Probable Probabilities[1] (with proofs)

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

Abstract

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

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

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

1. The Problem of Sparse Probability Knowledge

The uninitiated often suppose that if we know a few basic probabilities, we can compute the values of many others just by applying the probability calculus. Thus it might be supposed that familiar sorts of statistical inference provide us with our basic knowledge of probabilities, and then appeal to the probability calculus enables us to compute other previously unknown probabilities. The picture is of a kind of foundations theory of the epistemology of probability, with the probability calculus providing the inference engine that enables us to get beyond whatever probabilities are discovered by direct statistical investigation.

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

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

PROB($P \vee Q$) is that $.7 \leq$ PROB($P \vee Q$) ≤ 1.0. In general, the probability calculus imposes constraints on the probabilities of logical compounds, but it falls far short of enabling us to compute unique values. I will call this *the problem of sparse probability knowledge*.

I will argue that what we need are principles for reasoning about probabilities defeasibly. These will enable us to form reasonable expectations about the values of probabilities we cannot compute deductively, but as with any defeasible reasoning, there is always the logical possibility that new information will force us to change our minds and give up the initially reasonable expectations. For example, in applying probabilities to concrete problems, probability practitioners commonly adopt undefended assumptions of statistical independence. The independence assumption is a defeasible assumption, because obviously we can discover that conditions we thought were independent are unexpectedly correlated. The probability calculus can give us only necessary truths about probabilities, so the justification of such a defeasible assumption must have some other source.

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

2. Two Kinds of Probability

My solution to the problem of sparse probability knowledge requires that we start with objective probabilities. What I will call *generic probabilities* 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. I will refer to these probabilities as *singular probabilities*.

Most objective approaches to probability tie probabilities to relative frequencies in some essential way, and the resulting probabilities have the same logical form as the relative frequencies. That is, they are generic probabilities. The simplest theories identify generic probabilities with relative frequencies (Russell 1948; Braithwaite 1953; Kyburg 1961, 1974; Sklar 1970, 1973).[2] The simplest objection to such "finite frequency theories" is that we often make probability judgments that diverge from relative frequencies. For example, we can talk about a coin being fair (and so the generic probability of a flip landing heads is 0.5) even when it is flipped only once and then destroyed (in which case the relative frequency is either 1 or 0). For understanding such generic probabilities, we need a notion of probability that talks about *possible* instances of properties as well as actual instances. Theories of this sort are sometimes called "hypothetical frequency theories". C. S. Peirce was perhaps the first to make a suggestion of this sort. Similarly, the statistician R. A. Fisher, regarded by many as "the father of modern statistics", identified probabilities with ratios in a "hypothetical infinite population, of which the actual data is regarded as constituting a random sample" (1922, p. 311). Karl Popper (1956, 1957, and 1959) endorsed a theory along these lines and called the resulting probabilities *propensities*. Henry Kyburg (1974a) was the first to construct a precise version of this theory (although he did not endorse the theory), and it is to him that we owe

[2] 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.

the name “hypothetical frequency theories”. Kyburg (1974a) also insisted that von Mises should be considered a hypothetical frequentist. There are obvious difficulties for spelling out the details of a hypothetical frequency theory. More recent attempts to formulate precise versions of what might be regarded as hypothetical frequency theories are van Fraassen (1981), Bacchus (1990), Halpern (1990), Pollock (1990), Bacchus et al (1996). I will take my jumping-off point to be the theory of Pollock (1990), which I will sketch briefly in section three.

It has always been acknowledged that for practical decision-making we need singular probabilities rather than generic probabilities. So theories that take generic probabilities as basic need a way of deriving singular probabilities from them. Theories of how to do this are theories of *direct inference*. Theories of objective generic probability propose that statistical inference gives us knowledge of generic probabilities, and then direct inference gives us knowledge of singular probabilities. Reichenbach (1949) pioneered the theory of direct inference. The basic idea is that if we want to know the singular probability PROB(Fa), we look for the narrowest reference class (or reference property) G such that we know the generic probability prob(Fx/Gx) and we know Ga, and then we identify PROB(Fa) with prob(Fx/Gx). For example, actuarial reasoning aimed at setting insurance rates proceeds in roughly this fashion. Kyburg (1974) was the first to attempt to provide firm logical foundations for direct inference. Pollock (1990) took that as its starting point and constructed a modified theory with a more epistemological orientation. The present paper builds upon some of the basic ideas of the latter.

