Probabilities for AI[1]

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

Abstract

Probability plays an essential role in many branches of AI, where it is typically assumed that we have a complete probability distribution when addressing a problem. But this is unrealistic for problems of real-world complexity. 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. Introduction

AI aims at multiple goals, and probability plays an essential role in most of them. One of the ultimate aspirations of AI is the construction of agents of human-level intelligence, capable of operating in environments of real-world complexity (in short, "generally intelligent agents", or GIAs). For many years this problem was largely set aside as being too hard for existing AI technology, although there has been a recent resurgence of interest in GIAs. A more modest goal of AI is the construction of applications that can provide intelligent assistance to human agents. Either goal frequently requires AI systems to use and make inferences about probabilities, and as we will see, both goals encounter similar problems in their use of probabilities.

GIAs are faced with environments about which they have only limited knowledge. They must be able to expand their knowledge base, and use that knowledge to guide their activity. Just like human beings, they will have to be able to discover new regularities in the world, but these will not generally be exceptionless regularities. Much of that knowledge will be probabilistic. Their

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

reasoning about how to act must then be based on this probabilistic knowledge.

On the other hand, at least some kinds of AI assistants may not be required to discover new generalizations. Perhaps the human operators can be relied upon to provide the requisite general knowledge of the world, and then the AI assistants will reason from there. However, for humans too, most of our general knowledge of the world is probabilistic. We know that if there are certain kinds of clouds, it will probably rain, if we turn the key in the ignition of our car, it will probably start, and so forth. Very little of our general knowledge is of exceptionless laws of nature.

In their reasoning about probabilities, both GIAs and AI-assisted humans will face a general epistemological problem that have not been adequately addressed in the AI literature. AI researchers often assume that when a problem is addressed by either a GIA or an AI-assisted human, they will come to the problem equipped with knowledge of a complete probability distribution. The first problem for this assumption is that in a sufficiently complex environment it would be impossible to store a complete probability distribution in an AI system. In general, given n simple propositions, it will take 2^n logically independent probabilities to specify a complete probability distribution. For a rather small number of simple propositions, this 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. 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 constituents of a complete probability distribution, we would have to know 5 – 10 orders of magnitude more logically independent probabilities than the number of elementary particles in the universe.

An obvious problem is that if an AI system had to store all of these probabilities explicitly, it would have to have more memory than there are elementary particles in the universe. Sometimes this problem can be alleviated by assuming that most of the propositions under consideration are statistically independent of each other. That enables us to store the probabilities in a Bayesian net, which only requires us to explicitly store probabilities where independence fails. It can reasonably be oubted that there will always be enough statistical independence for this problem to be solved using Bayesian nets. But let us set that aside and focus on the epistemological problem. To use Bayesian nets in this way, we have to know what propositions are statistically independent of each other. So the human agent, or the GIA, would still have to know the values of all the 2^n logically independent probabilities required for specifying a complete probability distribution. In other words, the use of Bayesian nets may alleviate the storage problem, but not the epistemological problem of knowing the values of the probabilities required for constructing a Bayesian net.

In applying probabilities to real-world problems, researchers typically fill in many of the gaps in their knowledge by simply assuming statistical independence when they have no information to the contrary. This is strategy is often employed in the construction of Bayesian nets, but such assumptions are also made more generally. When they see no apparent connection between two kinds of events A and B, researchers assume that the probability of A occurring is independent of whether B occurs, i.e., $\text{prob}(A\&B) = \text{prob}(A)\cdot\text{prob}(B)$. Such assumptions are "defeasible", in the sense that they may be reasonable assumptions given what the researcher knows initially, but further knowledge could, at least in principle, make it clear that A and B are not really statistically independent.

Defeasible assumptions of statistical independence can go a long way towards filling the gaps in our knowledge of probability distributions. However, deciding which independence assumptions to make has usually been based on nothing but untutored intuition. AI researchers have lacked formal tools for choosing independence assumptions. The reason this is a problem is that different sets of seemingly reasonable independence assumptions are often inconsistent with each other. How do we decide which set of assumptions to adopt? Untutored intuition often fails us here, and the probability calculus is of no help. For example, consider a community with building codes that specify that only commercial buildings can be painted grey, and also specify that only commercial buildings can be multi-storey. Let A = *painted grey*, B = *multi-storey*, C = *building in this community*, and D = *commercial building in this community*. Suppose $\text{prob}(A/C) = r$, $\text{prob}(B/C) = s$, and $\text{prob}(D/C) = d$. It is tempting to assume that A and B are independent relative to C, and so $\text{prob}(A\&B/C) = r\cdot s$. But it is equally tempting to assume that A and B are independent relative to D. However, it is impossible for both of these independences to hold. ($A\&C$) and ($B\&C$) are both subproperties of (i.e., logically entail) D. It follows by the probability calculus that $\text{prob}(A/D) = \frac{r}{d}$ and $\text{prob}(B/D) =$

$\frac{s}{d}$. So if A and B are independent relative to D, prob($A\&B/D$) = $\frac{r \cdot s}{d^2}$. However, it is also true that D entails C. It then follows from the probability calculus that prob($A\&B/C$) = $\frac{r \cdot s}{d}$. Thus if $d \neq 1$, A and B cannot be independent relative to both C and D. Once this conflict is discovered, intuition alone might leave us wondering which independence assumption we should make.

The preceding example illustrates that a blanket assumption of statistical independence for all cases in which such assumptions seem initially reasonable will often by inconsistent with the probability calculus. The following theorem of the probability calculus is another illustration of this phenomenon:

Theorem: If A,B,C each entail U and

(a) prob(C/B & A) = prob(C/A);
(b) prob(C/B & ~A) = prob(C/U & ~A);
(c) prob(C/A) ≠ prob(C/U); and
(d) prob(B/A) ≠ prob(B/U);
then prob(C/B) ≠ prob(C/U).

In other words, if (c) and (d) hold, then the pair of independence assumptions in (a) and (b) are inconsistent with the assumption that C is independent of B. The upshot is that defeasible assumptions of independence can help alleviate the epistemological problem, but we need a theory to guide us in making defeasible assumptions of statistical independence, because our untutored intuitions will often lead us into contradiction.

Of course, even with such assumptions of independence, there will be a vast number of useful probabilities we will not know. Discovering the values of interesting probabilities is a difficult epistemic task. In the sciences, researchers get journal publications out of the discovery of new probabilistic generalizations, and even in everyday life, we usually have to observe many repeated occurrences of events before we can estimate probabilities. This problem can be illustrated by the common need for "joint probabilities". Consider a medical diagnosis problem. Think of Bernard, who has symptoms suggesting a particular disease, and tests positive on two unrelated tests for the disease. Suppose the probability of a person with those symptoms having the disease is .6. Suppose the probability of a person with those symptoms having the disease if they also test positive on the first test is .7, and the probability of their having the disease if they have those symptoms and test positive on the second test is .75. What is the *joint probability* of their having the disease if they have those symptoms and test positive on both tests? The probability calculus is of no help here. Given the preceding assumptions, it is consistent with the probability calculus for the joint probability to be anything from 0 to 1. Humans, on the other hand, when faced with a problem like this, expect the joint probability to be higher than the probability of having the disease given only that one tests positive on one of the tests. Such problems of predicting joint probabilities are ubiquitous in the real-world use of probabilities. Statistical investigation gives us knowledge of the component probabilities, but we frequently have no concrete data enabling us to estimate the joint probabilities, and it is often the joint probabilities that we need — not the component probabilities by themselves. A complete probability distribution would contain explicit knowledge of all the joint probabilities, but that is unrealistic. We rarely have the data required to make explicit statistical inferences about joint probabilities.

The upshot is that for sufficiently complex problems, we will typically fall far short of having a complete probability distribution. Our GIAs and AI assistants must accomodate this fact. For either purpose, we need AI systems that do not require knowledge of complete probability distributions. This paper explores one possibility for dealing with this problem. It will be argued that, just as it often seems reasonable to make defeasible assumptions of statistical independence, it can also be reasonable to make other defeasible assumptions about probabilities that cannot be computed just by applying the probability calculus to probabilities we already know. The core idea will be that there are inferences not licensed by the probability calculus which are nevertheless almost certain to produce correct results regarding unknown probabilities. In other words, the second-order probability of the conclusion (about unknown probabilities) being true given that the premises (about known probabilities) are true is extremely high. Among these inferences will be inferences about statistical independence, so this promises to resolve the aforementioned problem of selecting which assumptions of statistical independence to make. To justify these claims, and to make sense

of the second-order probabilities involved, we must focus on what kind of probability we are talking about. Thus the next section briefly surveys the variety of kinds of probability discussed in the literature on the foundations of probability theory.

2. Varieties of Probability

2.1 Subjective Probability

Early approaches to probability theory tended to focus on objective probability, but in response to perceived difficulties for objective probability, subjective probability became the dominant variety of probability in the last half of the twentieth century, and retains that status today. The basic ideas underlying subjective probability were introduced first by Frank Ramsey (1926), but did not have much impact at the time. They were rediscovered by Leonard Savage (1954), and it is his work that caught on and led to the dominant role of subjective probability today. The basic idea is that cognizers have varying degrees of confidence in beliefs about different propositions, and these degrees of confidence should affect what bets they are willing to accept. "Degree of belief" is a technical term, defined as follows:

> A cognizer S has degree of belief $n/(n+r)$ in a proposition P iff S would accept any bet that P is true with odds better than r:n, and S would accept any bet that P is false with odds better than n:r.

Degree of belief is supposed to be a measure of the cognizer's degree of confidence. Subjectivists assume that a cognizer has a degree of belief in every proposition.

There is no guarantee that a cognizer's degrees of belief will conform to the probability calculus, but if they do they are said to be *coherent*. The *Dutch Book Argument* is standardly used to argue that a cognizer is being irrational if its degrees of belief are not coherent. This argument turns on the notion of a Dutch book, which is a combination of bets on which a person will suffer a collective loss no matter what happens. For instance, suppose you are betting on a coin toss and are willing to accept odds of 1:2 that the coin will land heads and are also willing to accept odds of 1:2 that the coin will land tails. I could then place two bets with you, betting 50 cents against the coin landing heads and also betting 50 cents against the coin landing tails, with the result that no matter what happens I will have to pay you 50 cents on one bet but you will have to pay me $1 on the other. In other words, you have a guaranteed loss — Dutch book can be made against you. The Dutch book argument (due originally to Ramsey 1926) consists of a mathematical proof that if an agent's degrees of belief do not conform to the probability calculus then Dutch book can be made against him. It is alleged that it is irrational to put oneself in such a position, so it cannot be rational to have degrees of belief that do not conform to the probability calculus. Thus a completely rational cognizer will have degrees of belief that conform to the probability calculus, and these are the cognizer's *subjective probabilities*.

A standard objection to the Dutch Book Argument is that it is impossible for a real (resource-bounded) cognizer to have coherent degrees of belief. The difficulty is that it follows from the probability calculus (and from the Dutch Book Argument) that a necessary truth has probability 1 and a necessarily false proposition has probability 0. I assume that a GIA or an AI-assisted human cognizer is capable of reasoning about the propositions expressed by a first-order language. However, by Church's theorem, there is no algorithm for determining whether such propositions are necessary truths or necessary falsehoods. Thus there is no computationally possible way to ensure that every necessary truth is assigned probability 1 and every necessary falsehood is assigned probability 0.

Faced with this argument, subjectivists generally retreat to the position that only ideal cognizers (unconstrained by limited memory or processing speed) have coherent degrees of belief. For ideal cognizers, subjective probabilities are then identified with their actual degrees of belief. The difficulty is that neither human beings nor AI agents are ideal cognizers. So this leaves subjective probability undefined for them. To get around this difficulty, subjectivists typically define the subjective probability of P for a non-ideal agent S to be the degree of belief S *would have* in P if S were ideally rational. But this is also problematic. Given a non-ideal agent S with incoherent degrees of belief, is there any reason to think there is a unique degree of belief S would have in a proposition P if S were ideally rational? This, of course, depends upon what constraints rationality imposes, but subjectivists typically claim that as long as an agent's degrees of belief are coherent,

they cannot be criticized on grounds of rationality. In particular, subjectivists give no guidance as to how an incoherent set of degrees of belief should be altered to make it coherent. Lacking rules for converting incoherent degrees of belief into coherent degrees of belief, there is no such thing as *the* degree of belief *S* would have in *P* if *S* were ideally rational. *S* could have any degree of belief in *P* and still be rational and long as *S*'s overall set of degrees of belief is coherent.

The upshot is that subjective probability only seems to make sense for ideal agents. However, AI does not deal in ideal agents. Both GIAs and AI-assisted humans have serious resource constraints, including bounded memory and processing speed. So it does not seem that subjective probability has a place in AI.

2.2 Objective Probability

If subjective probabilities are not useful for AI, it seems we should look to 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 probabilities 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 often overlooked by contemporary probability theorists, perhaps because of the popularity of subjective probability (which has no obvious 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.[3] However, it is often objected, fairly I think, that such "finite frequency theories" are at least sometimes inadequate because our probability judgments often 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, it has been suggested that 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 be considered a hypothetical frequentist. 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 (1983, 1984, 1990), and Bacchus et al (1996). I will sketch my own proposal below.

I do not think that it should be supposed that there is just one sensible kind of generic

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

[3] Examples are Russell (1948); Braithwaite (1953); Kyburg (1961, 1974); Sklar (1970, 1973). 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.

probability. However, in my (1990) I suggested that there is a central kind of generic probability in terms of which a number of other kinds can be defined. This central kind of generic probability is what I called *nomic probability*. Nomic probabilities are supposed to be the subject matter of statistical laws of nature. Exceptionless general laws, like "All electrons are negatively charged", are not just about actual electrons, but also about all physically possible electrons. We can think of such a law as reporting that any physically possible electron would be negatively charged. This is an example of a *nomic generalization*. We can think of nomic probabilities as telling us instead that a certain proportion of physically possible objects of one sort will also have some other property. For example, we might have a law to the effect that the probability of a hadron being negatively charged is .5. We can think of this as telling us that half of all physically possible hadrons would be negatively charged.

After brief thought, most people 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 often tempted to confuse generic probabilities with probability distributions over random 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 how to treat Herman, his doctor wants to know the probability of *his* being lactose intolerant, not the probability of Slavs in general being lactose intolerant. 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 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.

What I will argue in this paper is that new mathematical results, coupled with ideas from the theory of nomic probability (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 statistical independence and direct inference, but they include many new principles as well. For example, among them is a heretofore unnoticed principle enabling us to defeasibly estimate the joint 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 how probabilities can be truly useful even when we are 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. I will just sketch the theory here. 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. For this purpose, physically possible G's cannot be identified with possible objects that are G, because the same object can be a G at one possible world and fail to be a G at another possible world. Instead, 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

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

laws) and x has the property G at w. I assume that for any nomically possible property F (i.e., property consistent with the laws of nature), the set $\mathfrak{F}$ of physically possible F's will be infinite. This follows from there being infinitely many possible worlds in which there are F's. I also assume that properties are rather coarsely individuated, in the sense that nomically equivalent properties are identical. Equivalently, if F and G are properties, $F = G$ iff $\mathfrak{F} = \mathfrak{G}$.

For properties F and G, where $\mathfrak{F}$ and $\mathfrak{G}$ are the sets of physically possible F's and G's respectively, let us define the *subproperty relation* as follows:

$F \preccurlyeq G$ iff $\mathfrak{F} \subseteq \mathfrak{G}$, i.e., iff it is physically necessary (follows from true physical laws) that $(\forall x)(Fx \rightarrow Gx)$.

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, in which case $F \preccurlyeq G$ iff the universal closure of $(F \rightarrow G)$ is physically necessary.

Proportion functions are a generalization of *measure functions* studied in mathematics in measure theory. Proportion functions are "relative measure functions". Given a suitable proportion function ρ, we could stipulate that:

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

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 proportion functions. 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.

