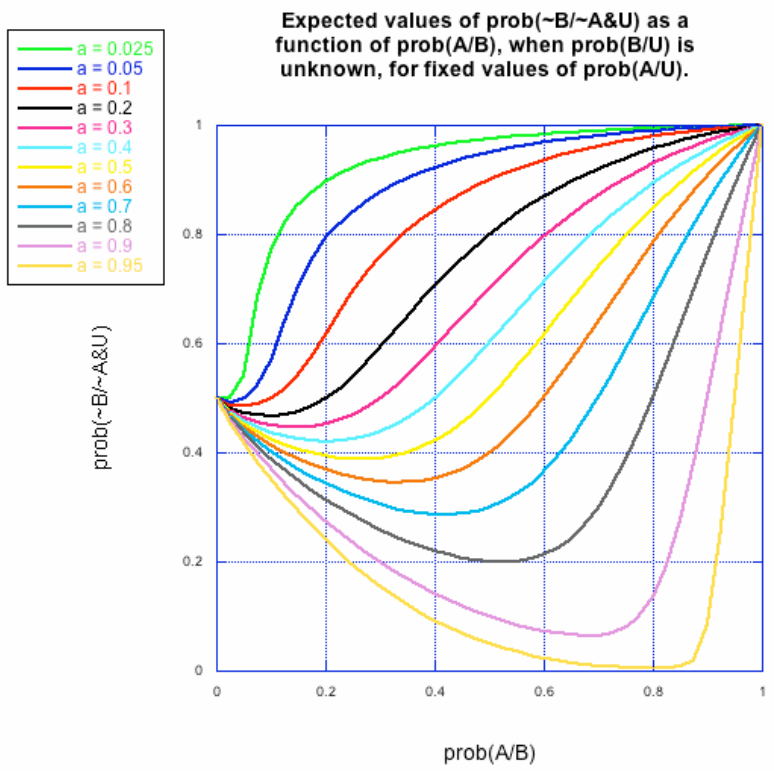

Figure 7. Expectable values of prob$(\sim B/\sim A \& U)$ as a function of prob(A/B), when prob(B/U) is unknown, for fixed values of prob(A/U).

The upshot is that even when we lack knowledge of the base rates, there is an expectable value for the inverse probability prob$(\sim B/\sim A \& U)$, and that expectable value tends to be high when prob(A/B) is high.

11. Meeting Some Objections

I have argued that mathematical results, coupled with the statistical syllogism, justify defeasible inferences about the values of unknown probabilities. Various worries arise regarding this conclusion. A few people are worried about any defeasible (non-deductive) inference, but I presume that the last 50 years of epistemology has made it amply clear that, in the real world, cognitive agents cannot confine themselves to conclusions drawn deductively from their evidence. We employ multitudes of defeasible inference schemes in our everyday reasoning, and the statistical syllogism is one of them.

Granted that we have to reason defeasibly, we can still ask what justifies any particular defeasible inference scheme. At least in the case of the statistical syllogism, the answer seems clear. If prob(A/B) is high, then if we reason defeasibly from things being B to their being A, we will generally get it right. That is the most we can require of a defeasible inference scheme. We cannot require that the inference scheme will always lead to true conclusions, because then it would not be defeasible. People sometimes protest at this point that they are not interested in the general case. They are concerned with some inference they are only going to make once. They want to know why they should reason this way in the single case. But all cases are single cases. If you reason in this way in single cases, you will tend to get them right. It does not seem that you can ask for any firmer guarantee than that. You cannot avoid defeasible reasoning.

But we can have a further worry. For any defeasible inference scheme, we know that there will be at possible cases in which it gets things wrong. For each principle of probable probabilities, the possible exceptions constitute a set of measure 0, but it is still an infinite set. The cases that actually interest us tend to be highly structured, and perhaps they also constitute a set of measure 0. How do we know that the latter set is not contained in the former? Again, there can be no logical guarantee that this is not the case. However, the generic probability of an arbitrary set of cases falling in the set of possible exceptions is 0. So without further specification of the structure of the cases that interest us, the probability of the set of those cases all falling in the set of exceptions is 0. Where defeasible reasoning is concerned, we cannot ask for a better guarantee than that.

We should resist the temptation to think of the set of possible exceptions as an amorphous unstructured set about which we cannot reason using principles of probable probabilities. The exceptions are exceptions to a single defeasible inference scheme. Many of the cases in which a particular inference fails will be cases in which there is a general defeater leading us to expect it to fail and leading us to make a different inference in its place. For example, knowing that prob(A/B) = r gives us a defeasible reason to expect that prob($A/B\&C$) = r. But if we also know that prob(A/C) = s and prob(A/U) = a, the original inference is defeated and we should expect instead that prob($A/B\&C$) = $\Upsilon(r,s \mid a)$. So this is one of the cases in which an inference by nonclassical direct inference fails, but it is a defeasibly expectable case.