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. I believe that this broad collection of new defeasible inference schemes provides the solution to the problem of sparse probability knowledge and explains how probabilities can be truly useful even when we are ignorant about most of them.

3. Nomic Probability

Pollock (1990) developed a possible worlds semantics for objective generic probabilities,[3] 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 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:

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

$\text{prob}_x(Fx/Gx) = \rho(\mathfrak{F},\mathfrak{G})$.[4]

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:

<u>Finite Proportions</u>*:*

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

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 $\rho(X \cap Y,Z) = \rho(X,Z) \cdot \rho(Y,X \cap Z)$.

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:

<u>Universality</u>*:*

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

<u>Finite Set Principle:</u>

For any set B, $N > 0$, and open formula Φ,

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

[4] 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.

Projection Principle:

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

Crossproduct Principle:

If C and D are nonempty, $\rho(A \times B, C \times D) = \rho(A,C) \cdot \rho(B,D)$.

These four principles are all theorems of elementary set theory when the sets in question are finite. 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.

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 $\ulcorner Gc\ \&\ \mathrm{prob}(F/G) \ge r \urcorner$ is a defeasible reason for $\ulcorner Fc \urcorner$, the strength of the reason being a monotonic increasing function of r.

I take it that the statistical syllogism is a very intuitive principle, and it is clear that we employ it constantly in our everyday reasoning. For example, suppose you read in the newspaper that 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 intuitively compelling examples, that the statistical syllogism must be so constrained. 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 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 $\ulcorner Hc\ \&\ \mathrm{prob}(F/G\&H) < \mathrm{prob}(F/G) \urcorner$ is an undercutting defeater for the inference by the statistical syllogism from $\ulcorner Gc\ \&\ \mathrm{prob}(F/G) \ge r \urcorner$ to $\ulcorner Fc \urcorner$.[5]

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 problem of sparse probability knowledge by justifying a large collection

[5] 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).

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 a bridge theorem that relates 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 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 $\dfrac{n}{2}$, 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.

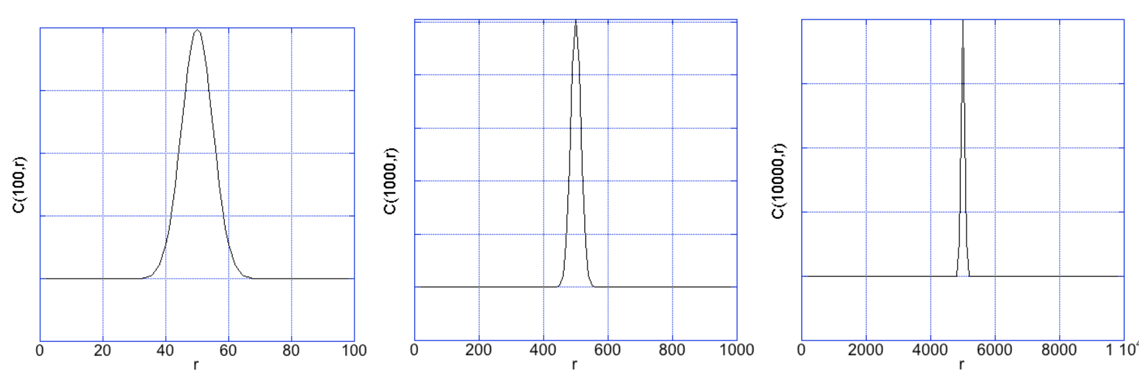

Figure 1. $C(n,r)$ for $n = 100$, $n = 1000$, and $n = 10000$.

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 $\#\left\{\left\langle X_1,\ldots,X_n\right\rangle \mid LC \;\&\; X_1,\ldots,X_n \subseteq U\right\} \geq K$.

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 about finite sets 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 $\#\left\{\left\langle X_1,\ldots,X_n\right\rangle \mid LC \;\&\; X_1,\ldots,X_n \subseteq U\right\} \geq N$ 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: 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. 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. It turns out that this can be done 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.