It is convenient to be able 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 prob_x and ρ_x are variable-binding operators, binding the variable x. When there is no danger of confusion, I will typically omit the subscript "x". To simplify expressions, I will often omit the variables, writing "prob(F/G)" for "prob(Fx/Gx)" when no confusion will result.

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:

Finite Proportions:

$$\text{For finite X, } \rho(A,X) = \frac{\#(A \cap X)}{\#X}.$$

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:

$0 \le \rho(X,Y) \le 1$;

If $Y \subseteq X$ then $\rho(X,Y) = 1$;

If $Z \neq \varnothing$ and $X \cap Y \cap Z = \varnothing$ then $\rho(X \cup Y,Z) = \rho(X,Z) + \rho(Y,Z)$;

If $Z \neq \varnothing$ then $\rho(X \cap Y,Z) = \rho(X,Z) \cdot \rho(Y,X \cap Z)$.

[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.

These axioms automatically hold for relative frequencies among finite sets, so the assumption is just that they also hold for proportions among infinite sets.

Finally, I need four assumptions about proportions that go beyond merely imposing the standard axioms for the probability calculus. The four assumptions I will make are:

Universality:

If $A \subseteq B$, then $\rho(B,A) = 1$.

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

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 four principles are all theorems of elementary set theory when the sets in question are finite. For instance, the projection principle tells us is that $\rho_x(Fx/\, (\exists y)(Rxy \;\&\; Gy))$ is a weighted average of the values of $\rho_x(Fx/\, Rxy)$ for different values of y. 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 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. The crossproduct principle holds for finite sets for the simple reason that $\#(A \times B) = (\#A) \cdot (\#B)$. For infinite sets we cannot just count members any more, but the algebra is the same. It is useful to axiomatize nomic probabilities indirectly by adopting axioms for proportions because the algebra of proportions is simpler than the algebra of probabilities.

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 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.[6]

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 the President 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

[6] The statistical syllogism was first expressed in this form in Pollock (1983a), but it has a long and distinguished history going back at least to C. S Peirce in the 1880's. See also Kyburg (1974, 1977).

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 projectibility. 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 the statistical syllogism can be applied to them

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, more specific information about c that lowers the probability of its being F constitutes a defeater.

4. Limit Theorems and Probable Probabilities

I propose to solve the epistemic problem of inadequate probability knowledge by justifying a large collection of defeasible inference schemes for reasoning about probabilities. The key to doing this lies in proving some limit theorems about the algebraic properties of proportions among finite sets, and proving some general theorems that relate those limit theorems to the algebraic properties of nomic probabilities.

4.1 Probable Proportions Theorem

Let us begin with a simple example. Suppose we have a set of 10,000,000 objects. I announce that I am going to select a subset, and ask you approximately 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 "$x \underset{\delta}{\approx} y$" 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

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

Proof: See appendix.

In other words, to any given degree of approximation, the proportion of subsets of U which are such that $\rho(X,U)$ is approximately equal to .5, 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

[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).

is illuminating to plot $C(n,r)$ for variable r and various fixed values of n. See figure 1. This illustrates that the sizes of subsets of U will cluster around $\frac{n}{2}$, and they cluster more tightly as n increases. $C(n,r)$ 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.

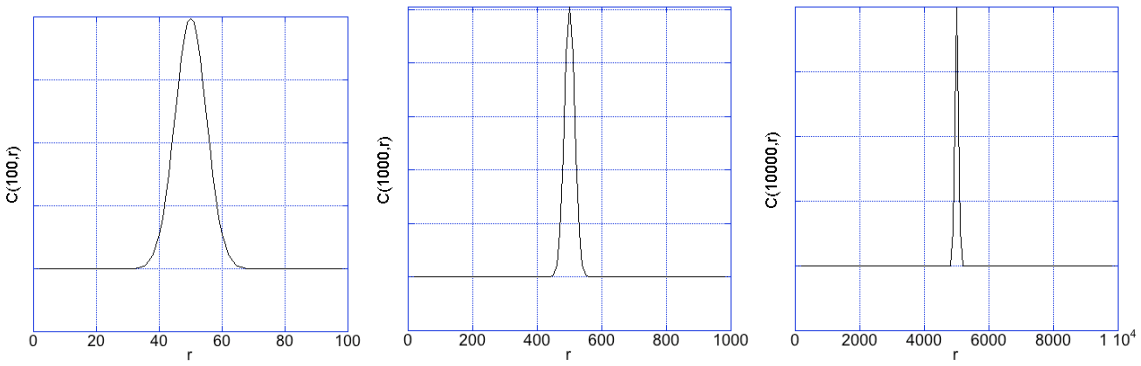

Figure 1. $C(n,r)$ for $n = 100$, $n = 1000$, and $n = 10000$.

The Finite Indifference Principle is our first example of an instance of a general combinatorial limit theorem. To state the general theorem, we need the notion of a linear constraint. Linear constraints 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 strategy will be to approximate the behavior of constraints applied to infinite domains by looking at their behavior in sufficiently large finite domains. Some linear constraints may be inconsistent with the probability calculus. We will want to rule those out of consideration, but we will need to rule out others as well. The difficulty is that there are sets of constraints that are satisfiable in infinite domains but not satisfiable in finite domains. For example, if r is an irrational number between 0 and 1, the constraint "$\rho(X,Y) = r$" is satisfiable in infinite domains but not in finite domains. Let us define:

> LC is *finitely unbounded* iff for every positive integer K there is a positive integer N such that if $\#U = N$ then $\#\left\{\langle X_1,...,X_n\rangle \mid LC \;\&\; X_1,...,X_n \subseteq U\right\} \geq K$.

For the purpose of approximating the behaviors of constraints in infinite domains by exploring their behavior in finite domains, I will confine my attention to finitely unbounded sets of linear constraints. If LC is finitely unbounded, it must be consistent with the probability calculus, but the converse is not true. I think that by appealing to limits, it should be possible to generalize the following results to all sets of linear constraints that are consistent with the probability calculus, but I will not pursue that here.

The key theorem we need is then:

Probable Proportions Theorem:

> Let $U,X_1,\ldots,X_n$ be a set of variables ranging over sets, and consider a finitely unbounded finite set LC of linear constraints on proportions between Boolean compounds of those variables. Then for any pair of relations P,Q whose variables are a subset of $U,X_1,\ldots,X_n$ there is a unique

real number r in [0,1] such that for every $\varepsilon,\delta > 0$, there is an N such that if U is finite and $\#\{\langle X_1,...,X_n\rangle \mid LC \;\&\; X_1,...,X_n \subseteq U\} \geq N$ then

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

Proof: See appendix.

Let us refer to this unique r as the *limit solution for* $\rho(P/Q)$ *given LC*. For all of the choices of constraints we will consider, finite unboundedness will be obvious, so the limit solution will exist. This theorem, which establishes the existence of the limit solution under very general circumstances, underlies all of the principles developed in this paper. It is important to realize that it is just a combinatorial theorem about finite sets, and as such is a theorem of set theory. It does not depend on any of the assumptions we have made about proportions in infinite sets. Thus far the mathematics is not philosophically questionable.

What we will actually want are particular instances of this theorem for particular choices of *LC* and specific values of *r*. An example is the Finite Indifference Principle. In general, *LC* generates a set of simultaneous equations, and the limit solution *r* can be determined by solving those equations. The simultaneous equations are the *term-characterizations* discussed in the appendix in the proof of the Probable Proportions Theorem. It turns out that these equations can be generated automatically and then solved automatically by a computer algebra program. To my surprise, neither Mathematica nor Maple has proven effective in solving these sets of equations, but I was able to write a special purpose LISP program that is fairly efficient. It computes the term-characterizations and solves them for the variable when that is possible. It can also be directed to produce a human-readable proof. If the equations constituting the term-characterizations do not have analytic solutions, they can still be solved numerically to compute the most probable values of the variables in specific cases. This software can be downloaded from http://oscarhome.soc-sci.arizona.edu/ftp/OSCAR-web-page/CODE/Code for probable probabilities.zip. I will refer to this as the *probable probabilities software*. The proofs of many of the theorems presented in this paper were generated using this software.

4.2 Limit Principle for Proportions

The Probable Proportions Theorem and its instances are mathematical theorems about finite sets. For example, the Finite Indifference Principles tells us that as $N \to \infty$, if U is finite but contains at least N members, then the proportion of subsets X of a set U which are such that $\rho(X,U) \underset{\delta}{\approx} 0.5$ goes to 1. This *suggests* that the proportion *is* 1 when U *is* infinite:

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 such infinitary principles from the corresponding finite principles. We first prove in familiar ways:

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

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

Proof**:** See appendix.

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.

The Law of Large Numbers for Proportions provides the link for moving from the behavior of linear constraints in finite sets to their behavior in infinite sets. It enables us to prove:

Limit Principle for Proportions:

Consider a finitely unbounded finite set *LC* of linear constraints on proportions between Boolean compounds of a list of variables $U,X_1,\ldots,X_n$. Let *r* be limit solution for $\rho(P/Q)$ given *LC*. Then for any infinite set *U*, for every $\delta > 0$:

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

Proof**:** See appendix.

This is our crucial "bridge theorem" that enables us to move from combinatorial theorems about finite sets to principles about proportions in infinite sets. This, together with the Probable Proportions Theorem, constitute the central theorems of this paper. They will allow us to establish many more concrete theorems. Thus, for example, from the Finite Indifference Principle we can derive:

Infinitary Indifference Principle:

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

4.4 Probable Probabilities

Nomic probabilities are proportions among physically possible objects. Recall that I have assumed that for any nomically possible property *F* (i.e., property consistent with the laws of nature), the set $\Im$ of physically possible *F*'*s* is be infinite. Thus the Limit Principle for Proportions implies an analogous principle for nomic probabilities:

Probable Probabilities Theorem:

Consider a finitely unbounded finite set *LC* of linear constraints on proportions between Boolean compounds of a list of variables $U,X_1,\ldots,X_n$. Let *r* be limit solution for $\rho(P/Q)$ given *LC*. Then for any infinite set *U*, for every $\delta > 0$,

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

Proof: See appendix.

I sometimes hear the objection that in proving theorems like the Probable Probabilities Theorem I must be making a hidden assumption about uniform distributions. It is not clear what lies behind this objection. I gave the proof. Where is the gap supposed to be? Talk of uniform distributions makes no sense as applied to either proportions or generic probabilities. I suspect that those who raise this objection are confusing generic probabilities with probability distributions over random variables, as discussed in section three.

Instances of the Probable Proportions Theorem tell us the values of the limit solutions for sets of linear constraints, and hence allow us to derive instances of the consequent of the Probable Probabilities Theorem. I will call the latter "probable probabilities principles". For example, from the Finite Indifference Principle we get:

Probabilistic Indifference Principle:

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

$\text{prob}_X\left(\text{prob}(X/G) \underset{\delta}{\approx} 0.5 \;/\; X \preccurlyeq G\right) = 1.$ [8]

[8] If we could assume countable additivity for nomic probability, the Indifference Principle would imply that $\text{prob}_X\left(\text{prob}(X/G) = 0.5 \;/\; X \preccurlyeq G\right) = 1.$ 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

4.5 Justifying Defeasible Inferences about Probabilities

Next note that we can apply the statistical syllogism to the second-order 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, if G is nomically possible then 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, if G is nomically possible then the expectable value of prob(F/G) = 0.5.

So in talking about expectable values, I am talking about this entire quadruple of principles. Our general theorem is:

Principle of Expectable Values

> Consider a finitely unbounded finite set LC of linear constraints on proportions between Boolean compounds of a list of variables $U,X_1,\ldots,X_n$. Let r be the limit solution for $\rho(P/Q)$ given LC. Then given LC, the expectable value of prob(P/Q) = r.

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. But 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) $\neq$ 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 the 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. This should be familiar from standard measure theory. 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

---

(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.

many rationals. The set of rationals is infinite, but it has 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:

$$\text{PROB}(P/Q) = \frac{\text{PROB}(P \,\&\, Q)}{\text{PROB}(Q)}.$$

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, PROB(P/Q) is undefined when 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. Statistical 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). 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. Recall that our general strategy is to formulate our assumptions as a set of finitely unbounded linear constraints, and then find the limit solution by solving the set of simultaneous equations generated by them (the term characterizations). This can usually be done using the probable probabilities software. In this case we get:

Finite Independence Principle:

For all rational numbers r,s between 0 and 1, given that $X,Y,Z \subseteq U$ & $\rho(X,Z) = r$ & $\rho(Y,Z) = s$,

the limit solution for $\rho(X \cap Y, Z)$ is $r{\cdot}s$.[9]

Proof: See appendix.

As before, this generates the four principles making up the following principle of expectable values:

[9] This illustrates that to get finite unboundedness, we often have to restrict the various parameters mentioned in *LC* to rational numbers. I am convinced that this restriction should be inessential. One can go ahead and solve the term characterizations in the same way for the cases in whichthe parameters are irrational, and I am inclined to endorse the resulting principles of probable probabilities. However, at this point I am unsure how to justify this.

Principle of Expectable Statistical Independence:

For rational numbers r,s between 0 and 1, given that 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 arbitrarily chosen 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. This solves one of the major unsolved problems of the application of probabilities to real-world problems.

6. Defeaters for Statistical Independence

Of course, 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 $r = 1$. In general, when A and B "overlap", in the sense that there is a D such that (A&C),(B&C) $\preccurlyeq$ D and prob(D/C) $\neq$ 1, then we should not expect that prob(A&B/C) = prob(A/C)·prob(B/C). This follows from the following principle of expectable probabilities:

Principle of Statistical Independence with Overlap:

If r,s,g are rational numbers between 0 and 1, given that prob(A/C) = r, prob(B/C) = s, prob(D/C) = g, (A&C) $\preccurlyeq$ D, and (B&C) $\preccurlyeq$ D, it follows that prob(A/C&D) = r/g, prob(B/C&D) = s/g, and the following values are expectable:

(1) prob(A&B/C) = $\dfrac{r \cdot s}{g}$;

(2) prob(A&B/C&D) = $\dfrac{r \cdot s}{g^2}$.

This can be proven automatically using the probable probabilities software. This proof, and the proofs of many other theorems left unproven in this paper, are also reproduced in the longer version of this paper, available on my website at http://oscarhome.soc-sci.arizona.edu/ftp/PAPERS/Probable Probabilities with proofs.pdf

To illustrate statistical independence with overlap using a simple and intuitive case, suppose $A = A_0$ & D and $B = B_0$ & D. Given no reason to think otherwise, we would expect A_0, B_0, and D to be statistically independent. But then we would expect that

$$\text{prob}(A\&B/C) = \text{prob}(A_0\&D\&B_0/C) = \text{prob}(A_0/C) \cdot \text{prob}(D/C) \cdot \text{prob}(B_0/C)$$

$$= \frac{\text{prob}(A_0 \,\&\, D \,/\, C) \cdot \text{prob}(B_0 \,\&\, D \,/\, C)}{\text{prob}(D \,/\, C)} = \frac{r \cdot s}{g}.$$

The upshot is that, given the overlap, we can expect A and B to be statistically independent relative to (C&D), but not relative to C. The second-order probability to which the statistical syllogism is applied to generate (1) in the Principle of Statistical Independence with Overlap is:

$$\text{prob}_{X,Y,Z,W}\left(\begin{array}{l}\text{prob}(X\,\&\,Y/Z)\underset{\delta}{\approx}\dfrac{r\cdot s}{g}\ /\\ X,Y,Z,W\preccurlyeq U \text{ and } \text{prob}(X/Z)=r \text{ and } \text{prob}(Y/Z)=s\\ \text{and } (X\,\&\,Z)\preccurlyeq W \text{ and } (Y\,\&\,Z)\preccurlyeq \text{ and } \text{prob}(W/Z)=g \text{ and } \text{prob}(W/U)=\zeta\\ \text{and } \text{prob}(X/U)=\alpha \text{ and } \text{prob}(Y/U)=\beta \text{ and } \text{prob}(Z/U)=\gamma\end{array}\right)=1.$$

On the other hand, the second-order probability to which the statistical syllogism is applied to generate the Principle of Statistical Independence was:

$$\text{prob}_{X,Y,Z}\left(\begin{array}{l}\text{prob}(X\,\&\,Y/Z)\underset{\delta}{\approx}r\cdot s\ /\\ X,Y,Z\preccurlyeq U \text{ and } \text{prob}(X/Z)=r \text{ and } \text{prob}(Y/Z)=s\\ \text{and } \text{prob}(X/U)=\alpha \text{ and } \text{prob}(Y/U)=\beta \text{ and } \text{prob}(Z/U)=\gamma\end{array}\right)=1.$$

The former probability takes account of more information than the latter, so it provides a subproperty defeater for the use of the statistical syllogism and hence an undercutting defeater for the Principle of Statistical Independence:

Overlap Defeat for Statistical Independence:

⌜(*A*&*C*) ≼ *D*, (*B*&*C*) ≼ *D*, and prob(*D*/*C*) ≠ 1⌝ is an undercutting defeater for the inference from ⌜prob(*A*/*C*) = *r* and prob(*B*/*C*) = *s*⌝ to ⌜prob(*A*&*B*/*C*) = *r* · *s*⌝ by the Principle of Statistical Independence.