There will also be cases that are not defeasibly expectable. This follows from the simple fact that there are primitive nomic probabilities representing statistical laws of nature. These laws are novel, and cannot be predicted defeasibly by appealing to other nomic probabilities. Suppose prob(A/B) = r, but $\ulcorner$prob($A/B\&C$) = $s\urcorner$ is a primitive law. The latter is an exception to nonclassical direct inference. Furthermore, we can expect that strengthening the reference property further will result in nomic probabilities like $\ulcorner$prob($A/B\&C\&D$) = $s\urcorner$, and these will also be cases in which the nonclassical direct inference from $\ulcorner$prob(A/B) = $r\urcorner$ fails. But, unlike the primitive law, the latter is a defeasibly expectable failure arising from subproperty defeat. So most of the cases in which a particular defeasible inference appealing to principles of probable probabilities fails will be cases in which the failure is defeasibly predictable by appealing to other principles of probable probabilities. This is an observation about how much structure the set of exceptions (of measure 0) must have. The set of exceptions is a set of exceptions just to a single rule, not to all principles of probable probabilities. The Probable Probabilities Theorem implies that even within the set of exceptions to a particular defeasible inference scheme, most inferences that take account of the primitive nomic probabilities will get things right, with probability 1.

12. Conclusions

The problem of sparse probability knowledge results from the fact that in the real world we lack direct knowledge of most probabilities. If probabilities are to be useful, we must have ways of making defeasible estimates of their values even when those values are not computable from known probabilities using the probability calculus. Within the theory of nomic probability, limit theorems from combinatorial mathematics provide the necessary bridge for these inferences. It turns out that in very general circumstances, there will be expectable values for otherwise unknown probabilities. These are described by principles telling us that although certain inferences from probabilities to probabilities are not deductively valid, nevertheless the second-order probability of their yielding correct results is 1. This makes it defeasibly reasonable to make the inferences.

I illustrated this by looking at indifference, statistical independence, classical and nonclassical direct inference, computational inheritance, and inverse probabilities. But these are just illustrations.

There are a huge number of useful principles of probable probabilities, some of which I have investigated, but most waiting to be discovered. I proved the first such principles laboriously by hand. It took me six months to find and prove the principle of computational inheritance. But it turns out that there is a uniform way of finding and proving these principles. I have written a computer program (in Common LISP) that analyzes the results of linear constraints and determines what the expectable values of the probabilities are. If desired, it will produce a human-readable proof. This makes it easy to find and investigate new principles.

This profusion of principles of probable probability is reminiscent of Carnap's logical probabilities (Carnap 1950, 1952; Hintikka 1966; Bacchus et al 1996). Historical theories of objective probability required probabilities to be assessed by empirical methods, and because of the weakness of the probability calculus, they tended to leave us in a badly impoverished epistemic state regarding probabilities. Carnap tried to define a kind of probability for which the values of probabilities were determined by logic alone, thus vitiating the need for empirical investigation. However, finding the right probability measure to employ in a theory of logical probabilities proved to be an insurmountable problem.

Nomic probability and the theory of probable probabilities lies between these two extremes. This theory still makes the values of probabilities contingent rather than logically necessary, but it makes our limited empirical investigations much more fruitful by giving them the power to license defeasible, non-deductive, inferences to a wide range of further probabilities that we have not investigated empirically. Furthermore, unlike logical probability, these defeasible inferences do not depend upon ad hoc postulates. Instead, they derive directly from provable theorems of combinatorial mathematics. So even when we do not have sufficient empirical information to deductively determine the value of a probability, purely mathematical facts may be sufficient to make it reasonable, given what empirical information we do have, to expect the unknown probabilities to have specific and computable values. Where this differs from logical probability is (1) that the empirical values are an essential ingredient in the computation, and (2) that the inferences to these values are defeasible rather than deductive.