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. 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 $\mathfrak{F}$ 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 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: See appendix.

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:

<u>Probabilistic Indifference Principle:</u>

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

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

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 $\mathrm{prob}(F/G) \underset{\delta}{\approx} 0.5$, 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:

<u>Indifference Principle:</u>

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 Limit Principle 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:

<u>Expectable Indifference Principle:</u>

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:

<u>Principle of Expectable Values</u>

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.

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

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

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

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. 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 $r \cdot s$.

Proof: See appendix.

Thus we get:

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($A\&B/C$) = $r \cdot s$.

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 $\text{prob}(D/C) \neq 1$, then we should not expect that $\text{prob}(A\&B/C) = \text{prob}(A/C)\cdot\text{prob}(B/C)$. 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 $\text{prob}(A/C) = r$, $\text{prob}(B/C) = s$, $\text{prob}(D/C) = g$, $(A\&C) \preccurlyeq D$, and $(B\&C) \preccurlyeq D$, it follows that $\text{prob}(A/C\&D) = r/g$, $\text{prob}(B/C\&D) = s/g$, and the following values are expectable:

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

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

Proof: See appendix.

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:

$\ulcorner(A\&C) \preccurlyeq D$, $(B\&C) \preccurlyeq D$, and $\text{prob}(D/C) \neq 1\urcorner$ is an undercutting defeater for the inference from $\ulcorner\text{prob}(A/C) = r$ and $\text{prob}(B/C) = s\urcorner$ to $\ulcorner\text{prob}(A\&B/C) = r\cdot s\urcorner$ by the Principle of Statistical Independence.

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

7. Nonclassical Direct Inference

Pollock (1984) introduced (using somewhat 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 $\text{prob}(A/B\&C) = r$.

Proof: See appendix.

The defense of this principle in Pollock (1990) was complex, but we can now derive it very simply from the Probable Proportions and Probable Probabilities Theorems. 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 $C \preccurlyeq D \preccurlyeq B$, $\text{prob}(A/D) = s$, and $\text{prob}(A/B) = r$, then the expectable value of $\text{prob}(A/C) = s$ (rather than r).

Proof: See appendix.

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

Proof: $\text{prob}(A\&B/C) = \text{prob}(A/B\&C)\cdot\text{prob}(B/C)$, and by nonclassical direct inference, the expectable value of $\text{prob}(A/B\&C) = p$. ■

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

This is an interesting generalization of Nonclassical Direct Inference. Although probabilists common 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.

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).[6] Let $\mathbf{K}$ be the conjunction of all the propositions the agent is warranted in believing,[7] 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 $\mathfrak{P}$ is the set of all physically possible P-worlds:

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

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

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

This makes singular probabilities sensitive to the agent's knowledge of his situation, which is what is needed for rational decision making.[8] Formally, this is 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) PROB(Fa) = 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 PROB(P/Q) = prob($Rx_1 \ldots x_n / Sx_1 \ldots x_n$ & $x_1 = a_1$ & … & $x_n = a_n$ & $\mathbf{K}$).

(3) PROB(P) = 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 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

[6] Bacchus (1990) gave a somewhat similar account of direct inference, drawing on Pollock (1983 ,1984).
[7] 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.
[8] For a further complication, see the literature on causal probability, as discussed for example in Pollock (2006).

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:

<u>Classical Direct Inference:</u>

$\ulcorner Sa_1\ldots a_n$ is warranted and $\text{prob}(Rx_1\ldots x_n / \ Sx_1\ldots x_n \ \& \ Tx_1\ldots x_n) = r\urcorner$ is a defeasible reason for $\ulcorner\text{PROB}(Ra_1\ldots a_n \ / \ Ta_1\ldots a_n) = r\urcorner$.

Similarly, we get subproperty defeaters:

<u>Subproperty Defeat for Classical Direct Inference:</u>

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

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:

<u>Overlap Defeat for Classical Direct Inference:</u>

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) $\text{prob}(Gx_1\ldots x_n / \ Sx_1\ldots x_n \ \& \ Tx_1\ldots x_n) \neq 1$

is an undercutting defeater for the inference by classical direct inference from $\ulcorner Sa_1\ldots a_n$ is warranted and $\text{prob}(Rx_1\ldots x_n / \ Sx_1\ldots x_n \ \& \ Tx_1\ldots x_n) = r\urcorner$ to $\ulcorner\text{PROB}(Ra_1\ldots a_n \ / \ Ta_1\ldots a_n) = r\urcorner$.