Suppose you know that prob(*A*/*C*) = *r* and prob(*B*/*C*) = *s*, and are inclined to infer that prob(*A*&*B*/*C*) = *r*·*s*. As long as *r*,*s* < 1, there will always be a *D* such that (*A*&*C*) ≼ *D*, (*B*&*C*) ≼ *D*, and prob(*D*/*C*) ≠ 1. Does this mean that the inference is always defeated? It does not, but understanding why is a bit complicated. First, what we know in general is the existential generalization (∃*D*)[(*A*&*C*) ≼ *D* and (*B*&*C*) ≼ *D* and prob(*D*/*C*) ≠ 1]. But the defeater requires knowing of a specific such *D*. The reason for this is that it is not true in general that prob(*Fx*/*Rxy*) = prob(*Fx*/(∃*y*)*Rxy*). For example, let *Fx* be "*x* = 1" and let *Rxy* be "*x* < *y* & *x*,*y* are natural numbers ≤ 2". Then prob(*Fx*/*Rxy*) = ⅓, but prob(*Fx*/(∃*y*)*Rxy*) = ½. Accordingly, we cannot assume that

$$\text{prob}_{X,Y,Z,W}\left(\begin{array}{l}\text{prob}(X\,\&\,Y/Z)\underset{\delta}{\approx}r\cdot s\ /\\ X,Y,Z,W\preccurlyeq U \text{ and } \text{prob}(X/Z)=r \text{ and } \text{prob}(Y/Z)=s \text{ and}\\ (X\,\&\,Z)\preccurlyeq W \text{ and } (Y\,\&\,Z)\preccurlyeq \text{ and } \text{prob}(W/Z)=g \text{ and } \text{prob}(W/U)=\zeta\\ \text{and } \text{prob}(X/U)=\alpha \text{ and } \text{prob}(Y/U)=\beta \text{ and } \text{prob}(Z/U)=\gamma\end{array}\right)$$

$$=\text{prob}_{X,Y,Z}\left(\begin{array}{l}\text{prob}(X\,\&\,Y/Z)\underset{\delta}{\approx}r\cdot s\ /\\ (\exists W)(\exists g)(\exists\zeta)[X,Y,Z,W\preccurlyeq U \text{ and } \text{prob}(X/Z)=r \text{ and } \text{prob}(Y/Z)=s \text{ and}\\ (X\,\&\,Z)\preccurlyeq W \text{ and } (Y\,\&\,Z)\preccurlyeq \text{ and } \text{prob}(W/Z)=g \text{ and } \text{prob}(W/U)=\zeta\\ \text{and } \text{prob}(X/U)=\alpha \text{ and } \text{prob}(Y/U)=\beta \text{ and } \text{prob}(Z/U)=\gamma]\end{array}\right)$$

and hence merely knowing that (∃*D*)[(*A*&*C*) ≼ *D* and (*B*&*C*) ≼ *D* and prob(*D*/*C*) ≠ 1] does not give us a defeater. In fact, it is a theorem of the calculus of nomic probabilities that if □[*B* → *C*] then prob(*A*/*B*) = prob(*A*/*B*&*C*). So because

□[(prob(*A*/*C*) = *r* and *r*,*s* < 1 and prob(*B*/*C*) = *s*)
→ (∃*D*)(∃*g*)(∃ζ)[(*A*&*C*) ≼ *D* and (*B*&*C*) ≼ *D* and prob(*D*/*C*) = *g* and prob(*D*/*U*) = ζ]]

it follows that

$$\text{prob}_{X,Y,Z}\left(\begin{array}{l}\text{prob}(X\,\&\,Y/Z)\underset{\delta}{\approx} r\cdot s\ /\\ (\exists W)(\exists g)(\exists\zeta)[X,Y,Z,W \preccurlyeq U \text{ and } \text{prob}(X/Z)=r \text{ and } \text{prob}(Y/Z)=s \text{ and}\\ (X\,\&\,Z)\preccurlyeq W \text{ and } (Y\,\&\,Z)\preccurlyeq \text{ and } \text{prob}(W/Z)=g \text{ and } \text{prob}(W/U)=\zeta\\ \text{and } \text{prob}(X/U)=\alpha \text{ and } \text{prob}(Y/U)=\beta \text{ and } \text{prob}(Z/U)=\gamma]\end{array}\right)=1.$$

Hence the mere fact that there always *is such a D* does not automatically give us a defeater for the application of the Principle of Statistical Independence. To get defeat, we must know of some *specific* D such that $(A\&C)\preccurlyeq D$ and $(B\&C)\preccurlyeq D$ and $\text{prob}(D/C)\neq 1$.

But now it may occur to the reader that there is a second strategy for generating automatic defeat. We can always construct a specific such D, namely, $(A\vee B)$. However, it turns out that this choice of D does not give us a defeater. In fact,

$$\text{prob}_{X,Y,Z}\left(\begin{array}{l}\text{prob}(X\,\&\,Y/Z)\underset{\delta}{\approx} r\cdot s\ /\\ X,Y,Z\preccurlyeq U \text{ and } \text{prob}(X/Z)=r \text{ and } \text{prob}(Y/Z)=s \text{ and}\\ (X\,\&\,Z)\preccurlyeq X\vee Y \text{ and } (Y\,\&\,Z)\preccurlyeq X\vee Y \text{ and } (\exists g)\text{prob}(X\vee Y/Z)=g \text{ and}\\ (\exists\zeta)\text{prob}(X\vee Y/U)=\zeta \text{ and } \text{prob}(X/U)=\alpha \text{ and } \text{prob}(Y/U)=\beta \text{ and } \text{prob}(Z/U)=\gamma\end{array}\right)$$

$$=\text{prob}_{X,Y,Z}\left(\begin{array}{l}\text{prob}(X\,\&\,Y/Z)\underset{\delta}{\approx} r\cdot s\ /\\ X,Y,Z\preccurlyeq U \text{ and } \text{prob}(X/Z)=r \text{ and } \text{prob}(Y/Z)=s\\ \text{and } \text{prob}(X/U)=\alpha \text{ and } \text{prob}(Y/U)=\beta \text{ and } \text{prob}(Z/U)=\gamma\end{array}\right)=1.$$

This is because, once again,

$$\begin{array}{l}\Box[(\text{prob}(A/C)=r \text{ and } \text{prob}(B/C)=s)\\ \quad\rightarrow(\exists g)(\exists\zeta)[(A\&C)\preccurlyeq(A\vee B) \text{ and } (B\&C)\preccurlyeq(A\vee B)\\ \qquad\qquad \text{and } \text{prob}(A\vee B/C)=g \text{ and } \text{prob}(A\vee B/U)=\zeta]].\end{array}$$

Notice that the latter depends upon our not knowing the value of g. If we do know that $\text{prob}(A/C)=r$, $\text{prob}(B/C)=s$, and $\text{prob}(A\vee B/C)=g$, then we can simply compute by the probability calculus that $\text{prob}(A\&B/C)=r+s-g$, in which case the application of the defeasible inference to the contrary conclusion is conclusively defeated.

The preceding can be generalized. There are many ways of automatically generating properties D such that $(A\&C)\preccurlyeq D$ and $(B\&C)\preccurlyeq D$. For example, given some fixed set E, we can define:

$$\mu(A,B)=A\vee B\vee E.$$

But again,

$$\begin{array}{l}\Box[(\text{prob}(A/C)=r \text{ and } \text{prob}(B/C)=s)\\ \quad\rightarrow(\exists g)(\exists\zeta)[(A\&C)\preccurlyeq\mu(A,B) \text{ and } (B\&C)\preccurlyeq\mu(A,B)\\ \qquad\qquad \text{and } \text{prob}(\mu(A,B)/C)=g \text{ and } \text{prob}(\mu(A,B)/U)=\zeta]]\end{array}$$

so

$$\text{prob}_{X,Y,Z}\left(\begin{array}{l}\text{prob}(X\,\&\,Y/Z)\underset{\delta}{\approx} r\cdot s\ /\\ X,Y,Z\preccurlyeq U \text{ and } \text{prob}(X/Z)=r \text{ and } \text{prob}(Y/Z)=s \text{ and } (X\,\&\,Z)\preccurlyeq\mu(X,Y)\\ \text{and } (Y\,\&\,Z)\preccurlyeq\mu(X,Y) \text{ and } (\exists g)\text{prob}(\mu(X,Y)/Z)=g \text{ and } (\exists\zeta)\text{prob}(\mu(X,Y)/U)=\zeta\\ \text{and } \text{prob}(X/U)=\alpha \text{ and } \text{prob}(Y/U)=\beta \text{ and } \text{prob}(Z/U)=\gamma\end{array}\right)=1.$$

These observations illustrate a general phenomenon that will recur for all of our defeasible principles of expectable probabilities. Defeaters cannot be generated by functions that apply automatically to the properties involved in the inference. For example, in obtaining overlap defeaters for the Principle of Statistical Independence, we must have some substantive way of picking out D that does not pick it out simply by reference to A, B, and C.

In sections seven and nine we will encounter additional undercutting defeaters for the Principle of Statistical Independence.

7. Nonclassical Direct Inference

Pollock (1984) noted (using different terminology) the following principle of probable probabilities:

Nonclassical Direct Inference:

If r is a rational number between 0 and 1, and $\text{prob}(A/B) = r$, the expectable value of $\text{prob}(A/B\&C) = r$.

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 eight, principles of classical direct inference can be derived from it.

Probability theorists have not taken formal note of the Principle of Nonclassical Direct Inference, but they often reason in accordance with it. 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. In fact, the Principle of Nonclassical Direct Inference is equivalent (with one slight qualification) to the defeasible Principle of Statistical Independence. This turns upon the following simple theorem of the probability calculus:

Independence and Direct Inference Theorem:

If $\text{prob}(C/B) > 0$ then $\text{prob}(A/B\&C) = \text{prob}(A/B)$ iff $\text{prob}(A\&C/B) = \text{prob}(A/B)\cdot\text{prob}(C/B)$.

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.

Nonclassical Direct Inference is a principle of defeasible reasoning, 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:

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).

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:

Principle of Statistical Independence with Subproperties:

If prob(A/C) = r, prob(B/C) = s, (B&C) $\preccurlyeq D \preccurlyeq C$, and prob($A/D$) = $p \neq r$, then the expectable value of prob(A&B/C) = $p \cdot s$ (rather than $r \cdot s$).

Subproperty Defeat for Statistical Independence:

⌜(B&C) $\preccurlyeq D \preccurlyeq C$ and prob(A/D) = $p \neq r$⌝ is an undercutting defeater for the inference by the Principle of Statistical Independence from ⌜prob(A/C) = r & prob(B/C) = s⌝ to ⌜prob(A&B/C) = $r \cdot s$⌝.

Consider an example of subproperty defeat for Statistical Independence. Suppose we know that prob(x is more than a year old/x is a vertebrate) = 0.15, and prob(x is a fish/x is a vertebrate) = 0.8, and we want to know the value of prob(x is more than a year old & x is a fish/x is a vertebrate). In the absence of any other information it would be reasonable to assume that being a fish and being more than a year old are statistically independent relative to "x is a vertebrate", and hence prob(x is more than a year old & x is a fish/x is a vertebrate) = $0.15 \cdot 0.8 = 0.12$. But suppose we also know prob(x is more than a year old/x is an acquatic animal) = 0.2. Should this make a difference? Relying upon untutored intuition may leave one unsure. However, being a vertebrate and a fish entails being an acquatic animal, so additional information gives us a subproperty defeater for the assumption of statistical independence. What we should conclude instead is that prob(x is more than a year old & x is a fish/x is a vertebrate) = $0.2 \cdot 0.8 = 0.16$.

By virtue of the equivalence of the principles of Nonclassical Direct Inference and Statistical Independence, defeaters for the Principle of Statistical Independence also yield defeaters for Nonclassical Direct Inference. In particular, overlap defeaters for the Principle of Statistical Independence yield overlap defeaters for Nonclassical Direct Inference. We have the following theorem:

Principle of Nonclassical Direct Inference with Overlap:

If B&$D \preccurlyeq G$ and C&$D \preccurlyeq G$ then the expectable value of prob(B/C&D) = $\dfrac{\text{prob}(B/D)}{\text{prob}(G/D)}$.

Note that if $G \preccurlyeq D$ then $\dfrac{\text{prob}(B/D)}{\text{prob}(G/D)} = \text{prob}(B/G)$, so ⌜$B$&$D$,$C$&$D \preccurlyeq G \preccurlyeq D$⌝ is a defeasible reason for ⌜prob(B/C&D) = prob(B/G)⌝.

This is an interesting generalization of Nonclassical Direct Inference. Although probabilists commonly reason in accordance with Nonclassical Direct Inference in practical applications (without endorsing the formal principle), untutored intuition is not apt to lead them to reason in accordance with Nonclassical Direct Inference with Overlap. To the best of my knowledge, Nonclassical Direct Inference with Overlap has gone unnoticed in the probability literature. Nonclassical Direct Inference with Overlap yields the standard principle of Nonclassical Direct Inference when D is tautologous.

Nonclassical Direct Inference with Overlap is subject to both subproperty defeat and overlap defeat, just as the standard principle is:

Subproperty Defeat for Nonclassical Direct Inference with Overlap:

⌜(C&D) $\prec E \prec D$ and prob(B/E) $\neq r$⌝ is an undercutting defeater for the inference by Nonclassical Direct Inference with Overlap from ⌜B&D,C&$D \preccurlyeq G$⌝ to ⌜prob(B/C&D) = $\dfrac{\text{prob}(B/D)}{\text{prob}(G/D)}$⌝.

Overlap Defeat for Nonclassical Direct Inference with Overlap:

$\ulcorner B\&D \preccurlyeq H,\ C\&D \preccurlyeq H$ and $\text{prob}(Gx/Dx) \neq \text{prob}(Hx/Dx)\urcorner$ is an undercutting defeater for the inference by Nonclassical Direct Inference with Overlap from $\ulcorner B\&D \preccurlyeq G$ and $C\&D \preccurlyeq G\urcorner$ to $\ulcorner \text{prob}(B/C\&D) = \dfrac{\text{prob}(B/D)}{\text{prob}(G/D)}\urcorner$.

