CARNAP AND THE LOGIC OF INDUCTIVE INFERENCE

S. L. Zabell

1 INTRODUCTION

This chapter discusses Carnap's work on probability and induction, using the notation and terminology of modern mathematical probability, viewed from the perspective of the modern Bayesian or subjective school of probability. (It is a much expanded and more mathematical version of [Zabell, 2007]). Carnap initially used a logical notation and terminology that made his work accessible and interesting to a generation of philosophers, but it also limited its impact in other areas such as statistics, mathematics, and the sciences. Using the notation of modern mathematical probability is not only more natural, but also makes it far easier to place Carnap's work alongside the contributions of such other pioneers of epistemic probability as Frank Ramsey, Bruno de Finetti, I. J. Good, L. J. Savage, and Richard Jeffrey.

Carnap's interest in logical probability was primarily as a tool, a tool to be used in understanding the quantitative confirmation of an hypothesis based on evidence and, more generally, in rational decision making. The resulting analysis of induction involved a two step process: one first identified a broad class of possible confirmation functions (the *regular c-functions*), and then identified either a unique function in that class (early Carnap) or a parametric family (later Carnap) of specific confirmation functions. The first step in the process put Carnap in substantial agreement with subjectivists such as Ramsey and de Finetti; it is the second step, the attempt to limit the class of probabilities still further, that distinguishes Carnap from his subjectivist brethren.

So: precisely what are the limitations that Carnap saw as natural to impose? In order to discuss these, we must begin with his concept**S** of probability.

2 PROBABILITY

The word 'probability' has always had a multiplicity of meanings. In the beginning mathematical probability had a meaning that was largely *epistemic* (as opposed to *aleatory*); thus for Laplace probability relates in part to our knowledge and in part to our ignorance. During the 19th century, however, empirical alternatives

Handbook of the History of Logic. Volume 10: Inductive Logic. Volume editors: Dov M. Gabbay, Stephan Hartmann and John Woods.
General editors: Dov M. Gabbay, Paul Thagard and John Woods.

arose. In the years 1842 and 1843, no fewer than four independent proposals for an objective or frequentist interpretation were first advanced: those of Jakob Friedrich Fries in Germany, Antoine Augustin Cournot in France, and John Stuart Mill and Robert Leslie Ellis in England. Less than a quarter of a century later, John Venn's *Logic of Chance* [Venn, 1866], the first book in English devoted exclusively to the philosophical foundations of probability, took a purely frequentist view of the subject.

Ramsey, in advancing his view of a quantitative subjective probability based on a consistent system of preferences [Ramsey, 1926], deftly side-stepped the debate by conceding that the frequency interpretation of probability was a perfectly reasonable one, one which might have considerable value in science, but argued that this did not preclude a subjective interpretation as well. During the 20th century the debate became increasingly more complex, von Mises, Reichenbach, and Neyman advancing frequentist views, and Keynes, Ramsey, and Jeffreys competing logical or subjective theories.

Carnap sought to bring order into this chaos by introducing the concepts of *explicandum* and *explicatum*. Sometimes philosophical debates arise unnecessarily due to the use of ill-defined (or even undefined) concepts. For example, an argument about whether or not viruses constitute a form of life can only really arise from a failure to define just what one means by life; define the term and the status of viruses (whose structure and function are in many cases very well understood) will become clear one way or the other. This is essentially an operationalist or logical positivist perspective, a legacy of Carnap's days in the Vienna Circle. For Carnap the explicandum was the ill-defined concept; the explicatum the clarification of it that someone advanced.

But probability did not involve just a dispute over the explication of a term. The term itself did double duty, being used by some in an epistemic fashion (the degree of belief in a proposition or event), and by others in an aleatory fashion (a frequency in a class or series). To unravel the Gordian knot of probability, one had to sever the two concepts and recognize that there are two distinct explicanda, each requiring separate exegesis.

2.1 Early views

In his paper "The two concepts of probability" [1945b], Carnap introduced the terms *probability*$_1$ and *probability*$_2$, the first referring to probability in its guise as a measure of confirmation, the second as a measure of frequency. This had twin advantages: putting the issue so clearly, debates about the one true meaning of probability became less credible; and the more neutral terminology helped shift the argument from issues of linguistic useage (which, after all, vary from one language to another), to conceptual explication. These ideas were developed at great length in Carnap's magisterial *Logical Foundations of Probability* [1950], probabilities being assigned to sentences in a formal language. In his later work Carnap discarded sentences (which he viewed as insufficiently expressive for his purposes) in favor of events or propositions, which he regarded as essentially equivalent, and we shall adopt this viewpoint. (The main technical complication in working at the level of sentences is that more than one sentence can assert the same proposition; for example, $\alpha \wedge \beta$ and $\neg(\neg\alpha \vee \neg\beta)$.)

Carnap's approach was a direct descendant of Wittgenstein's relatively brief remarks on probability in the *Tractatus*, later developed at some length by Waismann [1930]. Carnap, following Waismann, assumed the existence of a *regular measure function* $m(x)$ on sentences, defining these by first assuming a normalized nonnegative function on molecular sentences and then extending these to all sentences. Carnap then defined in the usual way $c(h, e)$, the conditional probability of a proposition h given the proposition e, as the ratio $m(h \wedge e)/m(e)$.

Carnap interpreted the conditional probabilities $c(h, e)$ as a measure of the extent to which evidence e confirms hypothesis h. Such functions had already been studied by Janina Hosiasson-Lindenbaum [1940] a decade earlier. Unlike Carnap, Hosiasson-Lindenbaum took a purely axiomatic approach: she studied the general properties of confirmation functions $c(h, e)$, assuming only that they satisfied a basic set of axioms. There are several equivalent versions of this set appearing in the literature; here is one particularly natural formulation:

The axioms of confirmation

1. $0 \leq c(h, e) \leq 1$.
2. If $h \leftrightarrow h'$ and $e \leftrightarrow e'$, then $c(h, e) = c(h', e')$.
3. If $e \to h$, then $c(h, e) = 1$.
4. If $e \to \neg(h \wedge h')$, then $c(h \vee h', e) = c(h, e) + c(h', e)$.
5. $c(h \wedge h', e) = c(h, e) \cdot c(h', h \wedge e)$.

Carnap's conditional probabilities $c(h, e)$ satisfied these axioms (and so were plausible candidates for confirmation functions).

2.2 Betting odds and Dutch books

But just what do the numbers $m(e)$ or $c(h, e)$ represent? It was one of the great contributions of Ramsey and de Finetti to advance *operational definitions* of subjective probability; for Ramsey, primarily as arising from preferences, for de Finetti as fair odds in a bet. By then imposing rationality criteria on such quantities, both were able to derive the standard axioms for finitely additive probability. Ramsey, in a remarkable *tour-de-force*, was able to demonstrate the simultaneous existence of utility and probability functions $u(x)$ and $p(x)$. He did this by imposing natural consistency constraints on a (sufficiently rich) set of preferences, introducing the device of the *ethically neutral proposition* (the philosophical equivalent of tossing a fair coin) as a means of interpolating between competing alternatives. The functions $u(x)$ and $p(x)$ track one's preferences in the sense that one action is preferred to another if and only if its expected utility is greater than the other. (Jeffrey [1983] discusses Ramsey's system and presents an extremely interesting variant of it.)

De Finetti, in contrast, initially gave primacy to probabilities interpreted as betting odds. (If p is a probability, then the corresponding odds are $p/(1-p)$.) The odds represent a bet *either side of which one is willing to take*. (Thus, the odds of 2 : 1 in favor of an event means that one would accept either a bet of 2 : 1 for, or a bet of 1 : 2 against. This is somewhat akin to the algorithm for two children dividing a cake: one divides the cake into two pieces, the other chooses one of the two pieces.) De Finetti imposed as his rationality constraint the requirement that these odds be *coherent*; that is, that it be impossible to construct a *Dutch book* out of them. (In a Dutch book, an opponent can choose a portfolio of bets such that he is assured of winning money. The existence of a Dutch book is analogous to the existence of arbitrage opportunities in the derivatives market.) A conditional probability $P(A \mid B)$ in de Finetti's system is interpreted as a *conditional bet* on A, available only if B is determined to have happened. De Finetti was able to show that the probabilities corresponding to a coherent set of bettings odds must satisfy the standard axioms of finitely additive probability. For example, if one takes the axioms for confirmation listed in the previous subsection, all are direct consequences of coherence.

John Kemeny, one of Carnap's collaborators in the 1950s, proved a beautiful converse to this result [Kemeny, 1955]. He showed that the above five properties of a confirmation function are at once both necessary and sufficient for coherence. That is, although de Finetti had in effect shown that coherence implies the five axioms, in principle there might be other, incoherent confirmation functions also satisfying the five axioms. If one did not begin by accepting (coherent) betting odds as the operational interpretation of $c(h, e)$, this left open the possibility of other confirmation functions, ones not falling into the Ramsey and de Finetti framework. The power of Kemeny's result is that if one accepts the five axioms above as necessary desiderata for any confirmation function $c(h, e)$, then such functions necessarily assign coherent betting odds to the universe of events. This was a powerful argument in favor of the betting odds interpretation, and it persuaded Carnap, who adopted it. Thus, while in *The Logical Foundations of Probability* Carnap had advanced no fewer than three possible interpretations for probability$_1$ — evidential support, fair betting quotients, and estimates of statistical frequencies — in his later work he explicitly abandoned the first of these, and wrote almost exclusively in terms of the second. (The "normative" force of Dutch book arguments has of course been the subject of considerable debate. Armendt [1993] contains a balanced discussion of the issues and provides a useful entry into the literature.)

Nevertheless, even accepting the subjective viewpoint, the issue remains: can the inductive confirmation of hypotheses be understood in quantitative terms? It was this later question that was of primary interest to Carnap, and the one to which he turned in a second paper "On inductive logic" [1945a].

3 CONFIRMATION

In order to better appreciate Carnap's analysis of the inductive process, let us briefly review the background against which he wrote.

First some basic mathematical probability. Suppose we have an uncertain event that can have one of two possible outcomes, arbitrarily termed "success" and "failure", and let S_n denote the number of successes in n instances ("trials"). If the trials are independent, and have a constant probability p of success, then the probability of k successes in the n trials is given by the *binomial distribution*:

$$P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad 0 \leq k \leq n.$$

Here

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

is the *binomial coefficient*, and $n! = n \cdot (n-1) \cdot (n-2) \ldots 3 \cdot 2 \cdot 1$.

Suppose next that the probability p is itself random, with some probability distribution $d\mu(p)$ on the unit interval. For example, success and failure might correspond to getting a head or tail when tossing a ducat, and the ducat is chosen from a bag of ducats having variable probability p of coming up heads (reflecting the composition of coins in the bag). In this case the probability $P(S_n = k)$ is obtained by averaging the binomial probabilities over the different possible values of p. This average is standardly given by an integral, namely

$$P(S_n = k) = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} d\mu(p), \qquad 0 \leq k \leq n,$$

In our example $d\mu(p)$ is aleatory in nature, tied to the composition of the bag. But it could just as well be taken to be epistemic, reflecting our degree of belief regarding the different possible values of p.