Because singular probabilities are generic probabilities in disguise, we can also use Nonclassical Direct Inference to infer singular probabilities from singular probabilities. Thus $\ulcorner\text{PROB}(P/Q) = r\urcorner$ gives us a defeasible reason for expecting that $\text{PROB}(P/Q\&R) = r$. We can employ principles of statistical independence similarly. For example, $\ulcorner\text{PROB}(P/R) = r \ \& \ \text{PROB}(Q/R) = s\urcorner$ gives us a defeasible reason for expecting that $\text{PROB}(P\&Q/R) = r{\cdot}s$. 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 $\text{PROB}(Fc) = \Upsilon(r,s\,|a)$. I call this *computational inheritance*, because it computes a new value for $\text{PROB}(Fc)$ from previously known generic probabilities. Direct inference, by contrast, is a kind of "noncomputational inheritance". It is *direct* in that $\text{PROB}(Fc)$ 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.

$$\text{prob}(F/U) = a \text{ and } G,H \preccurlyeq U$$

$$\text{prob}(F/G) = r \qquad \text{prob}(F/H) = s$$

$$\text{prob}(F/G\&H) = \Upsilon(r,s\,|a)$$

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:

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

Then we can establish:

Y-Principle:

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

Proof: See appendix.

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.

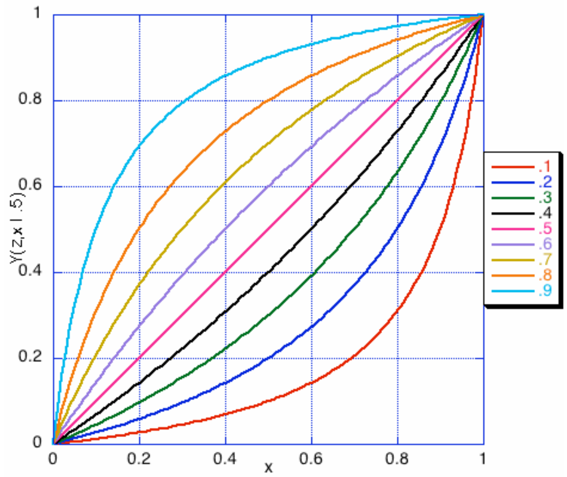

Figure 3. Y(z,x |.5), holding z constant
(for several choices of z as indicated in the key).