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, it was shown in Pollock (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 Pollock (1984 and 1990).[10] Let $\mathbf{K}$ be the conjunction of all the propositions the agent is warranted in believing,[11] and let $\mathfrak{K}$ 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) (written in small caps) to be the proportion of $\mathbf{K}$-worlds at which P is true. Where $\mathcal{P}$ is the set of all physically possible P-worlds:

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

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

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

This makes singular probabilities sensitive to the agent's knowledge of his situation, which is what is needed for rational decision making.[12] Formally, singular probabilities become analogous to Carnap's (1950, 1952) logical probability, with the important difference that Carnap took ρ to be logically specified, whereas here the identity of ρ is taken to be a 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) $\text{PROB}(Fa) = \text{prob}(Fx/x = a\ \&\ \mathbf{K})$;

(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 $\text{PROB}(P/Q) = \text{prob}(Rx_1\ldots x_n/Sx_1\ldots x_n\ \&\ x_1 = a_1\ \&\ \ldots\ \&\ x_n = a_n\ \&\ \mathbf{K})$.

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

Proof: See appendix.

PROB(P) is a kind of "mixed physical/epistemic probability", because it combines background knowledge in the form of $\mathbf{K}$ with nomic probabilities.

The probability $\text{prob}(Fx/x = a\ \&\ \mathbf{K})$ is a peculiar–looking nomic probability. It is a 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 $\text{PROB}(Fa) = \text{prob}(Fx/x = a\ \&\ \mathbf{K}) = \text{prob}(Fx/Gx)$. This is a familiar form of "classical" direct inference — that is, direct inference from generic probabilities to singular probabilities. More generally, we can derive:

[10] Bacchus (1990) gave a somewhat similar account of direct inference, drawing on Pollock (1983 ,1984).
[11] What an agent is justified in believing at a time depends on how much reasoning he has done. A proposition is warranted for an agent iff the agent would be justified in believing it if he could do all the relevant reasoning.
[12] For a further complication, see the literature on causal probability, as discussed for example in Pollock (2006).

Classical Direct Inference:

$\ulcorner Sa_1\ldots a_n$ is known and prob$(Rx_1\ldots x_n/\ Sx_1\ldots x_n \,\&\, Tx_1\ldots x_n) = r\urcorner$ is a defeasible reason for $\ulcorner$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 prob$(Rx_1\ldots x_n/\ Vx_1\ldots x_n \,\&\, Tx_1\ldots x_n) \neq r\urcorner$ is an undercutting defeater for the inference by classical direct inference from $\ulcorner Sa_1\ldots a_n$ is known and prob$(Rx_1\ldots x_n/\ Sx_1\ldots x_n \,\&\, Tx_1\ldots x_n) = r\urcorner$ to $\ulcorner$PROB$(Ra_1\ldots a_n\ /\ Ta_1\ldots a_n) = r\urcorner$.

Classical Direct Inference and Subproperty Defeat are (versions of) the two best known principles of direct inference. Pollock (1983) proposed them as precizations of Reichenbach's seminal principles of direct inference, and Kyburg (1974) and Bacchus (1990) built their theories around similar principles. However, as Kyburg was the first to observe, these two principles do not constitute a complete theory of direct inference. This is illustrated by overlap defeat, and we will find other defeaters too as we proceed:

Overlap Defeat for Classical Direct Inference:

The conjunction of

(i) $Rx_1\ldots x_n \,\&\, Sx_1\ldots x_n \,\&\, Tx_1\ldots x_n \preccurlyeq Gx_1\ldots x_n$ and

(ii) $(Sx_1\ldots x_n \,\&\, Tx_1\ldots x_n \,\&\, x_1 = a_1 \,\&\, \ldots \,\&\, x_n = a_n \,\&\, \mathbf{K}) \preccurlyeq Gx_1\ldots x_n$ and

(iii) prob$(Gx_1\ldots x_n/\ Sx_1\ldots x_n \,\&\, Tx_1\ldots x_n) \neq 1$

is an undercutting defeater for the inference by classical direct inference from $\ulcorner Sa_1\ldots a_n$ is known and prob$(Rx_1\ldots x_n/\ Sx_1\ldots x_n \,\&\, Tx_1\ldots x_n) = r\urcorner$ to $\ulcorner$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 $\ulcorner$PROB$(P/Q) = r\urcorner$ gives us a defeasible reason for expecting that PROB$(P/Q\&R) = r$. We can employ principles of statistical independence similarly. For example, $\ulcorner$PROB$(P/R) = r$ & PROB$(Q/R) = s\urcorner$ gives us a defeasible reason for expecting that PROB$(P\&Q/R) = r{\cdot}s$. And we get principles of subproperty defeat and overlap defeat for these applications of Nonclassical Direct Inference and Statistical Independence that are exactly analogous to the principles for generic probabilities.

9. Computational Inheritance

The biggest problem faced by most theories of direct inference concerns what to do if we have information supporting conflicting direct inferences. For example, suppose Bernard has symptoms suggesting, with probability .6, that he has a certain rare disease. Suppose furthere that we have two seemingly unrelated diagnostic tests for a disease, and Bernard tests positive on both tests. We know that the probability of a person with his symptoms having the disease if he tests positive on the first test is .7, 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. It is consistent with the probability calculus for the "joint probability" of his having the disease if he tests positive on both tests to be anything from 0 to 1. The Principle of Classical Direct inference as formulated in section eight is no help either. Direct inference gives us one reason for thinking the probability of Bernard having the disease is .7, and it gives us a different reason for drawing the conflicting conclusion that the probability is .75. The result, endorsed in Pollock (1990), is that both instances of Classical Direct Inference are defeated (it is a case of collective defeat), and we are left with no conclusion to draw about the singular probability of Bernard's having the disease. Because this sort of situation is so common, Classical

Direct Inference is not generally very useful. Kyburg (1974) tried to do better by proposing that Direct Inference locates singular probabilities in intervals. In this case his conclusion would be that the probability of Bernard having the disease is (or lies in the interval) [.7,.75]. But intuitively, this also seems unsatisfactory. If Bernard tests positive on both tests, the probability of his having the disease should be higher than if he tests positive on just one, so it should lie *above* the interval [.7,.75]. But how can we justify this?

Knowledge of generic probabilities would be vastly more useful in real application if there were a function $\Upsilon(r,s\,|a)$ such that when $\text{prob}(F/U) = a$, $G,H \preccurlyeq U$, $\text{prob}(F/G) = r$ and $\text{prob}(F/H) = s$ we could defeasibly expect that $\text{prob}(F/G\&H) = \Upsilon(r,s\,|a)$, and hence (by Nonclassical Direct Inference) that $\textsc{prob}(Fc) = \Upsilon(r,s\,|a)$. I call this *computational inheritance*, because it computes a new value for $\textsc{prob}(Fc)$ from previously known generic probabilities. Direct inference, by contrast, is a kind of "noncomputational inheritance". It is *direct* in that $\textsc{prob}(Fc)$ 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.

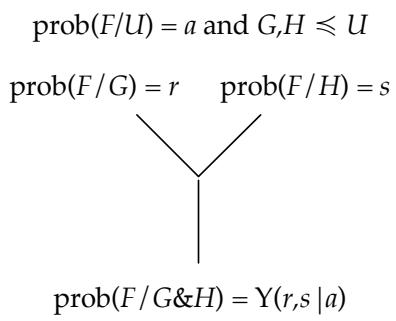

Figure 2. The Y-function

Following Reichenbach (1949), it has generally been assumed that there is no such function as the Y-function. Certainly, there is no function $\Upsilon(r,s\,|a)$ such that we can conclude *deductively* that $\text{prob}(F/G\&H) = \Upsilon(r,s\,|a)$. For any r, s, and a that are neither 0 nor 1, $\text{prob}(F/G\&H)$ can take any value between 0 and 1. However, that is equally true for Nonclassical Direct Inference. That is, if $\text{prob}(F/G) = r$ we cannot conclude deductively that $\text{prob}(F/G\&H) = r$. 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 $\Upsilon(r,s\,|a)$ such that we can defeasibly expect $\text{prob}(F/G\&H)$ to be $\Upsilon(r,s\,|a)$? It follows from the Probable Probabilities Theorem that the answer is "Yes". Let us define:

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

I use the non-standard notation "$\Upsilon(r,s\,|a)$" rather than "$\Upsilon(r,s,a)$" 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) $\text{prob}(C/\,B\,\&\,A) = \text{prob}(C/A)$

and

(b) $\text{prob}(C/B\,\&\,{\sim}A) = \text{prob}(C/U\,\&\,{\sim}A)$.

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

Y-Theorem:

Let r = prob(A/B), s = prob(A/C), a = prob(A/U), and $0 < a < 1$. If B and C are Y-independent for A relative to U then prob(A/B&C) = 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 combine 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 expectable values (which can also be proven directly using the probable probabilities software):

Y-Principle:

If B,$C \preccurlyeq U$, prob(A/B) = r, prob(A/C) = s, prob(A/U) = a, prob(B/U) = b, prob(C/U) = c, and $0 < a < 1$, then the expectable value of prob(A/B & C) = Y(r,s |a).

Note that the expectable value of prob(A/B & C) is independent of b and c.

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

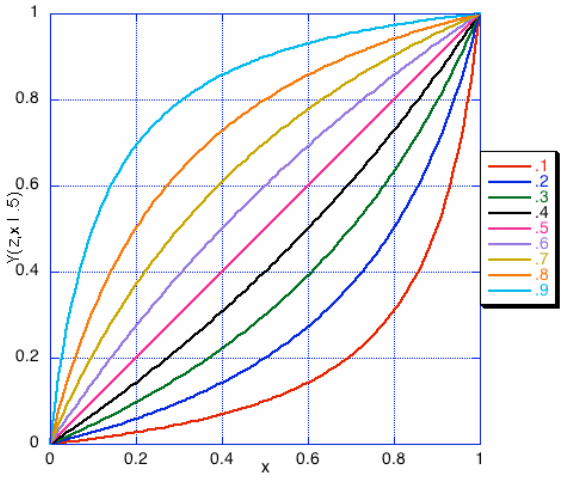

Figure 3. Y(z,x |.5), holding z constant
(for several choices of z as indicated in the key).

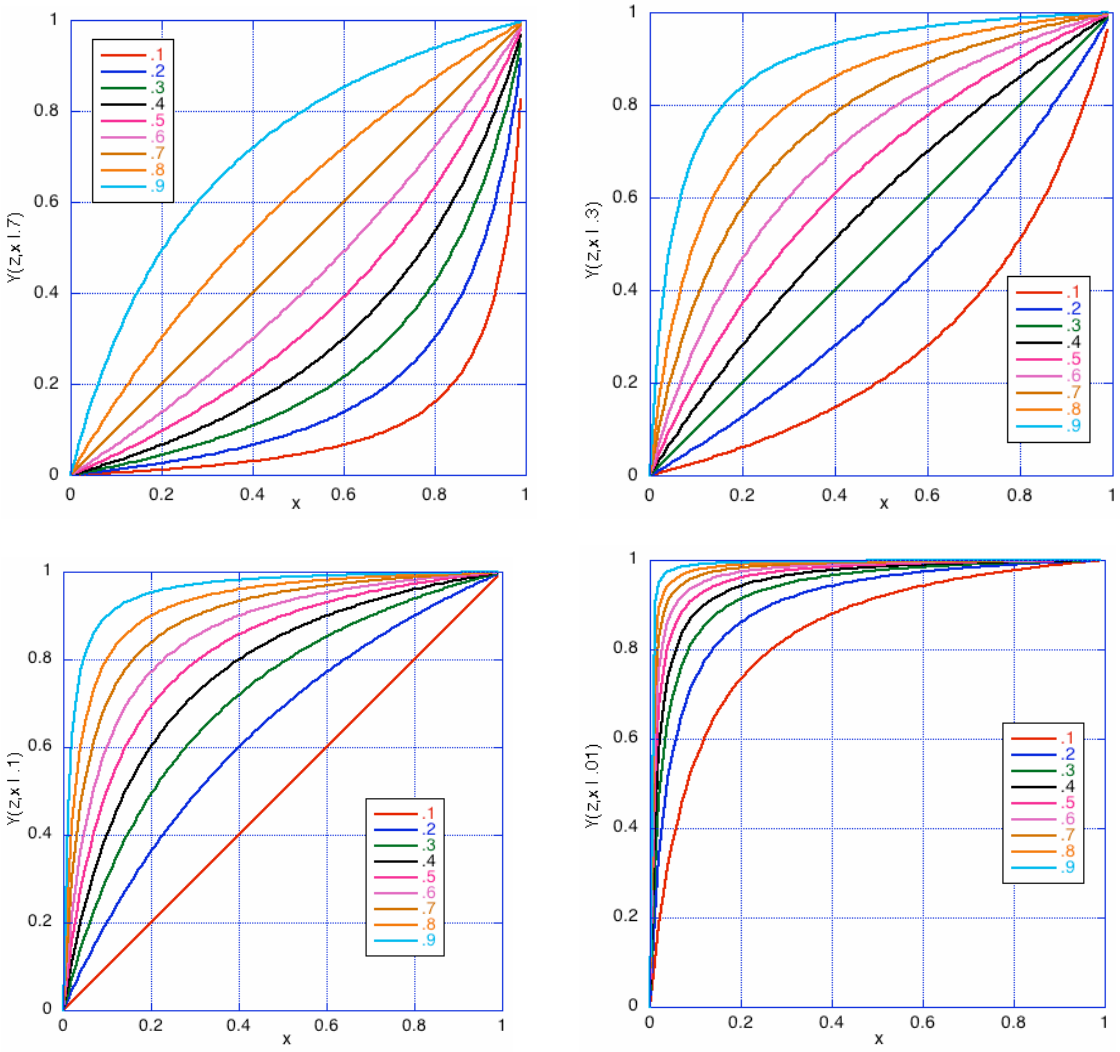

Figure 4. $Y(z,x\,|a)$ holding z constant (for several choices of z), for $a = .7$, $a = .3$, $a = .1$, and $a = .01$.

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.

Note that, in general, when $r,s < a$ then $Y(r,s\,|a) < r$ and $Y(r,s\,|a) < s$, and when $r,s > a$ then $Y(r,s\,|a) > r$ and $Y(r,s\,|a) > s$.

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:

Y-commutativity: $Y(r,s\,|a) = Y(s,r\,|a)$.

Y-associativity: $Y(r,Y(s,t\,|a)\,|a) = Y(Y(r,s\,|a),t\,|a)$.

---

[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.

Commutativity and associativity are 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 comutativity and associativity 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):

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

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

Compound Y-Principle:

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

If we know that prob(A/B) = r and prob(A/C) = s, we can also use Nonclassical Direct Inference to infer defeasibly that prob($A/B\&C$) = r. If $s \neq a$, $Y(r,s \,|a) \neq r$, so this conflicts with the conclusion that prob($A/B\&C$) = $Y(r,s \,|a)$. 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:

Y-Defeat Defeat for Nonclassical Direct Inference:

⌜$A,B,C \preccurlyeq U$ and prob(A/C) ≠ prob(A/U)⌝ is an undercutting defeater for the inference from ⌜prob(A/B) = r⌝ to ⌜prob($A/B\&C$) = r⌝ by Nonclassical Direct Inference.

It follows that we also have a defeater for the Principle of Statistical Independence:

Y-Defeat Defeat for Statistical Independence:

⌜$A,B,C \preccurlyeq U$ and prob(A/B) ≠ prob(A/U)⌝ is an undercutting defeater for the inference from ⌜prob(A/C) = r & prob(B/C) = s⌝ to ⌜prob($A\&B/C$) = $r\cdot s$⌝ 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. We are supposing that the probability of a person with his symptoms having the disease is .6. We also suppose the probability of such a person having the disease if 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 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 either the probability calculus alone or from Classical Direct Inference. 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 PROB(P) = a, PROB(P/Q) = r, and PROB(P/R) = s then PROB($P/Q\&R$) = $Y(r,s \,|a)$.