3.1 The rule of succession

In this analysis there are several important questions as yet unanswered. In particular, the nature of p (is it a physical probability or a degree of belief?) has not been specified, and no guidance has been given regarding the origin of the *initial* or *prior* distribution $d\mu(p)$. In particular, even if the nature of p is specificed, how does one determine the prior distribution $d\mu(p)$? For Laplace and his school, one had resort to the *principle of indifference*: lacking any reason to favor one value of p over another, the distribution was taken to be uniform over the unit interval: $d\mu(p) = dp$. In this case the integral simplifies to give:

$$P(S_n = k) = \frac{1}{n+1}, \qquad 0 \leq k \leq n.$$

But in fact the Reverend Thomas Bayes, the eponymous founder of the subject of Bayesian statistics, employed a subtler argument that paralleled Carnap's later approach. Bayes [1764] reasoned that in a case of complete ignorance ("an event concerning the probability of which we absolutely know nothing antecedently to any trials made concerning it"), one has $P(S_n = k) = 1/(n+1)$ for all $n \geq 1$ and $0 \leq k \leq n$ (in effect Bayes takes the later to be the definition of the former), and this in turn implies that the prior must be uniform.

The argument can in fact be made rigorous. Let $k = n$; then Bayes's postulate $P(S_n = k) = 1/(n+1)$ tells us that

$$\int_0^1 p^n d\mu(p) = \frac{1}{n+1} = \int_0^1 p^n dp, \quad n \geq 1.$$

Thus the as yet unknown probability $d\mu(p)$ has the same moments as the so-called "flat" prior dp. But the Hausdorff moment theorem tells us that a probability measure on a compact set (here $[0,1]$) is characterized by its moments. Thus $d\mu(p)$ and dp, having the same moments, must coincide.

Given the Bayes-Laplace formula $P(S_n = k) = 1/(n+1)$, it is a simple matter to derive the corresponding *predictive probabilities*. If, for example, X_j is a so-called indicator variable taking the values 1 or 0, depending on whether the outcome of the j-th trial is a success or failure, respectively (so that the number of successes S_n is $X_1 + ... + X_n$), then $P(X_{n+1} = 1 \mid S_n = k)$ is the conditional probability of a success on the next trial, based on the experience of the past n trials. Since the formula for conditional probability is $P(A \mid B) = P(A \text{ and } B)/P(B)$, it follows after a little algebra that

$$P(X_{n+1} \mid S_n = k) = \frac{k+1}{n+2}.$$

This is the celebrated (or infamous) *rule of succession*. Both it and the controversial principle of indifference on which it was based were the subject of harsh criticism beginning in the middle of the 19th century; see Zabell [1989]. Stigler [1982] argues that Bayes's form of the indifference postulate, applying as it does to the discrete outcome k, does not entail the same paradoxes as the principle of indifference applied to the continuous parameter p. But Bayes's ingenious argument was forgotten, and Laplace's approach became the focus of controversy. The Cambridge phenom Robert Leslie Ellis objected in the 1840s that one could not conjure something out of nothing: *ex nihilo nihil*; the German Johann von Kries countered in 1886 that one could invoke instead the principle of *cogent reason*: alternatives are judged equipossible because our knowledge is distributed equally among them; the point is the equi-distribution of knowledge rather than nihilist ignorance. In pragmatic England the Oxford statistician and economist F. Y. Edgeworth argued the use of flat priors was justified on approximate empirical grounds; the Cambridge logician and antiquarian John Venn ridiculed the use of the rule of succession. In France the distinguished Joseph Bertrand challenged the cogency of subjective probability; the even more distinguished Henri Poincaré championed it.

This was the decidedly unsatisfactory state of affairs in 1921, the year when John Maynard Keynes's *Treatise on Probability* appeared. Keynes's *Treatise* contains a useful summary of much of this debate. The next several decades saw increasing clarification of the foundations of probability and its use in inductive inference. But the particular thread we are interested in here involves a curious development that took place in two independent stages.

4 EXCHANGEABILITY

In 1924 William Ernest Johnson, an English logician and philosopher at King's College, Cambridge, published the third volume of his *Logic*. In an appendix at the end, Johnson suggested an alternative analysis to the one just discussed, one which represented a giant step forward. But despite the respect accorded him in Cambridge, Johnson had only limited influence outside it, and after his death in 1931, his work was little noted. It is one of the ironies of this subject that Carnap later followed essentially the same route as Johnson, but to much greater effect, in part because Carnap's *Logical Foundations of Probability* embedded his analysis in a much more detailed setting, and in part because he continued to refine his treatment of the subject for nearly two decades (whereas Johnson died only a few years after the appearance of his book).

Johnson's analysis contained several elements of novelty. The first two of these were designed to meet the two basic objections that had been raised regarding the classical rule of succession: its appeal to the so-called "principle of indifference", and its appeal by way of analogy to drawing balls from an urn.

4.1 Multinomial sampling

First, Johnson considered the case of $t \geq 2$ equipossible cases (instead of just two). This was no mere technical generalization. In many of the most telling attacks on the principle of indifference, situations were considered where it was unnatural to think of the outcome of interest as being one of two equipossible competing alternatives. By encompassing the multinomial case (several possible categories rather than just two) Johnson's analysis applied to situations in which the multiple competing outcomes are either naturally viewed as equipossible (for example, rolling a fair, six-sided die), or can be further broken down into equipossible subcases.

4.2 The permutation postulate

Second, Johnson presciently introduced the concept of *exchangeability*. Let us consider a sequence of random outcomes $X_1, ..., X_n$, each taking on one of t possible types $c_1, ..., c_t$. (For example, you are on the Starship Enterprise, and each time you encounter someone, they are either Klingon, Romulan, or Vulcan, so that $t = 3$.) Then a typical probability of interest is of the form

$$P(X_1 = e_1, X_2 = e_2, ..., X_n = e_n), \qquad e_i \in \{c_1, ..., c_t\}, \quad 1 \leq i \leq t.$$

In the classical inductive setting, the *order* of these observations is irrelevent, the only thing that matters being the counts or frequencies observed for each of the t categories. (More complex situtations will be discussed later.) Thus, if n_i is the number of X_j falling into the i-th category, it is natural to assume that all sequences $X_1 = e_1, X_2 = e_2, ..., X_n = e_n$ having the same frequency counts $n_1, n_2, ..., n_t$ have the same probability. Johnson termed this assumption the *permutation* postulate. (Carnap called the sequences $e_1, ..., e_n$ *state descriptions*, the frequency counts $n_1, ..., n_t$ *structure descriptions*, and made the identical symmetry assumption.)

The valid application of the rule of succession presupposes, as Boole notes, the aptness of the analogy between drawing balls from an urn — the *urn of nature*, as it was later called — and observing an event [Boole 1854, p. 369]. As Jevons [1874, p. 150] put it, "nature is to us like an infinite ballot-box, the contents of which are being continually drawn, ball after ball, and exhibited to us. Science is but the careful observation of the succession in which balls of various character present themselves . . . ".

The importance of Johnson's "permutation postulate" is that *it is no longer necessary to refer to the urn of nature*. To what extent is observing instances like drawing balls from an urn? Answer: to the extent that the instances are judged exchangeable. Venn and others, having attacked the rote use of the rule of succession, rightly argued that some additional assumption, other than mere repetition of instances, was necessary for valid inductive inference. From time to time various names for such a principle have been advanced: Mill's "Uniformity of Nature"; Keynes's "Principle of Limited Variety"; Goodman's "projectibility". It was Johnson's achievement to have realized both that 'the calculus of probability does not enable us to infer any probability-value unless we have some probabilities or probability relations given' [Johnson, 1924, p. 182]; and that the vague, verbal formulations of his predecessors could be captured in the mathematically precise formulation of exchangeability.

The permutation postulate (the assumption of *exchangeability* in modern parlance) was later independently introduced by the Italian Bruno de Finetti (see, for example, [de Finetti, 1937]), and became a centerpiece of his theory. For our purposes here, the basic point is that if the sequence is assumed to be exchangeable, then an assignment of probabilities to sequences of outcomes $e_1, e_2, ..., e_n$ reduces to assigning probabilities $P(n_1, n_2, ..., n_t)$ to sequences of frequency counts $n_1, n_2, ..., n_t$. This is because there are (using the standard notation for the *multinomial coefficient*)

$$\binom{n}{n_1\ n_2\ ...\ n_t} = \frac{n!}{n_1!\, n_2! \ldots n_t!}$$

different possible sequences $e_1, e_2, ..., e_n$ having the same set of frequency counts $n_1, n_2, ..., n_t$, and each of these is assumd to be equally likely, so by exchangeability and the additivity of probability

$$P(n_1, n_2, ..., n_t) = \left(\frac{n!}{n_1!\, n_2! \dots n_t!}\right) P(e_1, e_2, ..., e_n).$$

(That is, the probability of a state description $e_1, ..., e_n$, times the number of state descriptions having the same corresponding structure description $n_1, ..., n_t$, gives the probability of that structure description.)

It is a simple but nevertheless instructive exercise to verify that the predictive probabilities in this case take on a simple form:

$$P(X_{n+1} = c_i \mid X_1 = e_1, X_2 = e_2, ..., X_n = e_n) = P(X_{n+1} = c_i \mid n_1, n_2, ..., n_t).$$

(That is, although the conditional probability apparently depends on the entire state description $e_1, ..., e_n$, in fact it only depends on the corresponding structure description $n_1, ..., n_t$.)

In statistical parlance this last property is summarized by saying that the frequencies $n_1, ..., n_t$ are *sufficient statistics*: no information is lost in summarizing the sequence $e_1, ..., e_n$ by the counts $n_1, ..., n_t$. Such statistics turn out to be a powerful tool in extensions of exchangeability discovered in recent decades; see, e.g., [Diaconis and Freedman, 1984].

4.3 The combination postulate

But what do we choose for $P(n_1, n_2, ..., n_t)$? In the case $t = 2$, this reduces to assigning probabilities to the pairs (n_1, n_2). A little thought will show that Bayes's postulate (that the different possible frequencies k are equally likely) is equivalent to assuming that the different pairs (n_1, n_2) are equally likely (since $n_1 = k, n_2 = n - n_1$ and n is fixed). This in turn suggests the probability assignment that takes each of the possible structure descriptions to be equally likely, and this is in fact the path that both Johnson and Carnap initially took (Johnson termed this the *combination postulate*). Since there are

$$\binom{n+t-1}{t}$$

possible structure descriptions (also known as "ordered t-partitions of n", a well-known combinatorial fact, see, e.g., [Feller, 1968, p. 38]), and each of these is assumed equally likely, one has

$$P(n_1, n_2, ..., n_t) = \frac{1}{\binom{n+t-1}{t}}.$$

Together, the combination and permutation postulates uniquely determine the probability of any specific finite sequence; if a state description $e_1, e_2, ..., e_n$ has structure description $n_1, n_2, ..., n_t$ then its probability is

$$P(e_1, e_2, ..., e_n) = \frac{1}{\binom{n+t-1}{t}\binom{n}{n_1\ n_2\ ...\ n_t}};$$

see Johnson [1924, appendix on eduction]. This is Carnap's $m^\star$ function.