References

Bacchus, Fahiem
1990 *Representing and Reasoning with Probabilistic Knowledge*, MIT Press.
Bacchus, Fahiem, Adam J. Grove, Joseph Y. Halpern, Daphne Koller
1996 "From statistical knowledge bases to degrees of belief", *Artificial Intelligence* **87**, 75-143.
Braithwaite, R. B.
1953 *Scientific Explanation*. Cambridge: Cambridge University Press.
Carnap, Ruldolph
1950 *The Logical Foundations of Probability*. Chicago: University of Chicago Press.
1952 *The Continuum of Inductive Methods*. Chicago: University of Chicago Press.
de Finetti, B.
1974 *Theory of Probability*, vol. 1. New York: John Wiley and Sons.
Dombi, J.
1982 "Basic concepts for a theory of evaluation: The aggregative operator", *European Journal of Operational Research* **10**, 282-293.
Fisher, R. A.
1922 "On the mathematical foundations of theoretical statistics." *Philosophical Transactions of the Royal Society A*, 222, 309-368.
Fodor, J., R. Yager, A. Rybalov
1997 "Structure of uninorms", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **5**, 411 - 427.
Goodman, Nelson
1955 *Fact, Fiction, and Forecast*, Cambridge, Mass.: Harvard University Press.
Halpern, J. Y.
1990 "An analysis of first-order logics of probability", *Artificial Intelligence* **46**, 311-350.
Harman, Gilbert
1986 *Change in View*. MIT Press, Cambridge, Mass.
Hintikka, Jaakko

1966 "A two-dimensional continuum of inductive methods". In *Aspects of Inductive Logic*, ed. J. Hintikka and P. Suppes, 113-132. Amsterdam: North Holland.
Jeffrey, Richard
1983 *The Logic of Decision, 2nd edition*, University of Chicago Press.
Klement, E. P., R. Mesiar, E. Pap, E.
1996 "On the relationship of associative compensatory operators to triangular norms and conorms", *Int J. of Unc. Fuzz. and Knowledge-Based Systems* **4**, 129-144.
Kneale, William
1949 *Probability and Induction*. Oxford: Oxford University Press.
Kushmerick, N., Hanks, S., and Weld, D.
1995 "An algorithm for probabilistic planning", *Artificial Intelligence* **76**, 239-286.
Kyburg, Henry, Jr.
1961 *Probability and the Logic of Rational Belief*. Middletown, Conn.: Wesleyan University Press.
1974 *The Logical Foundations of Statistical Inference*, Dordrecht: Reidel.
1974a "Propensities and probabilities." *British Journal for the Philosophy of Science* **25**, 321-353.
Levi, Isaac
1980 *The Enterprise of Knowledge*. Cambridge, Mass.: MIT Press.
Pearl, Judea
1988 *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Mateo, CA: Morgan Kaufmann.
Pollock, John L.
1983 "A theory of direct inference", *Theory and Decision* **15**, 29-96.
1984 "Foundations for direct inference". *Theory and Decision* **17**, 221-256.
1984a *Foundations of Philosophical Semantics*, Princeton: Princeton University Press.
1990 *Nomic Probability and the Foundations of Induction*, New York: Oxford University Press.
1995 *Cognitive Carpentry*, Cambridge, MA: Bradford/MIT Press.
2006 *Thinking about Acting: Logical Foundations for Rational Decision Making*, New York: Oxford University Press.
2006a "Defeasible reasoning", in *Reasoning: Studies of Human Inference and its Foundations*, (ed) Jonathan Adler and Lance Rips, Cambridge: Cambridge University Press.
Popper, Karl
1938 "A set of independent axioms for probability", *Mind* **47**, 275ff.
1956 "The propensity interpretation of probability." *British Journal for the Philosophy of Science* **10**, 25-42.
1957 "The propensity interpretation of the calculus of probability, and the quantum theory." In *Observation and Interpretation*, ed. S. Körner, 65-70. New York: Academic Press.
1959 *The Logic of Scientific Discovery*, New York: Basic Books.
Reichenbach, Hans
1949 *A Theory of Probability*, Berkeley: University of California Press. (Original German edition 1935.)
Reiter, R., and G. Criscuolo
1981 "On interacting defaults", in *IJCAI81*, 94-100.
Renyi, Alfred
1955 "On a new axiomatic theory of probability". *Acta Mathematica Academiae Scientiarum Hungaricae* **6**, 285-333.
Russell, Bertrand
1948 *Human Knowledge: Its Scope and Limits*. New York: Simon and Schuster.
Savage, Leonard
1954 *The Foundations of Statistics*, Dover, New York.
Shafer, G.
1976 *A Mathematical Theory of Evidence*. Princeton: Princeton University Press.
Sklar, Lawrence
1970 "Is propensity a dispositional concept?" *Journal of Philosophy* **67**, 355-366.
1973 "Unfair to frequencies." *Journal of Philosophy* **70**, 41-52.
Skyrms, Brian
1980 *Causal Necessity*, Yale University Press, New Haven.
van Fraassen, Bas
1981 *The Scientific Image*. Oxford: Oxford University Press.
Venn, John

1888 *The Logic of Chance*, 3rd ed. London.