Somewhat surprisingly, when prob(C/A) ≠ prob(C/U) and prob(B/A) ≠ prob(B/U), Y-independence conflicts with ordinary independence:

Theorem 3: If B and C are Y-independent for A relative to U and prob(C/A) ≠ prob(C/U) and prob(B/A) ≠ prob(B/U) then prob(C/B) ≠ prob(C/U).

Theorem 3 seems initially surprising, because we have a defeasible assumption of independence

for B and C relative to all three of A, $U\&\sim A$, and U. Theorem 3 tells us that if A is statistically relevant to B and C then we cannot have all three. However, this situation is common. Consider the example of two sensors B and C sensing the presence of an event A. Given that one sensor fires, the probability of A is higher, but raising the probability of A will normally raise the probability of the other sensor firing. So B and C are not statistically independent. However, knowing whether an event of type A is occurring screens off the effect of the sensors on one another. For example, knowing that an event of type A occurs will raise the probability of one of the sensors firing, but knowing that the other sensor is firing will not raise that probability further. So prob($B/C\&A$) = prob(B/A) and prob($B/C\&\sim A$) = prob($B/U\&\sim A$).

The defeasible presumption of Y-independence for A is based upon a probability that takes account of more information than the probability grounding the defeasible presumption of statistical independence relative to U, so the former takes precedence. In other words, in light of theorem 3, we get a defeater for Statistical Independence whenever we have an $A \preccurlyeq U$ such that prob(A/C) ≠ prob(A/U) and prob(A/B) ≠ prob(A/U):

Y-Defeat for Statistical Independence:

⌜prob(A/C) ≠ prob(A/U) and prob(A/B) ≠ prob(A/U)⌝ is an undercutting defeater for the inference from ⌜prob(A/C) = r and prob(B/C) = s⌝ to ⌜prob($A\&B/C$) = $r \cdot s$⌝ by the Principle of Statistical Independence.

The application of the Y-function presupposes that we know the base rate prob(A/U). But suppose we do not. Then what can we conclude about prob($A/B\&C$)? It might be supposed that we can combine Indifference and the Y-Principle and conclude that prob($A/B\&C$) = Y($r,s\,|.5$). That would be interesting because, as Joseph Halpern has pointed out to me (in correspondence), this is equivalent to Dempster's "rule of composition" for belief functions (Shafer 1976).[14] However, by ignoring the base rate prob(A/U), that theory will often give intuitively incorrect results. For example, in the case of the two tests for the disease, suppose the disease is rare, with a base rate of .1, but each positive test individually confers a probability of .4 that the patient has the disease. Two positive tests should increase that probability further. Indeed, Y(.4,.4 |.1) = .8. However, Y(.4,.4 |.5) = .3, so if we ignore the base rate, two positive tests would lower the probability of having the disease instead of raising it.

The reason the Dempster-Shafer rule does not give the right answer when we are ignorant of the base rate is that, although when we are completely ignorant of the value of prob(A/U) it is reasonable to expect it to be .5, knowing the values of prob(A/B) and prob(A/C) changes the expectable value of prob(A/U). Let us define $Y_0(r,s)$ to be Y($r,s\,|a$) where a, b, and c are the solutions to the following set of three simultaneous equations (for fixed r and s):

$$2a^3 - (b + c - 2b \cdot r - 2c \cdot s - 3)a^2$$
$$+(b \cdot c + 2b \cdot r - b \cdot cr + 2c \cdot s - b \cdot c \cdot s + 2b \cdot c \cdot r \cdot s - b - c + 1)a - b \cdot c \cdot r \cdot s = 0;$$

$$\left(\frac{1-s}{1+(s-a)c}\right)^{1-s}\left(\frac{s}{a-s\cdot c}\right)^{s} = 1;$$

$$\left(\frac{1-r}{1+(r-a)b}\right)^{1-r}\left(\frac{r}{a-r\cdot b}\right)^{r} = 1.$$

Then we have the following principle:

[14] See also Bacchus et al (1996). Given very restrictive assumptions, their theory gets the special case of the Y-Principle in which a = .5, but not the general case.

Y_0-Principle:

If prob(A/B) = r and prob(A/C) = s, then the expectable value of prob(A/B&C) = $Y_0(r,s)$.

If a is the expectable value of prob(A/U) given that prob(A/B) = r and prob(A/C) = s, then $Y_0(r,s) = Y(r,s\,|a)$. However, a does not in general have a simple analytic characterization. $Y_0(r,s)$ is plotted in figure 6, and the default values of prob(A/U) are plotted in figure 6. Note how the curve for $Y_0(r,s)$ is twisted with respect to the curve for $Y(r,s\,|.5)$ (in figure 3).

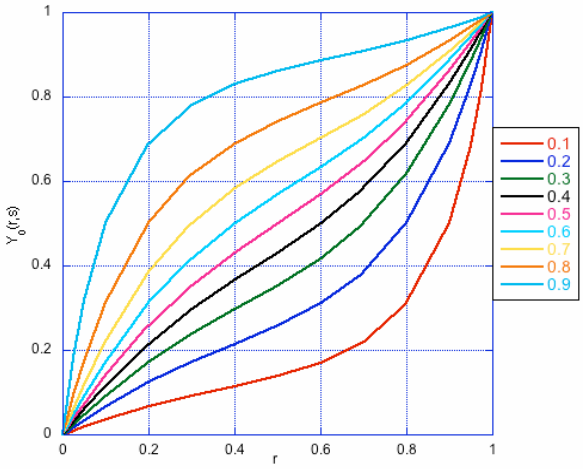

Figure 5. $Y_0(r,s)$, holding s constant (for several choices of s as indicated in the key)

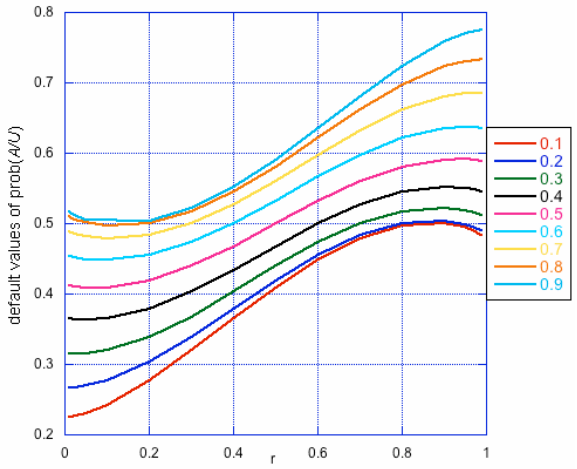

Figure 6. Default values of prob(A/U) (for several choices of s as indicated in the key)

10. Domination Defeaters for the Statistical Syllogism

The defeasible inferences licensed by our principles of probable probabilities are obtained by applying the statistical syllogism to second-order probabilities. It turns out that the principles of probable probabilities have important implications for the statistical syllogism itself. In stating the principle of the statistical syllogism in section three, the only primitive defeater that I gave was that of subproperty defeat. However, in Pollock (1990), it was argued that we must supplement subproperty defeaters with what were called "domination defeaters". Suppose that in the course of investigating a certain disease, Roderick's syndrome, it is discovered that 98% of the people having enzyme E in their blood have the disease. This becomes a powerful tool for diagnosis when used in connection with the statistical syllogism. However, the use of such information is complicated by the fact that we often have other sorts of statistical information as well. First, in statistical investigations of diseases, it is typically found that some factors are statistically irrelevant. For instance, it may be discovered that the color of one's hair is statistically irrelevant to the reliability of this diagnostic technique. Thus, for example, it is also true that 98% of all redheads having enzyme E in their blood have the disease. Second, we may discover that there are specifiable circumstances in which the diagnostic technique is unreliable. For instance, it may be found that of patients undergoing radiation therapy, only 32% of those with enzyme E in their blood have Roderick's syndrome. As we have found hair color to be irrelevant to the reliability of the diagnostic technique, we would not ordinarily go on to collect data about the effect of radiation therapy specifically on redheads. Now consider Jerome, who is redheaded, undergoing radiation therapy, and is found to have enzyme E in his blood. Should we conclude that he has Roderick's syndrome? Intuitively, we should not, but this cannot be explained directly by the statistical syllogism and subproperty defeat. We have statistical knowledge about the reference properties B = *person with enzyme E in his blood*, C = *redheaded person*, and D = *person who is undergoing radiation therapy*. Letting A be the property of having Roderick's syndrome, we know that:

(1) Bc & prob(Ax/Bx) = .98.

(2) Bc & Cc & prob(Ax/Bx&Cx) = .98.

(3) Dc & prob(Ax/Bx&Dx) = .32.

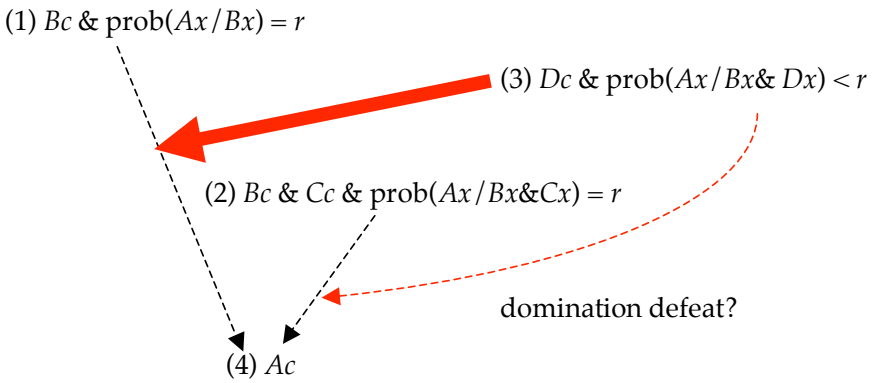

Figure 7. Domination defeat

(1), (2), and (3) are related as in figure 7, where the solid arrow indicates a defeat relation and the dashed arrows signify inference relations. By the statistical syllogism, both (1) and (2) constitute defeasible reasons for concluding that Jerome has Roderick's syndrome. (3) provides a subproperty defeater for the inference from (1), but it does not defeat the inference from (2). Thus it should be reasonable to infer that because Jerome is a redhead and most redheads with enzyme E in their blood have Roderick's syndrome, Jerome has Roderick's syndrome. Formally, the fact that Jerome is undergoing radiation therapy should not defeat the inference from (2), because that is not more specific information than the fact that Jerome has red hair. But, obviously, this is wrong. We regard Jerome's having red hair as irrelevant. The important inference is from the fact that most people with enzyme E in their blood have Roderick's syndrome to the conclusion that Jerome has Roderick's syndrome, and we regard *that* inference as undefeated. Pollock (1990) took this example to support the need for a new kind of defeater for the statistical syllogism. *Domination defeaters* were supposed to have the effect of making (3) defeat the inference from (2) to (4) by virtue of the fact that (2) defeats the inference the inference from (1) to (4) and prob(Ax/Bx) = prob(Ax/Bx&Cx):

Domination Defeat:

> If A is projectible with respect to D, then ⌜Dc & prob(A/B) = prob(A/B&C) & prob(A/B&D) < prob(A/B)⌝ is an undercutting defeater for the inference from ⌜Bc & Cc & prob(A/B&C) = r⌝ to ⌜Ac⌝ by the statistical syllogism.

What I will show here is that by appealing to the Y-Principle, we can derive domination defeaters from subproperty defeaters without making any further primitive assumptions. Applying the Y-Principle to (1), (2), and (3), we get:

Expectable Domination:

> If $B,C,D \preccurlyeq U$, prob(A/B) = r, prob(A/B&D) = $v < r$, and prob(A/U) = a, then the expectable value of prob(A/B&C&D) = $\Upsilon(v,a\,|a) = v < r$.

We can diagram the relations between these probabilities as in figure 8. The upshot is that we can infer defeasibly that prob(A/B&C&D) = prob(A/B&D), and this gives us a subproperty defeater for the inference from (2) to (4). Thus domination defeaters become derived defeaters. Note that this argument does not depend on the value of a. It works merely on the supposition that there is *some* base-rate a, and that is a necessary truth.

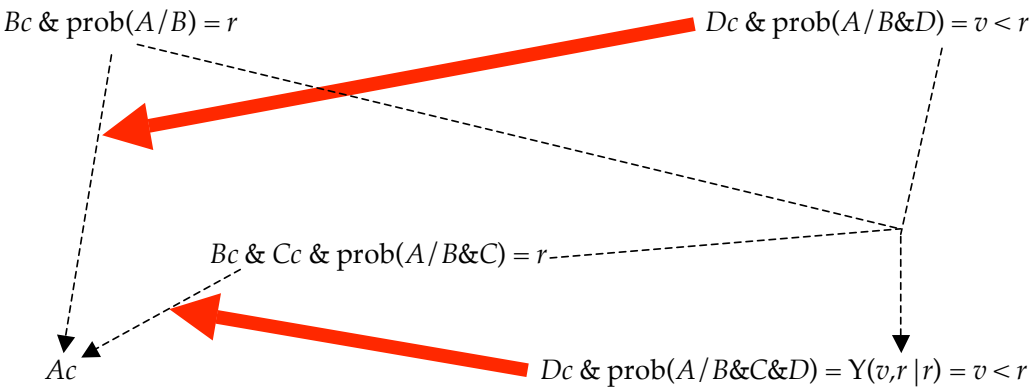

Figure 8. Reconstructing Domination Defeat

Somewhat surprisingly, domination defeat can be generalized. If we have prob(A/B&C) = $s \leq r$, we can still infer that prob(A/B&C&D) = Y(v,s |r). It is a general property of the Y-function that if $v < r$ and $s \leq r$ then Y(v,s |r) < v and Y(v,s |r) < s. Hence we get:

Generalized Domination Defeat:

If $s \leq r$ then ⌜Dc & prob(A/B) = r & prob(A/B&D) < r⌝ is an undercutting defeater for the inference from ⌜Bc & Cc & prob(A/B&C) = s⌝ to ⌜Ac⌝ by the statistical syllogism.

11. Inverse Probabilities

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 the Y-Principle 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 presents 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 prob(A/B). If we know the base rates prob(A/U) and prob(B/U), the probability calculus enables us to compute the value of the *inverse probability* prob(~B/~A&U):

Theorem 4: 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 9. Note that when prob(A/B) > prob(A/U), we can expect prob($\sim B/\sim A\&U$) to be almost as great as prob(A/B).

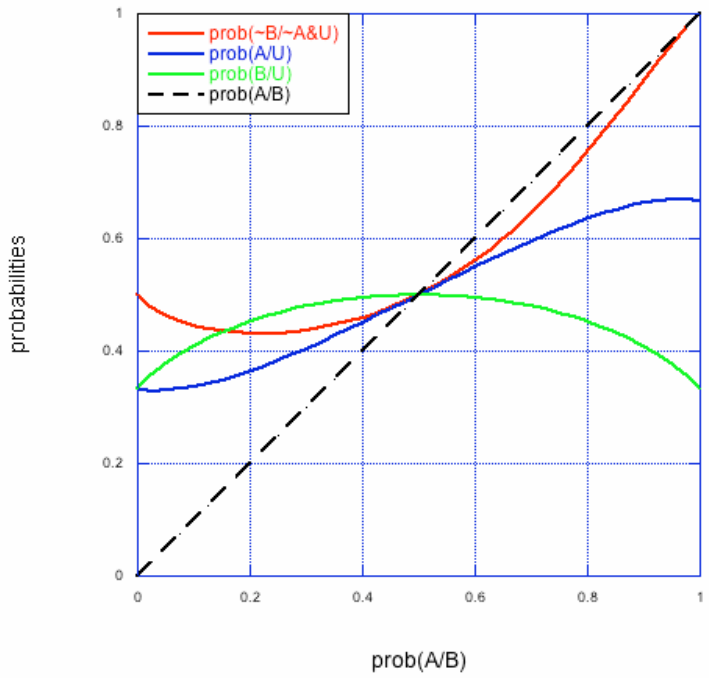

Figure 9. Expectable values of prob($\sim B/\sim A\&U$), prob(A/U), and prob(B/U), as a function of prob(A/B), when the base rates are unknown.