Having thus specified the probabilities of the "atomic" sequences, all other probabilities, including the rules of succession, are completely determined. Some simple algebra in fact yields

$$P(X_{n+1} = c_i \mid n_1, n_2, ..., n_t) = \frac{n_i + 1}{n + t};$$

see Johnson [1924]. This is Carnap's $c^\star$ function.

5 THE CONTINUUM OF INDUCTIVE METHODS

Although the mathematics of the derivation of the $c^\star$ system is certainly attractive, its assumption that all structure descriptions are equally likely is hardly compelling, and Carnap soon turned to more general systems.

It is ironic that here too his line of attack very closely paralleled that of Johnson. After criticisms from C. D. Broad [1924] and others, Johnson devised a more general postulate, later termed by I. J. Good [1965] the *sufficientness postulate*. This assumes that the predictive probabilities for a particular type i are a function of how many observations of the type have been seen already (n_i), and the total sample size n. It is a remarkable fact that this characterizes the predictive probabilities or rules of succession (and therefore the probability of any sequence).

5.1 The Johnson-Carnap continuum

Suppose $X_1, X_2, ..., X_n, ...$ represent an infinite sequence of observations, each assuming one of (the same) t possible values, and that at each stage n the sequence satisfies the permutation postulate. (In modern parlance, one has an infinitely exchangeable, t-valued sequence of random variables.) Assume the sequence satisfies the following three conditions:

1. Any state description $e, ..., e_n$ is *a priori* possible: $P(e_1, ...e_n) > 0$.
2. The "sufficientness postulate" is satisfied:

$$P(X_{n+1} = e_i \mid n_1, ..., n_t) = f_i(n_i, n).$$

3. There are at least three types of species; $t \geq 3$.

Then (unless the outcomes are independent of each other, so that observing one or more provides no predictive power regarding the others) the predictive probabilities have a very special form: there exist positive constants $\alpha_1, ..., \alpha_t$ such that if $\alpha = \alpha_1 + ... + \alpha_t$, then for all $n \geq 1$, states e_i, and structure descriptions $n_1, ..., n_t$,

$$P(X_{n+1} = e_i \mid n_1, ..., n_t) = \frac{n_i + \alpha_i}{n + \alpha}.$$

This truly beautiful result characterizes the predictive probabilities up to a finite sequence of positive constants $\alpha_1, \alpha_2, ..., \alpha_t$. Note Carnap's $c^\star$ measure of confirmation is a special case of the continuum, with $\alpha_i = 1$ for all i.

The assumption that all state descriptions have positive probability is needed to insure that the requisite conditional probabilities are well-defined. (In Carnap's terminology, the probability function is *regular*.) The restriction $t \geq 3$ is necessary because otherwise the sufficientness postulate would be vacuous. (One can recover the result in the case $t = 2$ by replacing the sufficientness postulate by the assumption that the predictive probabilities are linear in n_i; see, e.g., [Zabell, 1982].)

5.2 The de Finetti representation theorem

The assumption that arbitrarily long sequences satisfy the permutation postulate means their probabilities admit an integral representation of the type mentioned earlier in Section 3; this is the content of the celebrated *de Finetti representation theorem* [de Finetti, 1937]. Specifically, let Δ_t denote the set of probabilities on t elements:

$$\Delta_t := \{(p_1, ..., p_t) : p_j \geq 0, \sum_{j=1}^{t} p_j = 1\}.$$

De Finetti's theorem states that if $X_1, X_2, X_3, ...$ is an infinitely exchangeable sequence on t elements, then there exists a probability measure $d\mu$ on Δ_t, such that for every $n \geq 1$, if $n_1, ..., n_t$ are the frequency counts of $X_1, ..., X_n$, then

$$P(n_1, n_2, ..., n_t) = \int_{\Delta_t} \frac{n!}{n_1! n_2! ... n_t!} \, p_1^{n_1} p_2^{n_2} ... p_n^{n_t} \; d\mu(p_1, ..., p_t).$$

(Note that a single measure $d\mu$ simultaneously achieves this for all sample sizes n.)

There are a number of interesting foundational issues arising from this result. The integrand

$$\frac{n!}{n_1! n_2! ... n_t!} \, p_1^{n_1} p_2^{n_2} ... p_n^{n_t}$$

is a *multinomial probability*, and the theorem asserts that an exchangeable probability P can be represented as a integral mixture of multinomial probabilities. It is obvious that a multinomial probability and more generally any mixture of multinomials is exchangeable; the force of the theorem is that the converse holds:

every exchangeable probability is expressible as a mixture. There is no restriction placed on the mixing measure $d\mu$.

Many results in the literature of inductive inference are often easier to state, prove, or interpret in terms of such representations. For example, Johnson's theorem can be interpreted as telling us that when the sufficientness postulate is satisfied the averaging measure in the representation is a member of the classical *Dirichlet family* of prior distributions:

$$d\mu(p_1, ..., p_t) = \frac{\Gamma(\sum_{j=1}^t \alpha_j)}{\prod_{j=1}^t \Gamma(\alpha_j)} \prod_{j=1}^t p_j^{\alpha_j - 1} dp_1 ... dp_{t-1} \quad (\alpha_j > 0).$$

(Here Γ denotes the *gamma function*; if k is a positive integer, then $\Gamma(k) = (k-1)!$.)

The ability to characterize the predictive probabilities using Johnson's sufficientness postulate, however, means that in principle one can entirely pass over this interesting but more mathematically complex fact. As Johnson himself observed,

> I substitute, for the mathematician's use of Gamma Functions and α-multiple integrals, a comparatively simple piece of algebra, and thus deduce a formula similar to the mathematician's, except that, instead of for two, my theorem holds for α alternatives, primarily postulated as equiprobable. [Johnson, 1932, p. 418; Johnson's α corresponds to our t]

Why are rules of succession so important? Note the joint probability of a sequence of events can be built up from the corresponding sequence of conditional probabilities. For example: the joint probability

$$P(X_1 = e_1, X_2 = e_2, X_3 = e_3)$$

can be expressed as

$$P(X_1 = e_1) \cdot P(X_2 = e_2 \mid X_1 = e_1) \cdot P(X_3 = e_3 \mid X_1 = e_1, X_2 = e_2).$$

Thus one can express joint probabilities in terms of initial probabilities and rules of succession.

5.3 Interpretation of the Continuum

Let us consider a specific method in the continuum, say with parameters $\alpha_1, ..., \alpha_t$. Then one can write the rule of succession as

$$P(X_{n+1} = c_i \mid n_i) = \frac{n_i + \alpha_i}{n + \alpha} = \left(\frac{n}{n+\alpha}\right)\left[\frac{n_i}{n}\right] + \left(\frac{\alpha}{n+\alpha}\right)\left[\frac{\alpha_i}{\alpha}\right].$$

The two expressions in square brackets have obvious interpretations: the first, n_i/n is the *empirical frequency*, and represents the input of experience; the second,

α_i/α, is our initial or prior probability concerning the likelihood of seeing c_i (set $n_i = n = 0$ in the formula). The two terms in rounded brackets, $n/(n+\alpha)$ and $\alpha/(n+\alpha)$, sum to one and express the relative weight accorded to our observations versus our prior information. If α is small, then $n/(n+\alpha)$ is close to one, and the empirical frequencies n_i/n are accorded primacy; if α is large, then $n/(n+\alpha)$ is small, and the initial probabilties are accorded primacy.

Of course, "if α is large" must be understood relative to a *fixed* value of n; no matter how large α is, for a fixed value of α it is evident that

$$\lim_{n\to\infty} \frac{n}{n+\alpha} = 1,$$

reflecting the fact that no matter how large the initial weight assigned to our initial probabilities, these prior opinions are ultimately swamped by the overwhelming weight of empirical evidence.

5.4 History

The result itself has an interesting history. Johnson considered the special case when the function $f_i(n_i, n) = f(n_i, n)$; that is, it does not depend on the category or type i. In this case there is just one parameter, α, since $\alpha_i = \alpha/t$ for all i. Johnson did not publish his result in his own lifetime (shades of Bernoulli and Bayes!); he had planned a fourth volume of his *Logic*, but only completed drafts of three chapters of it at the time of his death. A (then very young) R. B. Braithwaite edited the chapters for publication, and they appeared as three separate articles in *Mind* in 1932 [Johnson, 1932]. (It is ironic that G. E. Moore, the editor of *Mind*, questioned the desirability of including a mathematical appendix giving the details of the proof in such a journal, but Braithwaite — fortunately — insisted.) Due to its posthumous character, the proof as published contained a few lacunae, and a desire to fill these led to [Zabell, 1982]. This paper shows that not only can the above-mentioned lacunae be filled, but that Johnson's method very naturally generalizes to cover the asymmetric case (when the predictive function $f_i(n_i, n)$ depends on i), the case $t = \infty$, and the case of finite exchangeable sequences that are not infinitely extendable.

Carnap followed much the same path as Johnson, initially considering the symmetric, category independent case, except that he *assumed* both the sufficientness postulate *and* the form of the predictive probabilities given in the theorem. It was only later that his collaborator John G. Kemeny was able to prove the equivalence of the two (assuming $t > 2$). Carnap subsequently extended these results, first to cover the case $t = 2$ [Carnap and Stegmüller, 1959]; and finally in Jeffrey (1980, Chapter 6) abandoned the assumption of symmetry between categories and derived the full result given above (see also [Kuipers, 1978]). The historical evolution is traced in [Schillp, 1963, pp. 74–75 and 979–980; Carnap and Jeffrey, 1971, pp. 1–4 and 223; Jeffrey, 1980, pp. 1–5 and 103–104].

6 CONFIRMATION OF UNIVERSAL GENERALIZATIONS

Suppose all n observations are of the same type; for example, that we are observing crows and thus far all have been black. In such situtations, it is natural to view our experience as evidence not just that most crows are black, but as confirming the "universal generalization" that *all* crows are black. This apparently natural expectation, however, leads to unexpected complexities.