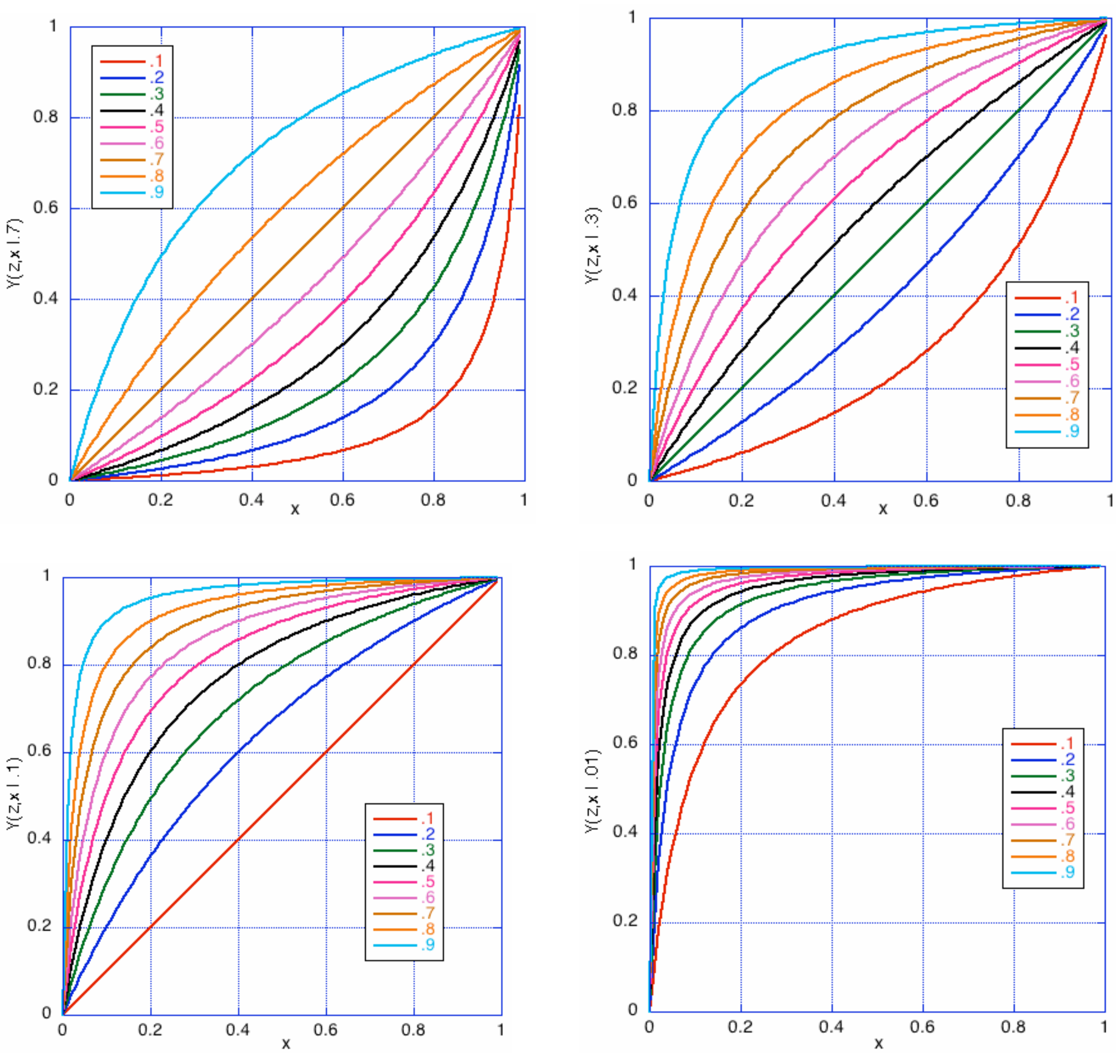

Figure 4. Y(z,x |a) holding z constant (for several
choices of z), for a = .7, a = .3, a = .1, and a = .01.

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 $Y(r,s\,|a) > s$.

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

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

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

Y-Defeat for Statistical Independence:

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

Proof: See appendix.

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 $\text{PROB}(P) = a$, $\text{PROB}(P/Q) = r$, and $\text{PROB}(P/R) = s$ then $\text{PROB}(P/Q\&R) = Y(r,s\,|a)$.

The application of the Y-function presupposes that we know the base rate $\text{prob}(A/U)$. But suppose we do not. Then what can we conclude about $\text{prob}(A/B\&C)$? It might be supposed that we can combine Indifference and the Y-Principle and conclude that $\text{prob}(A/B\&C) = Y(r,s\,|.5)$. 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).[9] However, by ignoring the base rate $\text{prob}(A/U)$, 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 $\text{prob}(A/U)$ it is reasonable to expect it to be .5, knowing the values of $\text{prob}(A/B)$ and $\text{prob}(A/C)$ changes the expectable value of $\text{prob}(A/U)$. Let us define $Y_0(r,s)$ to be $Y(r,s\,|a)$ where a, b, and c are the solutions

[9] 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.

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:

Y_0-Principle:

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

Proof: See appendix.

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 have a simple analytic characterization. $Y_0(r,s)$ is plotted in figure 5, 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).