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 and 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 10 plots the expectable values of prob($\sim B/\sim A\&U$) (for values 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) < $\frac{1}{3-2r}$.

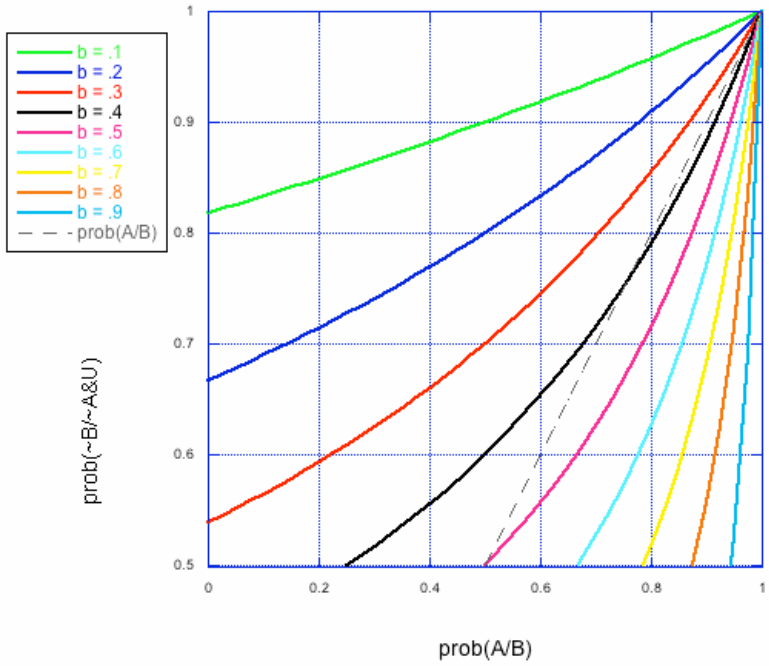

Figure 10. Expectable values of prob($\sim B/\sim A\&U$) as a function of prob(A/B), when prob(A/U) is unknown, for fixed values of prob(B/U).

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 can be computed using hill-climbing algorithms (included in the probable probabilities software). The results are plotted in figure 11. 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).

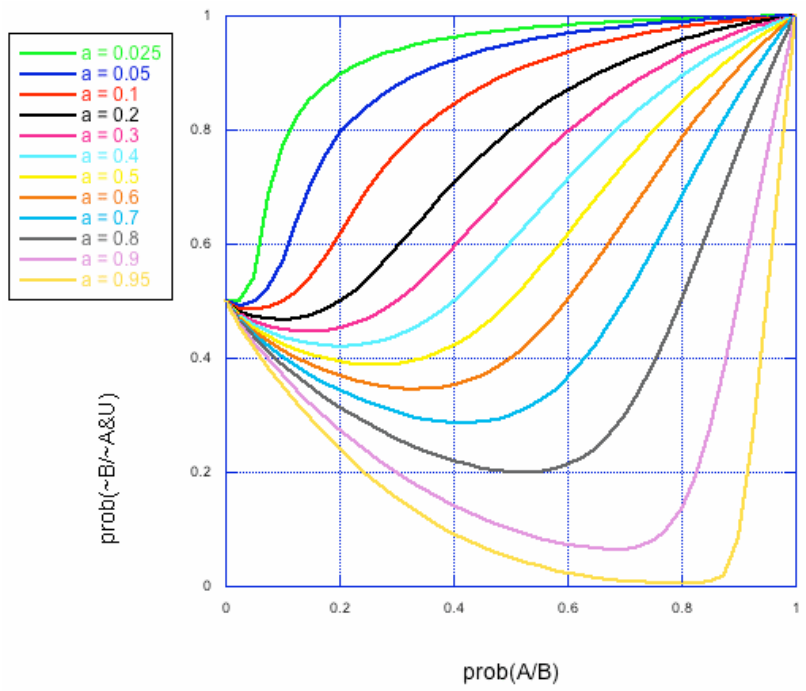

Figure 11. Expectable values of prob(~*B*/~*A*&*U*) as a function of prob(*A*/*B*), when prob(*B*/*U*) is unknown, for fixed values of prob(*A*/*U*).

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

12. 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 single defeasible inference schemes. 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 each to just 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.

13. 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 finite combinatorial mathematics provide the necessary basis 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, the Y-Principle, 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 Y-Principle. But it turns out that there is a uniform way of finding and proving these principles. This made it possible to write the probable probabilities software that analyzes the results of linear constraints and determines what the expectable values of the probabilities are. That software produces a proof of the Y-Principle in a matter of seconds.

Nomic probability and the principles of probable probability are 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 most 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.

Appendix: Proofs of Theorems

4. Limit Theorems and Probable Probabilities

Finite Indifference Principle:

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

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

We can prove the Finite Indifference Principle as follows. Theorem 1 establishes that $\frac{n}{2}$ is the size of r that maximizes Bin(n,r).

Theorem 1: If $0 < k \leq n$, $C\left(n,\frac{n}{2}\right) > C\left(n,\frac{n}{2}+k\right)$ and $C\left(n,\frac{n}{2}\right) > C\left(n,\frac{n}{2}-k\right)$.

Proof: $$\frac{C\left(n,\frac{n}{2}\right)}{C\left(n,\frac{n}{2}+k\right)} = \frac{\left(\frac{n!}{\left(\frac{n}{2}\right)!\left(\frac{n}{2}\right)!}\right)}{\left(\frac{n!}{\left(\frac{n}{2}-k\right)!\left(\frac{n}{2}+k\right)!}\right)} = \frac{\left(\frac{n}{2}-k\right)!\left(\frac{n}{2}+k\right)!}{\left(\frac{n}{2}\right)!\left(\frac{n}{2}\right)!} =$$

$$= \frac{\left(\frac{n}{2}+1\right)\cdot\ldots\cdot\left(\frac{n}{2}+k\right)}{\left(\frac{n}{2}-k+1\right)\cdot\ldots\cdot\left(\frac{n}{2}-k+k\right)} = \left(\frac{\left(\frac{n}{2}+1\right)}{\left(\frac{n}{2}+1-k\right)}\right)\cdot\ldots\cdot\left(\frac{\left(\frac{n}{2}+k\right)}{\left(\frac{n}{2}+k-k\right)}\right) > 1.$$

And $C\left(n,\frac{n}{2}+k\right) = C\left(n,\frac{n}{2}-k\right)$. ■

Theorem 2 shows that the slopes of the curves become infinitely steep as $n \to \infty$, and hence the sizes of subsets of an n-membered set cluster arbitrarily close to $\frac{n}{2}$:

Theorem 2: As $n \to \infty$, $\dfrac{C\left(n, \frac{n}{2}\right)}{C\left(n, \frac{n}{2}+kn\right)} \to \infty$ and $\dfrac{C\left(n, \frac{n}{2}\right)}{C\left(n, \frac{n}{2}-kn\right)} \to \infty$.

Proof: As in theorem 1,

$$\frac{C\left(n, \frac{n}{2}\right)}{C\left(n, \frac{n}{2}+kn\right)} = \frac{\left(\frac{n}{2}-kn\right)!\left(\frac{n}{2}+kn\right)!}{\left(\frac{n}{2}\right)!\left(\frac{n}{2}\right)!} = \frac{\left(\frac{n}{2}+1\right)}{\left(\frac{n}{2}+1-kn\right)} \cdot \ldots \cdot \frac{\left(\frac{n}{2}+kn\right)}{\left(\frac{n}{2}+kn-kn\right)}$$

$$> \left(\frac{\left(\frac{n}{2}+kn\right)}{\left(\frac{n}{2}\right)} \right)^{nk-1} = (1+2k)^{nk-1} \to \infty. \quad \blacksquare$$

Thus we have the Finite Indifference Principle .

Probable Proportions Theorem:

Let $U,X_1,\ldots,X_n$ be a set of variables ranging over sets, and consider a finitely unbounded finite set LC of linear constraints on proportions between Boolean compounds of those variables. Then for any pair of relations P,Q whose variables are a subset of $U,X_1,\ldots,X_n$ there is a unique real number r in [0,1] such that for every $\varepsilon,\delta > 0$, there is an N such that if U is finite and $\#\left\{\langle X_1,\ldots,X_n\rangle \mid LC \;\&\; X_1,\ldots,X_n \subseteq U\right\} \geq N$ then

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

Proof: Assume then that LC is finitely unbounded. For each intersection of elements of the set $\{X_1,\ldots,X_n\}$, e.g, $X \cap Y \cap Z$, let the corresponding lower-case variable xyz be $\rho(X \cap Y \cap Z/U)$. Given a set of linear constraints on these variables, the cardinality of an element X of the partition is a function $f(x)\cdot u$ of x (x may occur vacuously, in which case $f(x)$ is a constant function). I will refer to the $f(x)$'s as the *partition-coefficients*. Because the constraints are linear, for each $f(x)$ there is a positive or negative real number r such that $f(x+\varepsilon) = f(x)+r\cdot\varepsilon\cdot u$. If $r < 0$, I will say that x has a *negative occurrence*. Otherwise, x has a *positive occurrence*. It is a general characteristic of partitions that each variable has the same number k of positive and negative occurrences. Let $a_1(x),\ldots,a_k(x)$ be the partition-coefficients in which x has a positive occurrence, and let $b_1(x),\ldots,b_k(x)$ be those in which x has a negative occurrence. In most cases we will consider, $r = 1$ or $r = -1$, but not in all. The terms $r\cdot\varepsilon$ represent the amount a cell of the partition changes in size when x is incremented by ε. However, the sizes of the cells must still sum to u, so the sum or the r's must be 0. For each $i \leq k$, let r_i be the real number such that $a_i(x+\varepsilon) = a_i(x) + r_i\varepsilon$ and let s_i be the real number such that $b_i(x+\varepsilon) = b_i(x) - s_i\varepsilon$. So $r_1 + \ldots + r_k = s_1 + \ldots + s_k$. Note further that $a_1(x)\cdot u,\ldots,a_k(x)\cdot u, b_1(x)\cdot u,\ldots,b_k(x)\cdot u$ are the cardinalities of the elements of the partition, so they must be non-negative. That is, for any value ξ of x that is consistent with the probability calculus (an "allowable" value of x), $a_1(\xi),\ldots,a_k(\xi),b_1(\xi),\ldots,b_k(\xi)$ must be non-negative. It follows that if ξ is an allowable value of x, $\xi + \varepsilon$ is an allowable value only if for every $i \leq k$, $\varepsilon < \dfrac{b_i(\xi)}{s_1}$, and $\xi - \varepsilon$ is an allowable value only if for every $i \leq k$, $\varepsilon < \dfrac{a_i(\xi)}{r_1}$.

Define

$$\mathrm{Prd}(x) = (a_1(x)u)!\ldots(a_k(x)u)!(b_1(x)u)!\ldots(b_k(x)u)!$$

Our interest is in what happens as $u \to \infty$. For fixed values of the other variables, the most probable value of x occurs when $\frac{\mathrm{Prd}\left(x + 1/u\right)}{\mathrm{Prd}(x)} \to 1$ as $u \to \infty$. (This assumes that the curve has no merely local minima, but that is proven below in the course of proving the Probable Values Lemma.)

$$\frac{\mathrm{Prd}\left(x + 1/u\right)}{\mathrm{Prd}(x)} = \frac{\left(a_1\left(x + 1/u\right)u\right)!}{(a_1(x)u)!} \ldots \frac{\left(a_k\left(x + 1/u\right)u\right)!}{(a_k(x)u)!} \frac{\left(b_1\left(x + 1/u\right)u\right)!}{(b_1(x)u)!} \ldots \frac{\left(b_k\left(x + 1/u\right)u\right)!}{(b_k(x)u)!}$$

$$= \frac{\left(a_1(x)u + r_1\right)!}{(a_1(x)u)!} \ldots \frac{\left(a_k(x)u + r_k\right)!}{(a_k(x)u)!} \frac{\left(b_1(x)u - s_1\right)!}{(b_1(x)u)!} \ldots \frac{\left(b_k(x)u - s_k\right)!}{(b_k(x)u)!}.$$

For any positive or negative real number ε, $\frac{(N+\varepsilon)!}{N!} \to N^{\varepsilon}$ as $N \to \infty$, so the most probable value of x occurs when

$$\frac{(a_1(x)u)^{r_1} \ldots (a_k(x)u)^{r_k}}{(b_1(x)u)^{s_1} \ldots (b_k(x)u)^{s_k}} \to 1 \text{ as } u \to \infty.$$

$$\frac{(a_1(x)u)^{r_1} \ldots (a_k(x)u)^{r_k}}{(b_1(x)u)^{s_1} \ldots (b_k(x)u)^{s_k}} = \frac{(a_1(x))^{r_1} \ldots (a_k(x))^{r_k}}{(b_1(x))^{s_1} \ldots (b_k(x))^{s_k}} u^{r_1 + \ldots + r_k - s_1 - \ldots - s_k}.$$

As we are talking about finite sets, there is always at least one value of x that maximizes Prd(x) and hence that is a solution to this equation. It will follow from the proof of the Probable Values Lemma (below) that, in the limit, there is only one allowable value of x that is a solution. It is, in fact, the only real-valued solution within the interval [0,1]. It was noted above that $r_1 + \ldots + r_k - s_1 - \ldots - s_k = 0$, so more simply, the most probable value of x is a real-valued solution within the interval [0,1] of the following equation:

$$\frac{(a_1(x))^{r_1} \ldots (a_k(x))^{r_k}}{(b_1(x))^{s_1} \ldots (b_k(x))^{s_k}} = 1.$$

In the common case in which $r_1 = \ldots = r_k = s_1 = \ldots = s_k$, the most probable value of x occurs when

$$\frac{a_1(x) \cdot \ldots \cdot a_k(x)}{b_1(x) \cdot \ldots \cdot b_k(x)} = 1.$$

Whichever of these equations we get, I will call it the *term-characterization* of x. We find the most probable value of x by solving these equations for x.