6.1 Paradox feigned

This is due to an interesting property of the Johnson-Carnap continuum: (infinite) universal generalizations have zero probability! For example, having observed n black crows, it follows from k successive applications of the rule of succession that the probability the next k crows are also black is

$$P(X_{n+1} = X_{n+2} = ... = X_{n+k} = c_i \mid n_i = n) = \prod_{j=n}^{n+k-1} \frac{j+\alpha_i}{j+\alpha}.$$

It is not hard to see that this product tends to zero as k tends to infinity. It is a standard result that if $0 < a_n \leq 1 (n \geq 1)$ then the infinite product $\prod_{n\geq 1} a_n)$ diverges to zero if and only if the corresponding infinite series $\sum_{n\geq 1}(1 - a_n)$ diverges to infinity (see, e.g., [Knopp, 1947, pp. 218-221]). Because

$$\sum_{j=n}^{\infty} \frac{\alpha - \alpha_i}{j+\alpha}$$

diverges (it is essentially the harmonic series), one has

$$\lim_{k\to\infty} \prod_{j=n}^{n+k-1} \frac{j+\alpha_i}{j+\alpha} = 0.$$

This was viewed as a defect of Carnap's system by several critics, for example, [Barker, 1957, pp. 87-88; Ayer, 1972, pp. 37-38, 80-81]. But the phenomenon itself had been both noted and defended much earlier, by Augustus De Morgan [1838, p. 128] in the nineteenth century. ("No finite experience whatsoever can justify us in saying that the future shall coincide with the past in all time to come, or that there is any probability for such a conclusion"); and by C. D. Broad [1918] in a similar situation (the "finite rule of succession") in the twentieth. The obvious Bayesian response was advanced by Wrinch and Jeffreys [1919] a year after Broad wrote: one assigns non-zero initial probability to the generalization. As Edgeworth noted shortly after in his review of Keyens's *Treatise*, "pure induction avails not without some finite initial probability in favour of the generalisation, obtained from some other source than the instances examined" [Edgeworth 1922, p. 267].

But can one build such a "finite initial probability" into the Carnapian approach (that is, via axiomatic characterization)? In order to understand this, let us first consider the simplest case.

6.2 Paradox lost

It is possible to see what is going wrong in terms of the sufficientness postulate. Suppose there are three categories, 1, 2, and 3, and none of the observations thus far fall into the first. What can one say about

$$P(X_{2n+1} = c_1 \mid n_1, n_2, n_3)?$$

According to the sufficientness postulate, there is no difference between the three cases (*a*) $n_2 = 2n$, $n_3 = 0$, (*b*) $n_2 = 0$, $n_3 = 2n$, and (*c*) $n_2 = n_3 = n$. But from the point of universal generalizations there is an obvious difference: the first and second cases confirm different universal generalizations (which may have different initial probabilities), while the third case disconfirms both. Continua confirming universal generalizations must treat the cases differently.

Thus it is necessary to relax the sufficientness postulate, at least in the case when $n_i = n$ for some i. This diagnosis suggests a simple remedy. Suppose one modifies the sufficientness postulate so that the "representative functions" $f_i(n_1, ..., n_t)$ (to use yet another terminology sometimes employed) are assumed to be functions of n_i and n *unless* $n_i = 0$ and $n_j = n$ for some $j \neq i$. Then it can be shown (see, e.g., [Zabell, 1996]) that as long the observations are exclusively of one type, the representative function consists of two parts: a term corresponding to the posterior probability that future observations will continue to be of this type (the "universal generalization"), and a Johnson-Carnap term; and this continues to be the case as long as all observations are of a single type. If, however, at any stage a second type is observed, then the representative function reverts to a pure Johnson-Carnap form.

So this was a tempest in a teapot: this criticism of the continuum was easily answered even at the time it was initially made. In hindsight the reason Johnson's postulate gives rise to the problem is apparent, the minimal change to the postulate necessary to remedy the problem results in an expanded continuum confirming precisely the desired universal generalizations (and no others), and this can be demonstrated by a straightfoward modification of Johnson's original proof (for further discussion and references, see [Zabell, 1996]).

But in fact much more is true: such an extension of the original Carnap continuum is merely a special case of a much richer class of extensions due to Hintikka, Niiniluoto, and Kuipers.

6.3 Hintikka-Niiniluoto systems

In order to appreciate Hintikka's contribution, consider first the category symmetric case. Let $T_n(X_1, X_2, ..., X_n)$ denote the number of distinct types or species observed in the sample. In the continuum discussed in the previous subsection the predictive probabilities now depend not just on n_i and n, but also on T_n, the number of instantiated categories. Specifically: is $T_n = 1$ or is $T_n > 1$? Thus put, this suggests a natural generalization: let the predictive probabilities be *any* function of n_i, n, and T_n. The result is a very attractive extension of the Carnap continuum.

In brief, if the predictive probabilities depend on T_n, then in general they arise from mixtures of Johnson-Carnap continua concentrated on subsets of the possible types. Thus, given three categories a, b, c, the probabilities can be concentrated on a or b or c (universal generalizations), or Johnson-Carnap continua corresponding to the three pairs $(a, b), (a, c), (b, c)$, or a Johnson-Carnap continuum on all three. In retrospect, this is of course quite natural. If only two of the three possibilities are observed in a long sequence of observations (say a and b), then (in addition to giving us information about the relative frequency of a and b) this tentatively confirms the initial hypothesis that *only* a's and b's will occur. In the more general category asymmetric case, the initial probabilities for the six different generalizations (a, b, c, ab, ac, and bc) can differ, and the predictive probabilities are postulated to be functions of n_i, n, and the observed *constituent*: that is, the specific set of categories observed. (Thus in our example it is not enough to tell one that $T_n = 2$, but *which* two categories or species have been observed.)

This beautiful circle of results originates with Hintikka [1966], and was later extended by Hintikka and Niiniluoto [1979]. The monograph by Kuipers [1978] gives an outstanding survey and synthesis of this work, including discussion of Kuipers's own contributions; for a recent summary and evaluation, see Niiniluoto [2009].

6.4 Attribute symmetry

Both the original Johnson-Carnap continuum and its Hintikka-Niiniluoto-Kuipers generalizations are of great interest, but share a common weakness. *If* what one is trying to do is to capture precisely the notion of a category-symmetric state of knowledge – *no more and no less* — then the one and only constraint is that the resulting probabilities be invariant under permutation of the categories. Carnap referred to such invariance as *attribute symmetry*. If one writes an n-long sequence in compact form as

$$X : \{1, ..., n\} \to \{1, ..., t\},$$

and P is a probability on the possible sequences X, then exchangeability requires P to be invariant under permutations of $\{1, ..., n\}$ and attribute symmetry requires P to be invariant under permutations of $\{1, ..., t\}$.

Suppose one adds attribute symmetry to exchangeability as a restriction on P. The resulting class of probability functions is still infinite dimensional; see Zabell [1982, p. 1097, 1992; pp. 216–217]. At first sight this seems surprising: if our knowledge is category symmetric, surely the sufficientness postulate should hold. But it is not hard to construct counterexamples. For example, suppose we have a die and know one face is twice as likely to come up as another, but not which face. Then there are six hypotheses H_j: for $1 \leq j \leq 6$, $H_j : p_j = 2/7, p_k = 1/7, k \neq j$; and the six H_j are judged equiprobable. Consider the following two possible frequency vectors that could occur in a sample of size $n = 70$:

$$\mathbf{n}_1 = (20, 10, 10, 10, 10, 10), \qquad \mathbf{n}_2 = (20, 30, 5, 5, 5, 5).$$

Obviously $\mathbf{n}_1$ supports H_1 over H_2; and $\mathbf{n}_2$ supports H_2 over H_1, even though, if the sufficientness postulate held, the predictive probabilities for seeing a one on the next trial should be the same in each case.

So there exist natural category symmetric epistemic states in which the sufficientness postulate fails. In general, if there is attribute symmetry the sufficient statistics are the *frequencies of the frequencies* (denoted a_r): for each $r, 0 \leq r \leq t$, a_r is the number of categories j such that $n_j = r$. The recognition that even in these cases the entire list of frequencies n_i may contain relevant information concerning the individual categories via the a_r appears to go back to Turing; see [Good, 1965, Chapter 8].

Thus even assuming both exchangeability and attribute symmetry admits a rich family of possible probabilities; and it might be thought this would limit their utility. But even exchangeability by itself has many interesting qualitative consequences. The next section illustrates one of these.

7 INSTANTIAL RELEVANCE

One important desideratum of a candidate for confirmation is *instantial relevance*: if a particular type is observed, then it is more likely that such a type will be observed in the future. In its simplest form, this is the requirement that if $i < j$, then

$$P(X_j = 1 \mid X_i = 1) \geq P(X_j = 1)$$

(the X_k denoting indicators that take on the values 0 or 1).

It is not hard to see that exchangeability alone does not insure instantial relevance. Suppose, for example, one draws balls at random from an urn initially having three red balls and two black balls. If the sampling is *without replacement*, then the probability of selecting a red ball is initially 3/5, but the probability of selecting a second red ball, *given the first is red*, is 1/2.

In the past there was a small cottage industry devoted to investigating the precise circumstances under which the principle of instantial relevance does or does not hold for a sequence of observations. If the observations in question can be imbedded in an *infinitely* exchangeable sequence (that is, into an infinite sequence $X_1, X_2, ...$, any finite segment $X_1, ..., X_n$ of which is exchangeable), then instantial relevance does hold. After the power of the de Finetti representation theorem was appreciated, very simple proofs of this were discovered (see, e.g., [Carnap and Jeffrey, 1971, Chapters 4 and 5]).

There are also simple ways of seeing this without using the representation theorem. For example, the principle of instantial relevance is equivalent to the assertion that the observations are *nonnegatively correlated*. If $X_1, X_2, ..., X_n$ is an exchangeable sequence of random variables, then an elementary argument shows that the correlation coefficient $\rho = \rho(X_i, X_j)$ satisfies the simple inequality

$$\rho \geq -\frac{1}{n-1}.$$

This is because (using both the formula for the variance of a sum and the exchangeability of the sequence) if $\sigma^2 = Var[X_i]$, one has

$$0 \leq Var[X_1 + ... + X_n] = n\sigma^2 + n\,(n-1)\,\rho\sigma^2.$$

Thus, if the sequence can be indefinitely extended (so that one can pass to the limit $n \to \infty$), it follows that $\rho \geq 0$. The case $\rho = 0$ then corresponds to the case of independence (the past conveys no information about the future, inductive inference is impossible); and the case $\rho > 0$ corresponds to inductive inference and positive instantial relevance.

8 FINITE EXCHANGEABILITY

In the end, infinite sequences are really just fictions, so we would rather not incorporate them into our *Weltanschauung* in an essential way. In this section we take a closer look at this question.

8.1 Extendability

The de Finetti representation only holds for an *infinite* sequences; it is easy to construct counterexamples otherwise. Consider, for example, the exchangeable assignment

$$P(RB) = P(BR) = \frac{1}{2}; \qquad P(RR) = P(BB) = 0.$$

This corresponds to sampling *without* replacement from an urn containing one red ball (R) and one black ball (B). This exchangeable probability assignment on ordered pairs cannot be extended to one on ordered triples. To see this, suppose otherwise. Then

$$P(RBR) + P(RBB) = P(RB) = \frac{1}{2},$$

so either $P(RBR) > 0$ or $P(RBB) > 0$ (or both). Suppose without loss of generality that $P(RBR) > 0$. Then

$$P(RR) \geq P(RRB) = P(RBR) > 0$$

(the first inequality follows because probabilities are subadditive, that is, if $A \subseteq B$, then $P(A) \leq P(B)$; the equality because P is by assumption exchangeable). But this is impossible, since $P(RR) = 0$. (It is not hard to see this is typical: sampling without replacement from a finite population results in an exchangeable probability assignment that cannot be extended.)