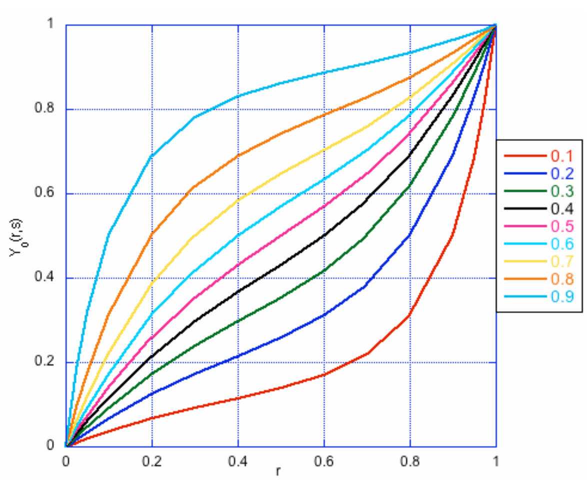

Figure 5. $Y_0(r,s)$, holding s constant (for several choices of s as indicated in the key)

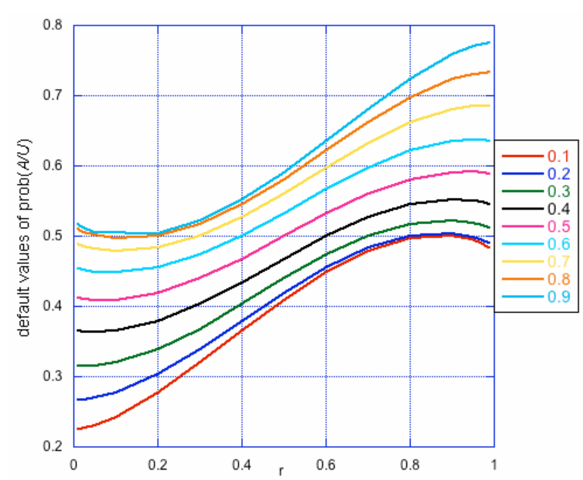

Figure 6. Default values of prob(A/U) (for several choices of s as indicated in the key)

10. 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, and the Y-Principle. 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

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,

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

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

Thus we have the Finite Indifference Principle .

<u>Probable Proportions Theorem:</u>

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 $\#\{\langle X_1,\ldots,X_n\rangle \mid LC \,\&\, X_1,\ldots,X_n \subseteq U\} \geq N$ 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 $i \leq k$, $\varepsilon < \frac{b_i(\xi)}{s_1}$, and $\xi - \varepsilon$ is an allowable value only if for every $i \leq k$, $\varepsilon < \frac{a_i(\xi)}{r_1}$.

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 $\frac{\mathrm{Prd}\left(x + \frac{1}{u}\right)}{\mathrm{Prd}(x)} \to 1$ 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.)

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

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

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 $\mathrm{Prd}(x)$ 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:

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

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:

<u>Partition Principle:</u>

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

$$a_1^{z_1}\cdot \ldots \cdot a_n^{z_n} < z_1a_1 + \ldots + z_na_n .$$

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 $\left(\dfrac{x_1}{x_1 - r_1\varepsilon}\right) > 1 > \left(\dfrac{y_1}{y_1 + s_1\varepsilon}\right)$, 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 $1-\varepsilon$.

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 $i\le k$, $\varepsilon<\frac{b_i(\xi)}{s_1}$, then $\frac{\mathrm{Prd}(\xi+\varepsilon)}{\mathrm{Prd}(\xi)}\to\infty$ as $u\to\infty$, and (2) if for every $i\le k$, $\varepsilon<\frac{a_i(\xi)}{r_1}$, 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.

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

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

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 $u\to\infty$,

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

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

Similarly,

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

Therefore,

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

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

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

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

Hence,

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

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

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

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

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

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

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

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

"$\rho(A/\{x_1,\ldots,x_N\}) = r$" is equivalent to the disjunction of $\dfrac{N!}{(rN)!((1-r)N)!}$ 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,

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

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

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

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

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

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

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 $\delta > 0$:

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

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 $\big\{\langle X_1,\ldots,X_n\rangle \mid LC \ \& \ X_1,\ldots,X_n \subseteq U^*\big\} \geq N$, then

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

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 $\{\langle X_1,\ldots,X_n\rangle / \; LC \;\&\; X_1,\ldots,X_n \subseteq U\}$ 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 $s = 1$. ■

<u>Probable Probabilities Theorem:</u>

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:

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

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,

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

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

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

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

Y-Principle:

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

Proof: Here is a machine-generated proof, produced by executing the following instruction:

```
(analyze-probability-structure
  :subsets '(A B C)
  :constants '(a b c r s)
  :constraints '((ab = (* r b))
                 (ac = (* s c)))
  :probability-queries '(prob(A / (B & C)))
  :independence-queries '((B C A) (B C ~A) (B C U))
  :parallel-term-characterizations t
  :display-details t
  :display-infix t)
```

```
Dividing U into 3 subsets A,B,C whose probabilities relative to U are a, b, c,
if the following constraints are satisfied:
    ab = (r * b)
    ac = (s * 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:
   {
    ((r * b) - abc) * u
    abc * u
    ((a + abc) - ((r * b) + (s * c))) * u
    ((s * c) - abc) * u
    ((b + abc) - ((r * b) + bc)) * u
    (bc - abc) * u
    (((r * b) + 1 + (s * c) + bc) - (a + b + abc + c)) * u
    ((c + abc) - ((s * c) + bc)) * u
   }
The subset of terms in the term-set that contain abc is:
        {
        ((r * b) - abc) * u
        abc * u
        ((a + abc) - ((r * b) + (s * c))) * u
        ((s * c) - abc) * u
        ((b + abc) - ((r * b) + bc)) * u
        (bc - abc) * u
        (((r * b) + 1 + (s * c) + bc) - (a + b + abc + c)) * u
        ((c + abc) - ((s * c) + bc)) * u
        }
The expectable-value of abc is then the real-valued solution to the following equation:
1 = ((((r * b) - abc) ^ (((r * b) - (abc + 1)) - ((r * b) - abc)))
        * (abc ^ ((abc + 1) - abc))
        * (((a + abc) - ((r * b) + (s * c))) ^ (((a + (abc + 1)) - ((r * b) + (s * c))) - ((a + abc) - ((r * b) + (s * c)))))
        * (((s * c) - abc) ^ (((s * c) - (abc + 1)) - ((s * c) - abc)))
        * (((b + abc) - ((r * b) + bc)) ^ (((b + (abc + 1)) - ((r * b) + bc)) - ((b + abc) - ((r * b) + bc))))
        * ((bc - abc) ^ ((bc - (abc + 1)) - (bc - abc)))
        * ((((r * b) + 1 + (s * c) + bc) - (a + b + abc + c))
           ^ ((((r * b) + 1 + (s * c) + bc) - (a + b + (abc + 1) + c)) - (((r * b) + 1 + (s * c) + bc) - (a + b + abc + c))))
        * (((c + abc) - ((s * c) + bc)) ^ (((c + (abc + 1)) - ((s * c) + bc)) - ((c + abc) - ((s * c) + bc)))))
   = ((((r * b) - abc) ^ (- 1))
        * (abc ^ 1)
```