What remains is to show, in the limit, that if ξ is the most probable value of x, then ξ has probability 1 of being the value of x. This is established by the Probable Values Lemma, stated below. To prove the Probable Values Lemma, we first need:

Partition Principle:

If $\varepsilon, x_1, \ldots, x_k, y_1, \ldots, y_k > 0$, $r_1\varepsilon < x_1, \ldots, r_k\varepsilon < x_k$, $x_1 + \ldots + x_k + y_1 + \ldots + y_k = 1$, and $r_1 + \ldots + r_k = s_1 + \ldots + s_k$ then

$$\left(1 - \frac{r_1\varepsilon}{x_1}\right)^{x_1 - r_1\varepsilon} \cdot \ldots \cdot \left(1 - \frac{r_k\varepsilon}{x_k}\right)^{x_k - r_k\varepsilon} \cdot \left(1 + \frac{s_1\varepsilon}{y_1}\right)^{y_1 + s_1\varepsilon} \cdot \ldots \cdot \left(1 + \frac{s_k\varepsilon}{y_k}\right)^{y_k + s_k\varepsilon} > 1.$$

Proof: By the inequality of the geometric and arithmetic mean, if $z_1, \ldots, z_n > 0$, $z_1 + \ldots + z_n = 1$, $a_1, \ldots, a_n > 0$,

and for some i,j, $a_i \neq a_j$, then

$$a_1^{z_1} \cdot \ldots \cdot a_n^{z_n} < z_1 a_1 + \ldots + z_n a_n .$$

We have:

$$x_1 - r_1\varepsilon + \ldots + x_k - r_k\varepsilon + y_1 + s_1\varepsilon + \ldots + y_k + s_k\varepsilon$$

$$= x_1 + \ldots + x_k + y_1 + \ldots + y_k + \varepsilon(s_1 + \ldots + s_k - r_1 - \ldots - r_k) = 1$$

and $\left(\frac{x_1}{x_1 - r_1\varepsilon}\right) > 1 > \left(\frac{y_1}{y_1 + s_1\varepsilon}\right)$, so

$$\left(\frac{x_1}{x_1 - r_1\varepsilon}\right)^{x_1 - r_1\varepsilon} \cdot \ldots \cdot \left(\frac{x_k}{x_k - r_k\varepsilon}\right)^{x_k - r_k\varepsilon} \cdot \left(\frac{y_1}{y_1 + s_1\varepsilon}\right)^{y_1 + s_1\varepsilon} \cdot \ldots \cdot \left(\frac{y_k}{y_k + s_k\varepsilon}\right)^{y_k + s_k\varepsilon}$$

$$< (x_1 - r_1\varepsilon)\left(\frac{x_1}{x_1 - r_1\varepsilon}\right) + \ldots + (x_k - r_k\varepsilon)\left(\frac{x_k}{x_k - r_k\varepsilon}\right) + (y_1 + s_1\varepsilon)\left(\frac{y_1}{y_1 + s_1\varepsilon}\right) + \ldots + (y_k + s_k\varepsilon)\left(\frac{y_k}{y_k + s_k\varepsilon}\right)$$

$$= 1.$$

Equivalently,

$$\left(1 - \frac{r_1\varepsilon}{x_1}\right)^{x_1 - r_1\varepsilon} \cdot \ldots \cdot \left(1 - \frac{r_k\varepsilon}{x_k}\right)^{x_k - r_k\varepsilon} \cdot \left(1 + \frac{s_1\varepsilon}{y_1}\right)^{y_1 + s_1\varepsilon} \cdot \ldots \cdot \left(1 + \frac{s_k\varepsilon}{y_k}\right)^{y_k + s_k\varepsilon} > 1. \quad \blacksquare$$

Now we can prove:

Probable Values Lemma:

If LC is an infinitely unbounded set of linear constraints, $a_1(x),\ldots,a_k(x),b_1(x),\ldots,b_k(x)$ are the resulting positive and negative partition coefficients, and

$$\frac{(a_1(\xi))^{r_1} \ldots (a_k(\xi))^{r_k}}{(b_1(\xi))^{s_1} \ldots (b_k(\xi))^{s_k}} = 1$$

then for every $\varepsilon, \delta > 0 > 0$, the probability that ξ is within δ of the actual the value of x is greater than 1–ε.

Proof: Where

$$\mathrm{Prd}(x) = (a_1(x)u)! \ldots (a_k(x)u)!(b_1(x)u)! \ldots (b_k(x)u)!$$

it suffices to show that when ξ is the most probable value of x, then (1) if for every $i \leq k$, $\varepsilon < \frac{b_i(\xi)}{s_1}$, then $\frac{\mathrm{Prd}(\xi + \varepsilon)}{\mathrm{Prd}(\xi)} \to \infty$ as $u \to \infty$, and (2) if for every $i \leq k$, $\varepsilon < \frac{a_i(\xi)}{r_1}$, then $\frac{\mathrm{Prd}(\xi - \varepsilon)}{\mathrm{Prd}(\xi)} \to \infty$ as $u \to \infty$. I will just prove the former, as the latter is analogous.

$$\frac{\mathrm{Prd}(\xi + \varepsilon)}{\mathrm{Prd}(\xi)} = \frac{\left(a_1(\xi + \varepsilon)u\right)!}{(a_1(\xi))!} \ldots \frac{\left(a_k(\xi + \varepsilon)u\right)!}{(a_k(\xi))!} \frac{\left(b_1(\xi + \varepsilon)u\right)!}{(b_1(\xi))!} \ldots \frac{\left(b_k(\xi + \varepsilon)u\right)!}{(b_k(\xi))!}$$

$$= \frac{\big((a_1(\xi)+\varepsilon r_1)u\big)!}{(a_1(\xi)u)!}\ldots\frac{\big((a_k(\xi)+\varepsilon r_k)u\big)!}{(a_k(\xi)u)!}\frac{\big((b_1(\xi)-\varepsilon s_1)u\big)!}{(b_1(\xi)u)!}\ldots\frac{\big((b_k(\xi)-\varepsilon s_k)u\big)!}{(b_k(\xi)u)!}.$$

By the Stirling approximation, $\sqrt{2\pi n}\left(\frac{n}{e}\right)^n < n! < \sqrt{2\pi n}\left(\frac{n}{e}\right)^n\left(1+\frac{1}{12n-1}\right)$. Thus as $u\to\infty$,

$$\frac{\big((a(\xi)+\varepsilon r)u\big)!}{(a(\xi)u)!} \to \frac{\big((a(\xi)+r\varepsilon)u\big)^{(a(\xi)+r\varepsilon)u+1/2}}{(a(\xi)u)^{a(\xi)u+1/2}}e^{-r\varepsilon u} = \frac{\left(a(\xi)u\left(1+\frac{r\varepsilon}{a(\xi)}\right)\right)^{(a(\xi)+r\varepsilon)u+1/2}}{(a(\xi)u)^{a(\xi)u+1/2}}e^{-r\varepsilon u}$$

$$= \left(\frac{a(\xi)u}{e}\right)^{r\varepsilon u}\left(1+\frac{r\varepsilon}{a(\xi)}\right)^{(a(\xi)+r\varepsilon)u+1/2}.$$

Similarly,

$$\frac{\big((b(\xi)-\varepsilon s)u\big)!}{(b(\xi)u)!} \to \frac{\left(1-\frac{s\varepsilon}{b(\xi)}\right)^{(b(\xi)-s\varepsilon)u+1/2}}{\left(\frac{b(\xi)u}{e}\right)^{s\varepsilon u}}.$$

Therefore,

$$\frac{\mathrm{Prd}(\xi+\varepsilon)}{\mathrm{Prd}(\xi)} \to \left(\begin{array}{l}\left(\frac{a_1(\xi)u}{e}\right)^{r_1\varepsilon u}\left(1+\frac{r_1\varepsilon}{a_1(\xi)}\right)^{(a_1(\xi)+r_1\varepsilon)u+1/2}\cdot\ldots\cdot\left(\frac{a_k(\xi)u}{e}\right)^{r_k\varepsilon u}\left(1+\frac{r_k\varepsilon}{a_k(\xi)}\right)^{(a_k(\xi)+r_k\varepsilon)u+1/2}\\ \qquad\cdot\frac{\left(1-\frac{s_1\varepsilon}{b_1(\xi)}\right)^{(b_1(\xi)-s_1\varepsilon)u+1/2}}{\left(\frac{b_1(\xi)u}{e}\right)^{s_1\varepsilon u}}\cdot\ldots\cdot\frac{\left(1-\frac{s_k\varepsilon}{b_k(\xi)}\right)^{(b_k(\xi)-s_k\varepsilon)u+1/2}}{\left(\frac{b_k(\xi)u}{e}\right)^{s_k\varepsilon u}}\end{array}\right)$$

$$= \left(\frac{\left(\frac{a_1(\xi)u}{e}\right)^{r_1}\cdot\ldots\cdot\left(\frac{a_k(\xi)u}{e}\right)^{r_k}}{\left(\frac{b_1(\xi)u}{e}\right)^{s_1}\cdot\ldots\cdot\left(\frac{b_k(\xi)u}{e}\right)^{s_k}}\right)^{\varepsilon u}\cdot\left(\begin{array}{l}\left(1+\frac{r_1\varepsilon}{a_1(\xi)}\right)^{(a_1(\xi)+r_1\varepsilon)u+1/2}\cdot\ldots\cdot\left(1+\frac{r_k\varepsilon}{a_k(\xi)}\right)^{(a_k(\xi)+r_k\varepsilon)u+1/2}\\ \cdot\left(1-\frac{s_1\varepsilon}{b_1(\xi)}\right)^{(b_1(\xi)-s_1\varepsilon)u+1/2}\cdot\ldots\cdot\left(1-\frac{s_k\varepsilon}{b_k(\xi)}\right)^{(b_k(\xi)-s_k\varepsilon)u+1/2}\end{array}\right)$$

$$= \left(\frac{\big(a_1(\xi)\big)^{r_1}\cdot\ldots\cdot\big(a_k(\xi)\big)^{r_k}}{\big(b_1(\xi)\big)^{s_1}\cdot\ldots\cdot\big(b_k(\xi)\big)^{s_k}}\right)^{\varepsilon u}\left(\frac{u}{e}\right)^{r_1+\ldots+r_k-s_1-\ldots-s_k}\cdot\left(\begin{array}{l}\left(1+\frac{r_1\varepsilon}{a_1(\xi)}\right)^{(a_1(\xi)+r_1\varepsilon)u+1/2}\cdot\ldots\cdot\left(1+\frac{r_k\varepsilon}{a_k(\xi)}\right)^{(a_k(\xi)+r_k\varepsilon)u+1/2}\\ \cdot\left(1-\frac{s_1\varepsilon}{b_1(\xi)}\right)^{(b_1(\xi)-s_1\varepsilon)u+1/2}\cdot\ldots\cdot\left(1-\frac{s_k\varepsilon}{b_k(\xi)}\right)^{(b_k(\xi)-s_k\varepsilon)u+1/2}\end{array}\right).$$

$r_1+\ldots+r_k-s_1-\ldots-s_k=0$, so $\left(\frac{u}{e}\right)^{r_1+\ldots+r_k-s_1-\ldots-s_k}=1$, and $\frac{\big(a_1(\xi)\big)^{r_1}\cdot\ldots\cdot\big(a_k(\xi)\big)^{r_k}}{\big(b_1(\xi)\big)^{s_1}\cdot\ldots\cdot\big(b_k(\xi)\big)^{s_k}}=1.$

Hence,

$$\frac{\mathrm{Prd}(\xi+\varepsilon)}{\mathrm{Prd}(\xi)} \to \begin{pmatrix} \left(1+\frac{r_1\varepsilon}{a_1(\xi)}\right)^{(a_1(\xi)+r_1\varepsilon)u+1/2} \cdot \ldots \cdot \left(1+\frac{r_k\varepsilon}{a_k(\xi)}\right)^{(a_k(\xi)+r_k\varepsilon)u+1/2} \\ \cdot \left(1-\frac{s_1\varepsilon}{b_1(\xi)}\right)^{(b_1(\xi)-s_1\varepsilon)u+1/2} \cdot \ldots \cdot \left(1-\frac{s_k\varepsilon}{b_k(\xi)}\right)^{(b_k(\xi)-s_k\varepsilon)u+1/2} \end{pmatrix}$$

$$= \left(1+\frac{r_1\varepsilon}{a_1(\xi)}\right)^{(a_1(\xi)+r_1\varepsilon)u} \cdot \ldots \cdot \left(1+\frac{r_k\varepsilon}{a_k(\xi)}\right)^{(a_k(\xi)+r_k\varepsilon)u} \cdot \left(1-\frac{s_1\varepsilon}{b_1(\xi)}\right)^{(b_1(\xi)-s_1\varepsilon)u} \cdot \ldots \cdot \left(1-\frac{s_k\varepsilon}{b_k(\xi)}\right)^{(b_k(\xi)-s_k\varepsilon)u}$$

$$= \left(\left(1+\frac{r_1\varepsilon}{a_1(\xi)}\right)^{(a_1(\xi)+r_1\varepsilon)} \cdot \ldots \cdot \left(1+\frac{r_k\varepsilon}{a_k(\xi)}\right)^{(a_k(\xi)+r_k\varepsilon)} \cdot \left(1-\frac{s_1\varepsilon}{b_1(\xi)}\right)^{(b_1(\xi)-s_1\varepsilon)} \cdot \ldots \cdot \left(1-\frac{s_k\varepsilon}{b_k(\xi)}\right)^{(b_k(\xi)-s_k\varepsilon)}\right)^u.$$

I will call the latter the *slope function for x*. By the Partition Principle:

$$\left(\left(1+\frac{r_1\varepsilon}{a_1(\xi)}\right)^{(a_1(\xi)+r_1\varepsilon)} \cdot \ldots \cdot \left(1+\frac{r_k\varepsilon}{a_k(\xi)}\right)^{(a_k(\xi)+r_k\varepsilon)} \cdot \left(1-\frac{s_1\varepsilon}{b_1(\xi)}\right)^{(b_1(\xi)-s_1\varepsilon)} \cdot \ldots \cdot \left(1-\frac{s_k\varepsilon}{b_k(\xi)}\right)^{(b_k(\xi)-s_k\varepsilon)}\right)^u > 1.$$

Hence $\frac{\mathrm{Prd}(\xi+\varepsilon)}{\mathrm{Prd}(\xi)} \to \infty$ as $u \to \infty$. So as $u \to \infty$, the probability that $x \underset{\delta}{\approx} \xi \to 1$. ■

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

Proof**:** Suppose $\rho(A/B) = p$, where B is infinite. By the finite-set principle:

$$\rho_X\left(\rho(A/X) = r \ / \ X \subseteq B \ \&\ \#X = N\right) =$$

$$\rho_{x_1,\ldots,x_N}\left(\rho(A/\{x_1,\ldots,x_N\}) = r \ / \ x_1,\ldots,x_N \text{ are pairwise distinct } \&\ x_1,\ldots,x_N \in B\right).$$

"$\rho(A/\{x_1,\ldots,x_N\}) = r$" is equivalent to the disjunction of $\frac{N!}{(rN)!((1-r)N)!}$ pairwise logically incompatible disjuncts of the form "$y_1,\ldots,y_{rN} \in A \ \&\ z_1,\ldots,z_{(1-r)N} \notin A$" where $\{y_1,\ldots,y_{rN},z_1,\ldots,z_{(1-r)N}\} = \{x_1,\ldots,x_N\}$. By the crossproduct principle,

$$\rho_{x_1,\ldots,x_{rN},y_1,\ldots,y_{(1-r)N}}\begin{pmatrix} x_1,\ldots,x_{rN} \in A \ \&\ y_1,\ldots,y_{(1-r)N} \notin A \ / \\ x_1,\ldots,x_{rN},y_1,\ldots,y_{(1-r)N} \in B \ \&\ x_1,\ldots,x_{rN},y_1,\ldots,y_{(1-r)N} \text{ are pairwise distinct} \end{pmatrix} =$$

$$= p^{rN}(1-p)^{(1-r)N}.$$

(For instance, $\rho_{x,y,z}\left(Ax \ \&\ Ay \ \&\ \sim Az \ / \ Bx \ \&\ By \ \&\ Bz\right) = \rho\left(A\times A\times(B-A), B\times B\times B\right) = p\cdot p\cdot(1-p)$.)
Hence by finite additivity:

$$\rho_X\left(\rho(A/X) = r \ / \ X \subseteq B \ \&\ \#X = N\right) = \frac{N!\,p^{rN}(1-p)^{(1-r)N}}{(rN)!((1-r)N)!}.$$

This is just the formula for the binomial distribution. It follows by the familiar mathematics of the binomial distribution according to which, like $C(n,r)$, it becomes "needle-like" in the limit, that for every $\varepsilon,\delta > 0$, there is an N such that

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

Limit Principle for Proportions:

Consider a finitely unbounded finite set LC of linear constraints on proportions between Boolean compounds of a list of variables $U,X_1,\ldots,X_n$. Let r be limit solution for $\rho(P/Q)$ given LC. Then for any infinite set U, for every $\delta > 0$:

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

Proof: Consider a finitely unbounded finite set LC of linear constraints, and let r be the limit solution for $\rho(P/Q)$ given LC. Thus for every $\varepsilon,\delta > 0$, there is an N such that if U^* is finite and $\left\{\langle X_1,\ldots,X_n\rangle \mid LC \;\&\; X_1,\ldots,X_n \subseteq U^*\right\} \geq N$, then

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

It follows by the projection principle that for every $\varepsilon,\delta > 0$, there is an N such that

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

Suppose that for some $\delta > 0$ and infinite U:

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

As LC is finitely unbounded, it follows that $\left\{\langle X_1,\ldots,X_n\rangle / \; LC \;\&\; X_1,\ldots,X_n \subseteq U\right\}$ is infinite. Hence by the Law of Large Numbers for Proportions, for every $\varepsilon > 0$, there is an N such that

(1) $$\rho_U\left(\rho_{X_1,\ldots,X_n}\left(\rho(P,Q) \underset{\delta}{\approx} r \;/\; LC \;\&\; X_1,\ldots,X_n \subseteq U^*\right) \underset{\delta}{\approx} s \;/\; U^* \subseteq U \;\&\; U^* \text{ is finite } \&\; \#U^* \geq N\right) \geq 1-\varepsilon$$

But we know that there is an N such that for every finite U^* such that $U^* \subseteq U$ and $\#U^* \geq N$,

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

So by the Universality Principle we get:

(3) $$\rho_{U^*}\left(\rho_{X_1,\ldots,X_n}\left(\rho(P,Q) \underset{\delta}{\approx} r \;/\; LC \;\&\; X_1,\ldots,X_n \subseteq U^*\right) \geq 1-\varepsilon \;/\; U^* \subseteq U \;\&\; U^* \text{ is finite } \&\; \#U^* \geq N\right) = 1$$

For every $\varepsilon,\delta > 0$ there is an N such that (1) and (3) hold. It follows that $s = 1$. $\blacksquare$

Probable Probabilities Theorem:

Consider a finitely unbounded finite set LC of linear constraints on proportions between Boolean compounds of a list of variables $U,X_1,\ldots,X_n$. Let r be limit solution for $\rho(P/Q)$ given LC. Then for any nomically possible property U, for every $\delta > 0$,

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

Proof: Assume the antecedent. $\text{prob}(P/Q) = \rho(\mathfrak{P},\mathfrak{Q})$, so the Limit Principle for Proportions immediately implies:

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

The crossproduct principle tells us:

$$\rho\left(A\times B/C\times D\right) = \rho(A/C)\cdot\rho(B/D).$$

The properties expressed by $\ulcorner LC \;\&\; X_1,\ldots,X_n \preccurlyeq U \urcorner$ and $\ulcorner \text{prob}(P/Q) \underset{\delta}{\approx} r \urcorner$ have the same instances in all physically possible worlds, so where $\mathfrak{W}$ is the set of all physically possible worlds,

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

$$= \rho\left(\left\{\langle w,X_1,\ldots,X_n\rangle \mid w\in\mathfrak{W} \;\&\; \text{prob}(P/Q) \underset{\delta}{\approx} r\right\},\left\{\langle w,X_1,\ldots,X_n\rangle \mid w\in\mathfrak{W} \;\&\; LC \;\&\; X_1,\ldots,X_n \preccurlyeq U\right\}\right)$$

$$= \rho\left(\mathfrak{W}\times\left\{X_1,\ldots,X_n \mid \text{prob}(P/Q) \underset{\delta}{\approx} r\right\},\mathfrak{W}\times\left\{X_1,\ldots,X_n \mid LC \;\&\; X_1,\ldots,X_n \preccurlyeq U\right\}\right)$$

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

5. Statistical Independence

Finite Independence Principle:

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

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

Proof: The limit value for $\rho(X\cap Y,Z)$ given that $X,Y,Z \subseteq U$ & $\rho(X,Z) = r$, $\rho(Y,Z) = s$, $\rho(X,U) = a$, $\rho(Y,U) = b$, and $\rho(Z,U) = c$ can be computed by executing the following instruction in the probable probabilities software:

```
(analyze-probability-structure
        :subsets '(A B C)
        :constants '(a b c r s)
        :probability-constraints '((prob(A / C) = r)
                                   (prob(B / C) = s))
        :probability-queries '(prob((A & B) / C))
        :display-details t
        :display-infix t)
```

"prob" and "&" are used in place of "ρ" and "$\cap$" because non-ASCII symbols are not supported in

most computer languages. However, in light of the Probable Probabilities Theorem, the result for the limit value implies the analogous principle for probabilities.

The software produces the following:

```
(
==========================================================
Dividing U into 3 subsets A,B,C whose cardinalities relative to U are a, b, c,
if the following constraints are satisfied:
    prob(A / C) = r
    prob(B / C) = s
and hence
    bc = (s * c)
    ac = (r * c)
and the values of a, b, c, r, s are held constant,
then the term-set consisting of the cardinalities of the partition of U is:
   {
    ((ab - abc) * u)
    (abc * u)
    (((a + abc) - (ab + (r * c))) * u)
    (((r * c) - abc) * u)
    (((b + abc) - (ab + (s * c))) * u)
    (((s * c) - abc) * u)
    (((ab + 1 + (r * c) + (s * c)) - (a + b + abc + c)) * u)
    (((c + abc) - ((r * c) + (s * c))) * u)
   }
```

For computing the most probable value of *abc*, we need only consider the members of the term-set that contain *abc*:

```
The subset of terms in the term-set that contain abc is:
        {
        ((ab - abc) * u)
        (abc * u)
        (((a + abc) - (ab + (r * c))) * u)
        (((r * c) - abc) * u)
        (((b + abc) - (ab + (s * c))) * u)
        (((s * c) - abc) * u)
        (((ab + 1 + (r * c) + (s * c)) - (a + b + abc + c)) * u)
        (((c + abc) - ((r * c) + (s * c))) * u)
        }
```

As shown in the Probable Proportions theorem, the most probable values of *ab* and *abc* are those that minimize the product of the factorials of these members of the term-set, and for any positive or negative real number ε, $\frac{(N+\varepsilon)!}{N!} \to N^{\varepsilon}$ as $N \to \infty$. So

```
The expectable-value of abc is then the real-valued solution to the following equation:

1 = (((ab - abc) ^ ((ab - (abc + 1)) - (ab - abc)))
    *
    (abc ^ ((abc + 1) - abc))
    *
    (((a + abc) - (ab + (r * c))) ^ (((a + (abc + 1)) - (ab + (r * c))) - ((a + abc) - (ab + (r * c)))))
    *
    (((r * c) - abc) ^ (((r * c) - (abc + 1)) - ((r * c) - abc)))
    *
    (((b + abc) - (ab + (s * c))) ^ (((b + (abc + 1)) - (ab + (s * c))) - ((b + abc) - (ab + (s * c)))))
    *
    (((s * c) - abc) ^ (((s * c) - (abc + 1)) - ((s * c) - abc)))
    *
    (((ab + 1 + (r * c) + (s * c)) - (a + b + abc + c)) ^ (((ab + 1 + (r * c) + (s * c)) - (a + b + (abc + 1) + c)) - ((ab + 1 + (r * c)
                                                                                                + (s * c)) - (a + b + abc + c))))
    *
    (((c + abc) - ((r * c) + (s * c))) ^ (((c + (abc + 1)) - ((r * c) + (s * c))) - ((c + abc) - ((r * c) + (s * c))))))
```

```
  = (((ab - abc) ^ (- 1)) * (abc ^ 1) * (((a + abc) - (ab + (r * c))) ^ 1) * (((r * c) - abc) ^ (- 1)) * (((b + abc) - (ab + (s * c))) ^ 1)
    * (((s * c) - abc) ^ (- 1)) * (((ab + 1 + (r * c) + (s * c)) - (a + b + abc + c)) ^ (- 1)) * (((c + abc) - ((r * c) + (s * c))) ^ 1))
  = ((1 / (ab - abc)) * abc * ((abc + a) - (ab + (r * c))) * (1 / ((r * c) - abc)) * ((abc + b) - (ab + (s * c))) * (1 / ((s * c) - abc)) *
    (1 / (((s * c) + (r * c) + 1 + ab) - (a + b + abc + c))) * ((abc + c) - ((r * c) + (s * c))))
  = (abc * ((abc + a) - (ab + (r * c))) * ((abc + b) - (ab + (s * c))) * ((abc + c) - ((r * c) + (s * c))) * (1 / (((s * c) + (r * c) + 1 +
    ab) - (a + b + abc + c))) * (1 / ((s * c) - abc)) * (1 / ((r * c) - abc)) * (1 / (ab - abc)))
  = ((((c + abc) - ((r * c) + (s * c))) * ((b + abc) - (ab + (s * c))) * ((a + abc) - (ab + (r * c))) * abc) / ((ab - abc) * ((r * c) –
    abc) * ((s * c) - abc) * ((ab + 1 + (r * c) + (s * c)) - (a + b + abc + c))))

The subset of terms in the term-set that contain ab is:
      {
      ((ab - abc) * u)
      (((a + abc) - (ab + (r * c))) * u)
      (((b + abc) - (ab + (s * c))) * u)
      (((ab + 1 + (r * c) + (s * c)) - (a + b + abc + c)) * u)
      }
The expectable-value of ab is then the real-valued solution to the following equation:
1 = (((ab - abc) ^ (((ab + 1) - abc) - (ab - abc)))
    * (((a + abc) - (ab + (r * c))) ^ (((a + abc) - ((ab + 1) + (r * c))) - ((a + abc) - (ab + (r * c)))))
    * (((b + abc) - (ab + (s * c))) ^ (((b + abc) - ((ab + 1) + (s * c))) - ((b + abc) - (ab + (s * c)))))
    * (((ab + 1 + (r * c) + (s * c)) - (a + b + abc + c)) ^ ((((ab + 1) + 1 + (r * c) + (s * c)) - (a + b + abc + c))
                                                            - ((ab + 1 + (r * c) + (s * c)) - (a + b + abc + c)))))
  = (((ab - abc) ^ 1) * (((a + abc) - (ab + (r * c))) ^ (- 1))
   * (((b + abc) - (ab + (s * c))) ^ (- 1))
   * (((ab + 1 + (r * c) + (s * c)) - (a + b + abc + c)) ^ 1))
  = ((ab - abc) * (1 / ((abc + a) - (ab + (r * c))))
    * (1 / ((abc + b) - (ab + (s * c)))) * (((s * c) + (r * c) + 1 + ab) - (a + b + abc + c)))
  = ((ab - abc) * (((s * c) + (r * c) + 1 + ab) - (a + b + abc + c))
    * (1 / ((abc + b) - (ab + (s * c)))) * (1 / ((abc + a) - (ab + (r * c)))))
  = ((((ab + 1 + (r * c) + (s * c)) - (a + b + abc + c)) * (ab - abc)) / (((a + abc) - (ab + (r * c))) * ((b + abc) - (ab + (s * c)))))
The preceding term-characterization for ab simplifies to:
. . ((((c * ab) + (u * abc) + (a * b) + (r * s * (c ^ 2))) - ((c * abc) + (u * ab) + (a * s * c) + (r * c * b))) = 0)
Solving for ab:
. . ab = ((((u * abc) + (a * b) + (r * s * (c ^ 2))) - ((r * c * b) + (a * s * c) + (c * abc))) / (u - c))
Substituting the preceding definition for ab into the previous term-characterizations
produces the new term-characterizations:
. . . . . abc: 1 = ((((c + abc) - ((r * c) + (s * c)))
                      * ((b + abc) - (((((u * abc) + (a * b) + (r * s * (c ^ 2)))
                        - ((r * c * b) + (a * s * c) + (c * abc))) / (u - c)) + (s * c)))
                      * ((a + abc) - (((((u * abc) + (a * b) + (r * s * (c ^ 2)))
                        - ((r * c * b) + (a * s * c) + (c * abc))) / (u - c)) + (r * c))) * abc)
                /
                ((((((u * abc) + (a * b) + (r * s * (c ^ 2))) - ((r * c * b) + (a * s * c) + (c * abc))) / (u - c)) - abc)
                    * ((r * c) - abc) * ((s * c) - abc) * ((((((u * abc) + (a * b) + (r * s * (c ^ 2)))
                      - ((r * c * b) + (a * s * c) + (c * abc))) / (u - c)) + u + (r * c) + (s * c)) - (a + b + abc + c))))
These term-characterizations simplify to yield the following term-characterizations:
. . . . . abc: 1 = ((abc * ((c + abc) - ((r * c) + (s * c)))) / (((r * c) - abc) * ((s * c) - abc)))
. The preceding term-characterization for abc simplifies to:
. . . (((r * s * (c ^ 2)) - (abc * c)) = 0)
. Solving for abc:
. . . abc = (r * s * (c ^ 1))

========================== EXPAND-DEFS ============================
Thus far we have found the following definitions:
abc = (r * s * (c ^ 1))
ab = ((((u * abc) + (a * b) + (r * s * (c ^ 2))) - ((r * c * b) + (a * s * c) + (c * abc))) / (u - c))

Substituting the definition for abc into the definition for ab and simplifying, produces:
ab = ((((a * b) + (u * r * s * c)) - ((r * c * b) + (a * s * c))) / (u - c))

Grounded definitions of the expectable values were found for all the variables.
---------------------------------------------------------------------------------------
The following definitions of expectable values were found that appeal only to the constants:
----------
abc = (r * s * c)
----------
```

ab = ((((r * s * c) + (a * b)) - ((r * c * b) + (a * s * c))) / (1 - c))
==========
Reconstruing a, b, c, etc., as probabilities relative to U rather than as cardinalities, the
following characterizations were found for the expectable values of the probabilities wanted:
---
prob((A & B) / C) = (r * s)

==========
) ■

8. Classical Direct Inference

Representation Theorem for Singular Probabilities:

$$\text{PROB}(Fa) = \text{prob}(Fx / x = a \;\&\; \mathbf{K}).$$

Proof:

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

$$= \rho\left(\{\langle w,x\rangle \mid w \in \mathfrak{W} \;\&\; (x = a \;\&\; Fx \;\&\; \mathbf{K}) \text{ at } w\}, \{\langle w,x\rangle \mid w \in \mathfrak{W} \;\&\; (x = a \;\&\; \mathbf{K}) \text{ at } w\}\right)$$

$$= \rho\left(\{\langle w,x\rangle \mid w \in \mathfrak{W} \;\&\; x = a \;\&\; (Fx \;\&\; \mathbf{K}) \text{ at } w\}, \{\langle w,x\rangle \mid w \in \mathfrak{W} \;\&\; x = a \;\&\; \mathbf{K} \text{ at } w\}\right)$$

$$= \rho\left(\{\langle w,a\rangle \mid w \in \mathfrak{W} \;\&\; (Fa \;\&\; \mathbf{K}) \text{ at } w\}, \{\langle w,a\rangle \mid w \in \mathfrak{W} \;\&\; \mathbf{K} \text{ at } w\}\right)$$

$$= \rho\left(\{w \mid w \in \mathfrak{W} \;\&\; (Fa \;\&\; \mathbf{K}) \text{ at } w\} \times \{a\}, \{w \mid w \in \mathfrak{W} \;\&\; \mathbf{K} \text{ at } w\} \times \{a\}\right)$$

$$= \rho\left(\{w \mid w \in \mathfrak{W} \;\&\; (Fa \;\&\; \mathbf{K}) \text{ at } w\}, \{w \mid w \in \mathfrak{W} \;\&\; \mathbf{K} \text{ at } w\}\right) \cdot \rho(\{a\},\{a\})$$

$$= \rho\left(\{w \mid w \in \mathfrak{W} \;\&\; (Fa \;\&\; \mathbf{K}) \text{ at } w\}, \{w \mid w \in \mathfrak{W} \;\&\; \mathbf{K} \text{ at } w\}\right)$$

$$= \text{PROB}(Fa) \;.\; \blacksquare$$

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
1947 *Meaning and Necessity*. Chicago: University of Chicago Press.
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.
1977, "Randomness and the right reference class", *Journal of Philosophy* **74**, 791-797
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.
1983a "Epistemology and Probability", *Synthese*, **55**, 231-252.
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.
Ramsey, Frank
1926 "Truth and probability", in *The Foundations of Mathematics*, ed. R. B. Braithwaite, Paterson, NJ: Littlefield, Adams.
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.