In general, if $X_1, X_2, ..., X_n$ is an exchangeable sequence, then it may or may not be possible to extend it to a longer exchangeable sequence $X_1, X_2, ..., X_n, ..., X_{n+r}$, $r \geq 1$. If it is possible to do so for *every* $r \geq 1$, then we can think of $X_1, X_2, ..., X_n$ as the initial sequence of an *infinitely* exchangeable sequence $X_1, X_2, X_3, ...$ (thanks to the *Kolmogorov existence theorem*). Thus the de Finetti representation theorem applies, the infinite sequence can be represented as a mixture of iid (*independent and identically distributed*) sequences, and hence *a fortiori* the initial segment of length n can be so represented.

On the other hand, if a finite exchangeable sequence of length n has a representation as a mixture of iid sequences, it is immediate that it is infinitely extendable. Thus:

> *A finite exchangeable sequence is infinitely extendable if and only if it is representable as a mixture of iid sequences.*

To summarize: in general a finite exchangeable sequence may or may not be extendable. Carnap alludes to this fact when he reports that while at the Institute for Advanced Studies in 1952–1953, he and his collaborator John Kemeny

> had talks with L. J. Savage. Among other things, Savage showed them that the use of a language L_N with a finite number of individuals is not advisable, because a symmetric M-function in L_N cannot always be extended to an M-function in a language with a greater number of individuals. [Carnap and Jeffrey, 1971, p. 3]

Note the curious phrase "not advisable". It is unclear why Savage thought this (if indeed he did): recall sampling without replacement from a finite population results in a perfectly respectable exchangeable assignment even though it cannot be extended. More generally think of any population which is naturally finite in extent, and to which we wish to extrapolate on the basis of a partial sample from it. (For example, think of a limited edition of a book, and whether or not such books are defective.) The phenomenon of non-extendability is no sense pathological.

Or course there is a price to pay: the loss of the de Finetti representation. Or is there?

8.2 The finite representation theorem

Given a set of counts $\mathbf{n} = (n_1, ..., n_t)$, imagine an urn containing n_j balls of each type, and suppose one successively draws out "at random" without replacement each ball in the urn ("at random" meaning that all possible sequences are judged equally likely). There are a total of $(n_1 + \quad + n_t)!/(n_1!...n_t!)$ such sequences; the exchangeable probability assignment $H_\mathbf{n}$ giving each of these equal probability is called the *hypergeometric distribution*. If, more generally, $X_1, , X_n$ is any exchangeable sequence whatsoever, and $P(\mathbf{n})$ the corresponding probability assignment on the set of counts $\mathbf{n}$, then the overall probability assignment P on the set of sequences is a mixture of the hypergeometric probabilities $H_\mathbf{n}$ using the weights

$P(\mathbf{n})$; compactly this can be expressed as

$$P = \sum_{\mathbf{n}} P(\mathbf{n}) H_{\mathbf{n}}.$$

This result is the *finite de Finetti representation theorem.* It is basically just the so-called "theorem of total probability" in disguise. It tells us that the structure of the generic finite exchangeable sequence is really quite simple. If the sequence is N long, and the outcomes can be of t different types, then you can think of it as a sequence of draws from an urn with N balls, each of which can be one of the t types, but the distribution of types of among the N balls (the $\mathbf{n}$) is unknown. *If* (as the Spartans would say), you knew the distribution of types, then your probability assignment would be the appropriate hypergeometric distribution. But since you don't, you assign a prior distribution to $\mathbf{n}$ and then average.

Although the finite representation theorem is not quite as well known (or appreciated) as its big brother, the representation theorem for an infinite exchangeable sequence, it would be a serious mistake to underestimate it. To begin, thanks to the representation, there is a drastic reduction in the number of independent probabilities to be specified; in the case of tossing a coin 10 times, for example, from $2^{10} - 1 = 1023$ to 11.

But there are also important conceptual and philosophical advantages to thinking in terms of the finite representation theorem.

8.3 The finite rule of succession

The classical rule of succession, that if in n trials there are k successes, then the probability of a success on the next trial is $(k{+}1)/(n{+}2)$, assumes you are sampling from an infinite population (see [Laplace, 1774]). (Strictly speaking the last makes no sense, but it can be viewed as a shorthand for either sampling with replacement (so that the population remains unaltered by the sampling) or as passing to the limit in the case case of sampling from a finite population.) In particular, if all n are of the same type, then the probability that the next is also of this type is $(n + 1)/(n + 2)$.

But it is clear that the basic relevant question is a different one: the probability if you are sampling *without replacement* from *a finite* population. This question was first asked and answered by Prevost and L'Huilier [1799]. To answer the question, of course, one must make some assumption regarding the composition of the urn (that is, adopt some set of prior probabilities regarding the different possible urn compositions). The natural assumption, parallel to the Bayes-Laplace analysis, is to assume all possible vectors of counts are equally likely. Doing this, Prevost and L'Huilier were able to first derive the posterior probabilities for the different urn constitutions of the urn; and then from this derive the rule of succession as a consequence, the final result being that (given p successes out of m to date) the probability of a success on the next trial is $(p + 1)/(m + 2)$, *exactly the same answer as the classical Laplace rule of succession*!

This result was subsequently independently rediscovered several times over the next century and a quarter, the last being by C. D. Broad in 1918, when it finally gained some traction in philosophical circles (see generally [Zabell, 1988]). The brute force mathematical derivation of this particular rule of succession requires the evaluation of a tricky combinatorial sum; and its history of successive rediscovery is a phenomenon that is sometimes seen in the mathematical literature when a result is interesting enough (so that it repeatedly attracts attention), hard enough (so that it is deemed worthy of publication), and obscure or technical enough (so that it is then subsequently easily forgotten or overlooked).

But our point here is that this striking coincidence between the finite and infinite rules of succession, which, when viewed through the prism of the combinatorial legerdemain required to evaluate the necessary sum, appears to be a minor miracle, *is in fact obvious when thought of in terms of the finite representation theorem.*

For consider. Suppose $X_1, X_2, ...$ is an infinite exchangeable sequence of 0*s* and 1*s* having mixing measure $dQ(p) = dp$ in the de Finetti representation (that is, the Bayes-Laplace process). If $S_n = X_1 + ... + X_n$ denotes the number of 1*s* in n trials, then, as noted earlier,

$$P(S_n = k) = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} dp = \frac{1}{n+1}.$$

Now consider the initial segment $X_1, X_2, .., X_n$ by itself. This is a finite exchangeable sequence, and so has a finite representation in terms of some mixture of hypergeometric probabilities. But the mixing measure for the finite representation in the dichotomous case is $P(S_n = k)$, which is, as just noted, $1/(n+1)$, the Prevost-L'Huilier prior (or, as Jack Good might put it, the Prevost-L'Huilier-Terrot-Todhunter-Ostrogradskii-Broad prior).

But the finite representation uniquely determines the stochastic structure of a finite exchangeable sequence; thus an n-long Prevost-L'Huiler sequence is stochastically *identically* to the initial, n-long segment of the Bayes-Laplace process, and therefore the two coincide in *all* respects, including (but not limited to) their rules of succession. No tricky sums!

Viewed from the perspective of the philosophical foundations of inductive inference the finite rule of succession is important for two reasons *vis-a-vis* the classical Laplacean analysis:

1. It eliminates a variety of possible concerns about the occurrence of the infinite in the Laplacean analysis (e.g., [Kneale, 1949, p. 205]): that is, attention is focused on a finite segment of trials, rather than a hypothetical infinite sequence or population.

2. The frequency, propensity, or objective chance p that appears in the integral is replaced by the fraction of successes in a finite population; thus a purely personalist or subjective analysis becomes possible and objections to "probabilities of probabilities" or "unknown probabilities" (e.g., [Keynes, 1921, pp. 372 –75]) are eliminated.

8.4 The finite continuum of inductive methods

As one final example of both the utility and interest of considering finite exchangeable sequences, we note in passing that Johnson's derivation of the continuum of inductive methods carries over immediately to the finite case, the chief element of novelty being that now the α parameters in the rule of succession can be *negative* (since, for example, when sampling without replacement from an urn, the more balls of a given color one sees, the *less* likely it becomes to see other balls of the same color); see [Zabell, 1982].

8.5 The proper role of the infinite

Aristotle (*Physics 3.6*, see, e.g., [Heath, 1949, pp. 102–113]) distinguishes between the actual infinite and the potential infinite, a useful distinction to keep in mind when thinking about the use of the infinite in probability. One might summarize Aristotle as saying that the use of the infinite is only appropriate in its potential rather than actual sense. Let us apply this to the case of probability: theories that depend in an essential way on the actual infinite are fatally flawed. Consider von Mises's frequency theory. In any theory of physical probability, if $0 < p < 1$ is the probability of an outcome in a sequence of independent trials, then any finite frequency k in n trials has a positive probability. Thus any observed value of k is consistent with any possible value of p. In von Mises's theory in order to achieve this consistency of any p with any k, it is essential that p be an *infinite* limiting frequency. But, being infinite in nature, p is unobservable, hence metaphysical (in the pejorative sense); see, e.g., [Jeffrey, 1977].

But, one might object, doesn't the infinite representation theorem also suffer from this defect, since it holds just for infinitely exchangeable sequences (rather than finitely exchangeable sequences, the only things we really see)? The answer is no, if one correctly understands it from both a mathematical and a philosophical standpoint.

Mathematical interpretation of the representation theorem

In applied mathematics one frequently uses infinite limit theorems as approximations to the large but finite. That is, the sequence, although of course necessarily finite, is viewed as *effectively* unlimited in length. (So, for example, in tossing a coin, there is no practical limit to how many times we can toss it, although it will certainly wear down after many googles of tosses.)

But the applied mathematician must also have some idea of when to use a limit theorem as an approximation and when not. This is the reason the central limit theorem (CLT) is is of practical use, but the law of the iterated logarithm (LIL) is not: the CLT provides an excellent approximation to sums of random variables for surprisingly small sample sizes; the LIL only for surprisingly large.

What this ultimately means is that what the applied mathematician needs is either a generous fund of experience or a more informative mathematical result:

not just the limiting value but the *rate* of convergence to that limit. Happily such a result is available for the de Finetti representation theorem, thanks to Persi Diaconis and David Freedman [1980a].

First some notation: if S is a set, let S^n denote its n-fold Cartesian product ($n \leq \infty$). If p is a probability on S, let p^n denote the corresponding n-fold product probability on S^n (corresponding to an n-long, p-iid sequence). If P is a probability on S^n, then P_k denotes its restriction to S^k, $k \leq n$. If Θ parametrizes the set of probabilities on S and μ is a a probability on Θ (to be thought of as a mixing measure), let $P_{\mu n}$ denote the resulting exchangeable probability on S^n; that is

$$P_{\mu n} = \int_{\Theta} p_\theta^n \, d\mu(\theta).$$

Finally, if P and Q are probabilities on S^n, let

$$||P - Q|| = \max_{A \subset S^n} |P(A) - Q(A)|$$

denote the *variation distance* between P and Q.