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

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 = ((s * c * r * b) / a)
----------
bc = ((((a + (s * r)) - ((s * a) + (r * a))) * c * b) / (a * (1 - a)))
============================================================
Reconstruing a prob(A / U), the following characterizations were found for the expectable values of the probabilities
wanted:
----------
prob(A / (B & C)) = (((1 - a) * s * r) / ((a + (s * r)) - ((r * a) + (s * a))))
----------
```

B and C are STATISTICALLY INDEPENDENT relative to A
B and C are STATISTICALLY INDEPENDENT relative to ~A

B and C are NOT statistically independent relative to U
(prob(B / U) * prob(C / U)) ÷ prob((B & C) / U) =
((a * (1 - a)) / ((a + (s * r)) - ((s * a) + (r * a))))

■

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.

This is established by lemma 2. To prove lemma 2 we first prove:

Lemma 1: If $x \neq x^*$ and $y \neq y^*$ then $xy + (1-x)y^* \neq x^*y + (1-x^*)y^*$.

Proof: Suppose $xy + (1-x)y^* = x^*y + (1-x^*)y^*$. $(x - x^*)y = (x - x^*)y^*$, so either $(x - x^*) = 0$ or $y = y^*$. ■

Lemma 2: 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).

Proof: Suppose the antecedent.

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

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

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

If prob(C/A) ≠ prob(C/U) then prob(C/U&A) ≠ prob(C/U&~A). Furthermore, prob(A/B) = prob(B/A)$\cdot\dfrac{\text{prob}(A/U)}{\text{prob}(B/U)}$, so prob($A$/$B$) = prob($A$/$U$) iff prob($B$/$A$) = prob($B$/$U$). Thus prob($A$/$B$) ≠ prob($A$/$U$), so by lemma 1, prob($C$/$B$) ≠ prob($C$/$U$). ■

Y_0-Principle:

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

Proof: Proceeding as usual, the expectable values of abp, a, b, p, and ab are the solutions to the following set of equations:

$$\frac{abp\ (a - ab + abp - a\,r)\ (-ab + abp + b - b\,s)\ (abp + p - a\,r - b\,s)}{(ab - abp)\ (-abp + a\,r)\ (-abp + b\,s)\ (1 - a + ab - abp - b - p + a\,r + b\,s)} = 1;$$

$$\left(\frac{a - ab + abp - a\,r}{1 + ab + r\,a + s\,b - a - b - abp - p}\right)^{1-r}\left(\frac{-abp + a\,r}{abp + p - a\,r - b\,s}\right)^{r} = 1;$$

$$\left(\frac{-ab + abp + b - b\,s}{1 - a + ab - abp - b - p + a\,r + b\,s}\right)^{1-s}\left(\frac{-abp + b\,s}{abp + p - a\,r - b\,s}\right)^{s} = 1;$$

$$\frac{abp + p - ar - bs}{1 - a + ab - abp - b - p + ar + bs} = 1;$$

$$\frac{(ab - abp)(1 - a + ab - abp - b - p + ar + bs)}{(a - ab + abp - ar)(-ab + abp + b - bs)} = 1.$$

As in the Y-principle, solving the first and fifth equation for *ab* and *abp* produces:

$$abp = \frac{abrs}{p};$$

$$ab = \frac{ab(p - pr - ps + rs)}{p(1 - p)}.$$

Substituting these values into the other equations produces the triple of equations that define $Y_0(r,s)$. ■

References

Bacchus, Fahiem
1990 *Representing and Reasoning with Probabilistic Knowledge*, MIT Press.
Bacchus, Fahiem, Adam J. Grove, Joseph Y. Halpern, Daphne Koller
1996 "From statistical knowledge bases to degrees of belief", *Artificial Intelligence* **87**, 75-143.
Braithwaite, R. B.
1953 *Scientific Explanation*. Cambridge: Cambridge University Press.
Carnap, Ruldolph
1950 *The Logical Foundations of Probability*. Chicago: University of Chicago Press.
1952 *The Continuum of Inductive Methods*. Chicago: University of Chicago Press.
Fisher, R. A.
1922 "On the mathematical foundations of theoretical statistics." *Philosophical Transactions of the Royal Society A*, 222, 309-368.
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.
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.
Kneale, William
1949 *Probability and Induction*. Oxford: Oxford University Press.
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.
Pollock, John L.
1983 "A theory of direct inference", *Theory and Decision* **15**, 29-96.
1984 "Foundations for direct inference". *Theory and Decision* **17**, 221-256.
1990 *Nomic Probability and the Foundations of Induction*, New York: Oxford University Press.
1995 *Cognitive Carpentry*, Cambridge, MA: Bradford/MIT Press.
2006 *Thinking about Acting: Logical Foundations for Rational Decision Making*, New York: Oxford University Press.
2006a "Defeasible reasoning", in *Reasoning: Studies of Human Inference and its Foundations*, (ed) Jonathan Adler and Lance Rips, Cambridge: Cambridge University Press.
Popper, Karl
1938 "A set of independent axioms for probability", *Mind* **47**, 275ff.
1956 "The propensity interpretation of probability." *British Journal for the Philosophy of Science* **10**, 25-42.

1957 “The propensity interpretation of the calculus of probability, and the quantum theory.” In *Observation and Interpretation*, ed. S. Körner, 65-70. New York: Academic Press.
1959 *The Logic of Scientific Discovery*, New York: Basic Books.
Reichenbach, Hans
1949 *A Theory of Probability*, Berkeley: University of California Press. (Original German edition 1935.)
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.
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.
van Fraassen, Bas
1981 *The Scientific Image*. Oxford: Oxford University Press.
Venn, John
1888 *The Logic of Chance*, 3rd ed. London.