Then one has the following result: *Suppose S is a finite set of cardinality t and P is an exchangeable probability on S^n. Then there exists a probability μ on the Borel sets of Θ and a constant c such that*

$$||P_k - P_{\mu k}|| = \left\| P_k - \int p_\theta^n \, d\mu(\theta) \right\| \leq \frac{2tk}{n} \qquad for\ all\ k \leq n.$$

This beautiful result has a number of interesting consequences. First, it makes precise the interrelationship between extendability and the existence of an integral representation. Given an exchangeable sequence of length k, if the sequence is extendable to a longer sequence of length n, then it can be approximated by an integral mixture to order k/n in variation distance. The more the sequence can be extended, the more it looks like an integral mixture. Thus it is not surprising (and Diaconis and Freedman in fact use the above theorem to prove) that a sequence which can be extended indefinitely (equivalently, is the initial segment of an infinitely exchangeable sequence) has an integral representation.

But the theorem also tells us how to think about the application of the representation theorem. Given a sequence that is the initial segment of a “potentially infinite” sequence (that is, unbounded in any practical sense), thinking of it as an integral mixture is a reasonable approximate procedure (in just the same way as summarizing a population of heights in terms of a normal distribution is a reasonable approximation to an ultimately discrete underlying reality). For a very readable discussion of this topic, see [Diaconis, 1977].

Philosophical interpretation of the representation theorem

From this perspective the representation is a tool used for mathematical approximation. The “parameter” p is a purely mathematical object, not a physical quantity. This was in fact de Finetti’s view: “it is possible... and to my mind preferable, to stick to the firm and unexceptionable interpretation that the limit distribution is merely the asymptotic expression of frequencies in a large, but finite, number of trials" [de Finetti, 1972, p. 216].

De Finetti was a finitist who rejected the use of countable additivity in probability as lacking a philosophical justification. (It is not a consequence of the usual Dutch book argument.) In particular, de Finetti's statement and proof of the representation theorem uses only finitely additive probability. See Cifarelli and Regazzini [1996] for an outstanding discussion of the role of the infinite in de Finetti's papers.

9 THE FIRST INDUCTION THEOREM

There is a very interesting result, which Good [1975, p. 62] terms the *first induction theorem*. Its interest is that it makes no reference at all to exchangeability, and yet it provides an account of enumerative induction, in that it tells us that confirming instances (in a sense to be made precise in a moment) increase the probability of other potential instances. To be precise, if $P(H) > 0$ and $P(E_j|H) = 1, j \geq 1$ (the E_j are "implications" of H), then (E_1E_2 denoting the conjunction of E_1 and E_2, and so on),

$$\lim_{n\to\infty} P(E_{n+1}E_{n+2}...E_{n+m}\big|E_1E_2...E_n) = 1$$

uniformly in m. The proof (due to [Huzurbazar, 1955]) is at once simple and elegant. Just note that for any $n \geq 1$, one has $P(E_1...E_n\big|H) = 1$, hence

$$P(E_1...E_n) \geq P(E_1...E_nE_{n+1}) \geq P(E_1...E_nE_{n+1}H) = P(H) > 0.$$

It follows that $u_n = P(E_1...E_n)$ is a decreasing sequence bounded from below by a positive number, and therefore has a positive limit. Thus

$$\lim_{n\to\infty} P(E_{n+1}E_{n+2}...E_{n+m}\big|E_1E_2...E_n) = \lim_{n\to\infty} \frac{u_{n+m}}{u_n} = 1;$$

and it is apparent that the convergence is uniform in m.

The result is not so surprising for sampling from a finite population, but for a potentially infinite sequence is at first startling. It tells us that observing a sufficiently long sequence of confirming instances makes *any* further finite sequence, *no matter how long*, as close to one as desired. Good [1975, p. 62] says "the kudology is difficult", but cites both Keynes [1921, Chapter 20] and Wrinch and Jeffreys [1921]; see also [Jeffreys, 1961, pp. 43–44].

10 ANALOGY

Simple enumeration is an important form of inductive inference but there are also others, based on analogy. Carnap distinguished between two forms of analogy:

analogy by proximity and *analogy by similarity*; that is, proximity in time (or sequence number) and similarity of attribute.

In the case of inductive analogy, Carnap wished to generalize his results, allowing for the possibility that the inductive strength of P varies depending on some measure of "closeness" of either time or attribute. In the case of attributes this required the specification of a "distance" on the attribute set; in the case of time such a metric is of course already present. But Carnap only obtained only partial results in this case (see [Carnap and Jeffrey, 1971, p. 1; Jeffrey, 1980, Chapter 6, Sections 16–18]).

De Finetti and his successors were more successful. De Finetti formulated early on a concept of *partial exchangeability* [de Finetti, 1938], differing forms of partial exchangeability corresponding to differing forms of analogy. He viewed matters in effect as a spectrum of possibilities; exchangeability representing one extreme, a limiting case of "absolute" analogy. At the other extreme all one has is Bayes's theorem, $P(E|A) = P(AE)/P(A)$; absent "particular hypotheses concerning the influence of A on E", nothing further can be said, "no determinate conclusion can be deduced". The challenge was to find "other cases ... more general but still tractable". For an English translation of de Finetti's paper, see [Jeffrey, 1980, Chapter 9]. Diaconis and Freedman [Jeffrey, 2004, pp. 82–97] provides a very readable introduction to de Finetti's ideas here.

10.1 Markov exchangeability

One example of building analogy by proximity into a probability function is the concept of *Markov exchangeability* (describing a form of analogy in time). Suppose $X_0, X_1, ...$ is an infinite sequence of random outcomes, each taking values in the set $S = \{c_1, ..., c_t\}$. For each $n \geq 1$, consider the statistics X_0 (the initial state of the chain) and the *transition counts* n_{ij} recording the number of transitions from c_i to c_j in the sequence up to X_n. (That is, the number of times $k, 0 \leq k \leq n-1$, such that $X_k = c_i$ and $X_{k+1} = c_j$.) If for all $n \geq 1$, all sequences $X_0, ..., X_n$ starting out in the same initial state x_0 and having the same transition counts n_{ij} have the same probability, then the sequence is said to be *Markov exchangeable.*

Suppose further that the sequence is *recurrent*: the probability is 1 that $X_n = X_0$ for infinitely many n. (That is, the sequence returns to the initial state infinitely often.) There is, it turns out, a de Finetti type representation theorem for the stochastic structure (probability law) of such sequences: they are precisely the mixtures of Markov chains, just as ordinary exchangeable sequences are mixtures of binomial or multinomial outcomes [Diaconis and Freedman, 1980b]. Furthermore there is also a Johnson-Carnap type rule of succession [Zabell, 1995].

Of course one might ask why Markov exchangeability is a natural assumption to make. Diaconis and Freedman [Jeffrey, 2004, p. 97] put it well: "If someone ... had never heard of Markov chains it seems unlikely that they would hit on the appropriate notion of partial exchangeability. The notion of symmetry seems strange at first ... A feeling of naturalness only appears after experience and reflection." For further discussion of Markov exchangeability and its relation to inductive logic, see [Skyrms, 1991].

10.2 Analogy by similarity

Given the tentative and limited nature of Carnap's attempt's to formulate an inductive logic that incorporated analogy by similarity, this stood as an obvious challenge and since Carnap's death there have been a number of attempts in this direction; see, e.g., [Romeijn, 2006] and the references there to earlier literature. Skyrms [1993; 1996] suggests using what he terms "hyperCarnapian" systems: finite mixtures of Dirichlet priors. He argues (p. 331): "In a certain sense, this is the only solution to Carnap's problem. ... HyperCarnapian inductive methods are the general solution to Carnap's problem of analogy by similarity".

But what if the outcomes are continuous in nature? In order to discuss this, it will be necessary to first revisit the definition of exchangeability.

10.3 The general definition of exchangeability

Consider first the general definition of exchangeability. A probability P on the space of sequences $x_1, x_2, ..., x_n$ of real numbers (that is, on $\mathbf{R}^n$) is said to be (finitely) *exchangeable* if it is invariant under all permutations σ of the index set $\{1, ..., n\}$; a probability P on the space of infinite sequences $x_1, x_2, ...$ (that is, on $\mathbf{R}^\infty$) is said to be *infinitely exchangeable* if its restriction P_n to finite sequences $x_1, x_2, ..., x_n$ is exchangeable for each $n \geq 1$. There is a sweeping generalization of the de Finetti representation theorem that characterizes such probabilities.

Some notation, briefly. Let $\{P_\theta : \theta \in \Theta\}$ denote the set of independent and identically distributed (iid) probabilities on infinite sequences. (That is, if p_θ is a probability measure on $\mathbf{R}$, then $P_\theta = (p_\theta)^\infty$ is the corresponding product measure on $\mathbf{R}^\infty$. Here θ is just an index for the probabilities on the real line. Certain measure-theoretic niceties are being swept under the carpet at this point to simplify the exposition.)

Now suppose that P is an infinitely exchangeable probability on infinite sequences. Then there exists a unique probability μ on Θ such that

$$P = \int_\Theta P_\theta \, d\mu(\theta).$$

That is, every exchangeable P on *infinite* sequences can be represented as a mixture of independent and identically distributed probabilities. (It is clear that every mixture of iid sequences is exchangeable; it is the point of the representation theorem that conversely *every* infinitely exchangeable probability arises thus. Aldous [1986] contains an outstanding survey of this and other generalizations of the original de Finetti theorem.)

Thus, in order to arrive at P, it suffices to specify μ. Unfortunately, Θ is an uncountably infinite set, and the representation usefully reduces the dimensionality of the problem of determining P only if one is able to exploit a difference in infinite cardinals!

10.4 The pragmatic Bayesian approach

In practical Bayesian statistics one sometimes proceeds as follows. Based on the background, training, and experience of the statistician, it is judged that the underlying but unknown distribution p_θ of a population of numbers is a member of some particular parametric family (for example, normal, exponential, geometric, or Poisson) and it is the task of the statistician to estimate the unknown parameter θ. The parameter space Θ is now finite dimensional, often one dimensional.

The mathematical model for a sample from such a population is an iid sequence of random variables $X_1, X_2, X_3, ...$, each X_j having distribution p_θ, so that $X_1, X_2, X_3, ...$ has distribution $P_\theta = (p_\theta)^\infty$. Being a Bayesian, the statistician assigns a "prior" or initial probability to Θ; the average over Θ using $d\mu$ then specifies a probability P as in the displayed formula above. Given a "random sample" (iid sequence) $X_1, ..., X_n$ from the population, the statistician then computes the "posterior" or final probability

$$P(\theta|X_1, ..., X_n)$$

using Bayes's theorem.

In general, the larger the sample, the more concentrated the posterior distribution is about some value of the parameter. For example, if the density of p_θ is

$$p_\theta(x) = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{(x-\theta)^2}{2}\right], \qquad -\infty < x < \infty,$$

(that is, normal, standard deviation one, unknown mean θ), then (except for certain "over-opinionated" priors) the posterior distribution for θ will be concentrated about $\bar{X}_n$, the sample mean for the random sample $X_1, ..., X_n$.

It is apparent that this procedure in fact captures precisely the form of analogical reasoning that Carnap had in mind. That is, if the sample mean is $\bar{X}_n = x$, then the resulting posterior distribution expresses support for the belief that the next observation will be in the vicinity of x, the strength of the evidence for different values y decreasing as the distance of y from x increases.

"But", the Carnapian may object, "this is an enterprise entirely different from the one Carnap envisaged! There is no logical justification proffered for the choice of the parametric family p_θ, or the choice of the prior $d\mu$"!! True, but how might such a justification–if it existed–proceed?

Consider the multinomial case in the continuum of inductive methods. There the de Finetti representation theorem tells us that the most general exchangeable sequence is a mixture of multinomial probabiities. The elegance of the Johnson-Carnap approach is that it replaces the essentially arbitrary, albeit mathematically convenient, *quantitative* assumption of the practicing Bayesian statistician that the prior is a member of a specific low-dimensional family (the Dirichlet priors on Δ_{t-1}) by the purely qualitative sufficientness postulate. That is, based on information received one might well arrive at the purely qualitative judgment that the probability that the next observation will be of a certain type should depend only on the number of that type already observed and the total number of observations to date. This is certainly a more principled approach to the problem of assigning a prior, in stark contrast to assuming the prior is Dirichlet purely for reasons of mathematical convenience.

Framed in this way, the form of a principled Bayesian approach to the more general problem (of deciding on priors for other parametric families) is also clear. Can one find, at least for the most common parametric families in statistics, a natural *qualitative* assumption on a sequence of observations *in addition to exchangeability* that implies the sequence is in fact not just an arbitrary mixture of iid probabilities, but a mixture of distributions strictly within the given parametric family? For example, what would be an analog of the sufficientness postulate ensuring that an exchangeable sequence is a mixture of normal, or exponential, or geometric, or Poisson distributions?

10.5 Group invariance and sufficient statistics

Thanks to some very deep and hard mathematics on the part of David Freedman, Persi Diaconis, Phil Dawid, and others, one can in fact answer this question for many of the most common statistical families. Here are some examples, followed by a brief summary of the currently known state of the theory.

Let $\phi_{\mu,\sigma^2}(x)$ denote the density of the normal distribution with mean μ and variance σ^2; that is,

$$\phi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

If a random variable X has such a distribution, then this is denoted $X \sim N(\mu, \sigma^2)$. The first example, characterizing exchangeable sequences that are a mixture of $N(0, \sigma^2)$, is admittedly not the most interesting from a statistical standpoint, but it provides a simple illustration of the type of results the theory provides.

EXAMPLE 1. An infinite sequence of random variables $X_1, X_2, X_3, ...$ is said to be *orthogonally invariant* if for every $n \geq 1$, the sequence $X_1, ..., X_n$ is invariant under all orthogonal transformations of $\mathbf{R}^n$. (An orthogonal transformation is a linear map that preserves distances. It can be thought as an n-dimensional rotation.)

Schoenberg's theorem tells us that every orthogonally invariant infinite sequence of random variables is a mixture of $N(0, \sigma^2)$ iid random variables. (Note that a coordinate permutation is a very special kind of orthogonal transformation; thus orthogonal invariance entails exchangeability and is much more restrictive.) In terms of the de Finetti representation, if P is the distribution of the orthogonally invariant sequence $X_1, X_2, ...,$ and P_σ the distribution of an iid sequence of $N(0, \sigma^2)$ random variables, then there exists a probability measure Q on $[0, \infty)$ such that

$$P = \int_0^\infty P_\sigma \, dQ(\sigma).$$

There is an equivalent formulation of Schoenberg's theorem in terms of sufficient statistics. Consider the statistic

$$T_n = \sqrt{X_1^2 + ... + X_n^2}.$$

Then the property of orthogonal invariance is equivalent to the property that, for each $n \geq 1$, conditional on T_n the distribution of $X_1, .., X_n$ is uniform on the $n-1$-sphere of radius T_n. Furthermore, the limit $T = \lim_{n\to\infty} T_n/\sqrt{n}$ exists almost surely and $P(T \leq \sigma) = Q([-\infty, \sigma))$; that is, the mixing measure Q is the distribution of the limit T.

This has (accepting for the moment that one is willing to talk about infinite sequences of random variables, about which more later), a striking consequence. The statistic $T_n/\sqrt{n}$ is the standard sample estimate of the standard deviation σ. Thus one has a natural interpretation of both the Q and the σ appearing in the de Finetti representation. Far from being merely mathematical objects in the representation theorem, they acquire a significance of their own. The "parameter" (σ) emerges as the limit of the sample standard deviation (note one is certain of the existence of the limit but not its value); Q is our degree of belief regarding the unknown parameter (our uncertainty regarding the value of σ); and conditional on the limit being σ the sequence is iid $N(0, \sigma^2)$.

Thus one has a complete explication of the role of parameters, parametric families, and priors used by the pragmatic Bayesian statistician in this case. The particular parametric family arises from the particular strengthening of exchangeability (here orthogonal invariance) reflecting the knowledge of the statistician in this case. (If he doesn't subscribe to orthogonal invariance, he shouldn't be using a mixture of mean zero normals!) The single parameter σ is interpreted as the large sample limit of the sample standard deviation; and the mixing measure Q reflects our degree of belief as to the value of this limit. Very neat!

EXAMPLE 2. Suppose P is a mixture of iid $N(\mu, \sigma^2)$ normals. Then it is easy to see that P is invariant under transformations that are orthogonal *and* preserves the line $L_n : x_1 = x_2 = ... = x_n$. *Dawid's theorem* states that this is in fact the necessary and sufficient condition for P to be such a mixture. In this case there are two sufficient statistics:

$$U_n = X_1 + ... + X_n, \qquad V_n = \sqrt{X_1^2 + ... + X_n^2};$$

and the symmetry assumption is equivalent to the property that, conditional on U_n, V_n, the distribution of $X_1, ..., X_n$ is uniform on the resulting $(n-2)$-sphere. Furthermore, one has that the limits

$$U = \lim_{n\to\infty} U_n/n, \qquad V = \lim_{n\to\infty} V_n/\sqrt{n}$$

exist almost surely and generate the mixing measure on the two-dimensional parameter space $\mathbf{R} \times [0, \infty)$.

Characterizations of this kind are known for a number of standard statistical distributions. Many of these form "exponential families"; Diaconis and Ylvisaker [1980] characterize the conjugate priors for such families in terms of the linearity of their posterior expectations. In other cases the challenge remains to find such characterizations, preferably in terms of both symmetry condition and sufficient statistics. Diaconis and Freedman [1984] is an outstanding exposition, describing many such results and placing them into a unified theoretical superstructure.

In sum: Carnap recognized the limited utility of the inductive inferences that the continuum of inductive methods provided, and sought to extend his analysis to the case of analogical inductive inference: an observation of a given type makes more probable not merely observations of the exact same type but also observations of a "similar" type. The challenge lies both in making precise the meaning of "similar", and in being able to then derive the corresponding continua.

Carnap sought to meet the first challenge by proposing that underlying judgements of similarity is some notion of "distance" between predicates; but then immediately hit the brick wall of how one could use a general notion of distance to derive plausible continua. Neither Carnap nor any of his successors were able to solve this problem (although not for want of trying).

The Diaconis-Freedman theory enables us to see why. *If* one recognizes that the problem of analogical reasoning is essentially that of justifying parametric Bayesian inference, then it is indeed possible to derive attractive results that parallel those for the multinomial case. But these results are not trivial; they involve very hard mathematics, and although many special cases have been successfully tackled, it is possible to argue that no complete theoretical superstructure yet exists.

11 THE SAMPLING OF SPECIES PROBLEM

Another important problem concerns the nature of inductive inference when the possible types or species are initially unknown (this is sometimes referred to in the statistical literature as the *sampling of species problem*). Carnap thought this could be done using the equivalence relation R: *belongs to the same species as*. (That is, one has a notion of equivalence or common membership in a species, without prior knowledge of that species.) Carnap did not pursue this idea further, however, thinking the attempt premature given the relatively primitive state of the subject at that time.

Carnap's intuition was entirely on the mark here. One *can* construct a theory for the sampling of species problem, one that parallels the classical continuum of inductive methods — *but* the attendant technical difficulties are considerable, exchangeable random *sequences* being replaced by exchangeable random *partitions*. (Two sequences generate the same partition if they have the same *frequencies of frequences* a_r defined earlier.) Fortunately, the English mathematician J. H. C.

Kingman did the necessary technical spadework in a brilliant series of papers a quarter of a century ago. Kingman's beautiful results enable one to establish a parallel inductive theory for this case, including a Johnson-type characterization of an analogous continuum of inductive methods; see [Zabell, 1992; 1997].

In brief, consider the following three axioms, that parallel (in two cases) or extend (in one case) those of Johnson.

1. All sequences of outcomes are possible (have positive probability).
2. The probability of seeing on the next trial the i-th species already seen, is a function of the number of times that species has been observed, n_i, and the total sample size n: $f(n_i, n)$.
3. The probability of observing a new species is a function only of the number of species already observed t and the sample size n: $g(t, n)$.

It is a remarkable fact that if these three assumptions are satisfied, then one can prove that the functions $f(n_i, n), g(t, n)$ are members of a three-dimensional continuum described by three parameters α, θ, γ.

The continuum of inductive methods for the sampling of species

Case 1: If $n_i < n$ for some i, then

$$f(n_i, n) = \frac{n_i - \alpha}{n + \theta}, \qquad g(t, n) = \frac{t\alpha + \theta}{n + \theta}.$$

Note that if $n_i < n$, then $t > 1$, there are at least two species, and the universal generalization is *dis*confirmed.

Case 2: If $n_i = n$ for some i, then

$$f(n_i, n) = \frac{n_i - \alpha}{n + \theta} + c_n(\gamma), \qquad g(t, n) = \frac{t\alpha + \theta}{n + \theta} - c_n(\gamma);$$

here

$$c_n(\gamma) = \frac{\gamma(\alpha + \theta)}{(n + \theta)\left[\gamma + (\alpha + \theta - \gamma)\prod_{j=1}^{n-1}\left(\frac{j - \alpha}{j + \theta}\right)\right]}$$

represents the increase in the probability of seeing the i-th species again due to the confirmation of the universal generalization. Not all parameter values are possible: one must have

$$0 \leq \alpha < 1; \quad \theta > -\alpha; \quad 0 \leq \gamma < \alpha + \theta.$$

There is a simple interpretation of the three parameters θ, α, γ. The first, θ, is related to the likelihood of new species being observed; the larger the value of θ, the more likely it is that the next observation is that of a new species.

Observation of a new species has a double inductive import: it is a *new* species, and it is a *particular* species. Observing it contributes both to the likelihood that a new species will again be observed and, if a new species is not observed, that the species just observed will again be observed (as opposed to another species already observed); this is the role of α. Finally, the parameter γ is related to the likelihood that only one species will be observed. If ϵ is the initial probability that there will only be one species, then $\gamma = (\alpha + \theta)\epsilon$.

The special case $\alpha = \gamma = 0$ is of particular interest. In this case the probability of an "allelic partition" (set of frequencies of frequencies a_r) has a particularly simple form: given a sample of size n,

$$P(a_1, a_2, ..., a_n) = \frac{n!}{\theta(\theta+1)...(\theta+n-1)} \prod_{r=1}^{n} \frac{\theta^{a_r}}{r^{a_r} a_r!};$$

this is the *Ewens sampling formula.* There is a simple urn model for such a process in this case, analogous to the Polya urn model [Hoppe, 1984]. Suppose we start out with an urn containing a single, black ball: the *mutator*. The first time we select a ball, it is necessarily the black one. We replace it, together with a ball of some color. As time progresses, the urn contains the mutator and a number of colored balls. Each colored ball has a weight of one, the mutator has weight θ. The likelihood of selecting a ball is proportional to its weight. If a colored ball is selected, it is replaced together with a ball of the same color; this corresponds to observing a species that has already been observed before (hence balls of its color are already present). If the mutator is selected, it is replaced, *together with a ball of a new color*; this corresponds to observing a new species. It is not difficult to verify that the rules of succession for this process are

$$f(n_i, n) = \frac{n_i}{n+\theta}; \qquad g(n) = \frac{\theta}{n+\theta}.$$

Note that in this case the probability of a new species does not depend on the number observed. Such predictive probabilities arguably go back to De Morgan; see [Zabell, 1992].

12 A BUDGET OF PARADOXES

Strictly speaking, true paradox (in the sense of a basic contradiction in the theory itself) is no more possible in the Bayesian framework than it is in propositional logic: both are theories of consistency of input. The term "paradox" is often used instead to describe either some unexpected (but reasonable) consequence of the theory (so that we learn something from it); or an inconsistency arising from conflicting sets of inputs (which is what the theory is supposed to detect); or an apparent failure of the theory to explain what we regard as a valid intuition (which should be viewed as more of a challenge than a paradox). Nevertheless, analyzing and understanding such conundrums often gives us much greater insight into a subject, and the theory of probability has certainly had its fair share of such "challenge problems".

In the following paragraphs a few of these paradoxes are briefly noticed, more by way of initial orientation and an entry into the literature, than any detailed analysis. Indeed the literature on all of these is considerable.

12.1 The paradoxes of conditional probability

There is an amusing and interesting literature concerning conditional probability paradoxes such as the paradox of the second ace [Shafer, 1985], the three prisoner paradox [Falk, 1992], and the two-envelope paradox [Katz and Olin, 2007]. The unnecessary controversies that sometimes arise over these (for example, in *Philosophy of Science* and *The American Statistician*, names omitted to protect the guilty) are object lessons in the pitfalls that can attend informal attempts to analyze problems based on vague intuitions without the rigor of first carefully defining the sample space of possibilities or modeling the way information is received. Properly understood these puzzles serve as examples of the utility of the theory, not its deficiencies.

12.2 Hempel's paradox of the ravens

Nicod's criterion states that an assertion "all A are B" is supported by an observation of an A that is also a B; *Hempel's equivalence condition* that two logically equivalent propositions are equally confirmed by the same evidence. Hempel's paradox [Hempel, 1945], in its best-known (or most notorious) form considers the assertion "all ravens are black". This is equivalent to its contrapositive, "all non-black objects are not ravens". If one then observes a pink elephant, does this confirm the proposition "all ravens are black"?

Strictly speaking this is not a paradox of logical or subjective probability, because it follows just from Nicod's criterion and the equivalence condition. It is in any case easily accommodated within the Bayesian framework which, in brief, notes that pink elephants can indeed confirm black ravens, albeit to a very slight degree; see, e.g., [Hosiasson-Lindenbaum, 1940; Good, 1960]. Vranas [2004a], Howson and Urbach [2006, pp. 99–103], Fitelson [2008] provide entries to the recent literature; Sprenger [2009] provides a general survey and assessment.

12.3 Goodman's new riddle of induction

For Carnap, probability$_1$ is analytic and syntactic; probability$_2$ synthetic and semantic. Returning in 1941 to Keynes's *Treatise on Probability* with increased appreciation, Carnap sought to provide a satisfactory technical and quantitative foundation for inductive inference he saw as absent in Keynes. But after his paper proposing a purely syntactic justification for inductive inference [Carnap, 1945b], Nelson Goodman [1946] immediately published a serious challenge to it. To use the example later put forward by Goodman in *Fact, Fiction, and Forecast* (1954), under the striking heading of “the new riddle of induction”, Goodman defined a predicate *grue*: say an object is grue if, for some fixed time t, it is green before t and blue after. If all emeralds observed prior to time t are green, then this is equally consistent with their being either green and grue, and therefore apparently supports to an equal degree the expectation that emeralds observed after time t will be either green or red.

Goodman’s conclusion was that inductive inference is not purely syntactic in nature; that to varying degrees predicates are more or less *projectible*, projectability depending on the extent to which a predicate is *entrenched* in natural language. Although Goodman and Carnap soon agreed to disagree, there was no escape; and Goodman’s point is now generally accepted. (Carnap sought to meet this objection by invoking his *requirement of total evidence*, of which more in a moment.)

Goodman’s “new riddle” has sparked a substantial literature (see, e.g., [Stalker, 1994]). For a recent survey, see Schwartz [2009]. From a Bayesian perspective, projectability is effectively a question of the presence of exchangeability (or partially exchangeability); and as such this literature may be viewed as a complement to, rather than rival of the subjectivist position (see, e.g., [Horwich, 1982, pp. 67–72]). For Carnap’s final views on grue, see [Carnap and Jeffrey, 1971, pp. 73–76].

12.4 The principle of total evidence

Carnap’s initial defense to Goodman’s example was to invoke a *requirement of total evidence*, that

> in the application of inductive logic to a given knowledge situation, the total evidence available must be taken as basis for determining the degree of confirmation. [Carnap, 1950, p. 211]

This closed one hole in the dike, only for another to arise. In 1957 Ayer raised a fundamental question: in any purely logical theory of probability, why are new observations important? This is an issue that, as Good [1967] observes, is both related to the principle of total evidence and relevant to subjective theories of probability. Good’s solution to the conundrum was a neat one:

> [I]n expectation, it pays to take into account further evidence, provided that the cost of collecting and using this evidence, although positive, can be ignored. In particular, we, should use all the evidence *already* available, provided that the cost of doing so is negligible. With this proviso then, the principle of total evidence follows from the principle of rationality [that is, of maximizing expected utility].

For further discussion of the principle of total evidence, see [Skyrms, 1987]; for the value of knowledge, see [Horwich, 1982, pp. 122–129; Skryms, 1990, Chapter 4].

Related questions here are Glymour's *problem of old evidence* (if a theory T entails an experimental outcome E, but one observes E before this is discovered, does this increase the probability of T?), see, e.g., [Garber, 1983; Jeffrey, 1992, Chapter 5; Earman, 1992, Chapter 5; Jeffrey, 2004, pp. 44-47; Howson and Urbach, 2006, pp. 197–20]; and I. J. Good's concept of *dynamic* (or *evolving*) probability [Good, 1983, Chapter 10]. Central to both is the issue of the appropriateness of the principle of *logical omniscience*: if H logically entails E, then $P(E \mid H) = 1$. As Good notes [1983, p. 107], invoking a standard chestnut, it makes sense for purposes of betting to assign a probability of 1/10 that the millionth digit of π is a 7, even though one can, given sufficient times and resources, compute the actual digit (so that some would argue that the probability is either 0 or 1 depending). Discussion of this issue goes back at least to Polya [1941]; Hacking [1967] deals with the issue in terms of sentences that are "personally possible". (Of course from a practical Bayesian perspective one simple solution is to work with probabilities defined on subsets of a sample space rather than logical propositions or sentences. Thus in the case of π, take the sample space to be the set $\{0, 1, ..., 9\}$, and assign a coherent probability to the elements of the set. Whether or not it is profitable to expand the sample space to accommodate further events then goes to the issue of the value of further knowledge.)

12.5 The Popper-Carnap controversy and Miller's paradox

Karl Popper was a lifelong and dogged opponent of Carnap's inductivist views. In Appendix 7 of his *Logic of Scientific Discovery* [Popper, 1968] Popper made the claim that the logical probability of a universal generalization must be zero; today this can only be regarded as an historical curiosity. For two critiques (among many) of Popper's claim, see [Howson, 1973; 1987].

For those interested in the more general debate between Popper and Carnap, their exchange in the Schillp volume on Carnap [Schillp, 1963] is a natural place to start. For a general overview, see [Niiniluoto, 1973]. One important thread in the debate was *Miller's paradox*; Jeffrey [1975] is at once a useful reprise of the initial debate, and a spirited rebuttal. Closely related to Miller's paradox is Lewis's "principal principle"; see [Vranas, 2004b] for a recent discussion and many earlier references. For a more sympathetic view of Popper than the one here, see [Miller, 1997].

13 CARNAP REDUX

Thus far we have discussed Carnap's basic views regarding probability and inductive inference, some of his technical contributions to this area, and some of the extensions of Carnap's approach that took place during his lifetime and after. In this final part of the chapter we return to the philosophical (rather than technical) underpinnings of Carnap's approach, and attempt to place them in the context of both his predecessors and his successors.

13.1 "Two concepts of probability"

In his 1945 paper "The Two Concepts of Probability", Carnap advanced his view of "the problem of probability'. Noting a "bewildering multiplicity" of theories that had been advanced over the course of more than two and a half centuries, Carnap suggested one had to carefully steer between the Scylla and Charybdis of assuming either too few or too many underlying explicanda, and settled on just two. These two underlying concepts Carnap called $probability_1$ and $probability_2$: degree of confirmation versus relative frequency in the long run.

Carnap's identification of these two basic kingdoms of probability was not however novel; it is clearly stated in Poisson's 1837 treatise on probability (where Poisson uses the terms *probability* and *chance* to distinguish the two). Thus Poisson writes:

> In this work, the word chance will refer to events in themselves, independent of our knowledge of them, and we will retain the word probability ... for the reason we have to believe. [Poisson, 1837, p. 31]

Much the same distinction was made shortly after by Cournot [1843], *Exposition de la theorie des chances et des probabilités*, where he notes its "double sense", which he refers to as subjective and objective, a terminology also found later in [Bertrand, 1890] and [Poincaré, 1896]. Hacking [1975, p. 14] sees the distinction as going even further back to Condorcet in 1785. For discussion of Poisson and Cournot, see [Good, 1986, pp. 157–160; Hacking, 1990, pp. 96–99].

In the 20th century, Frank Plumpton Ramsey, one of the great architects of the modern subjective theory, likewise noted the possible validity of both senses:

> In this essay the Theory of Probability is taken as a branch of logic, the logic of partial belief and inconclusive argument; but there is no intention of implying that this is the only or even the most important aspect of the subject. Probability is of fundamental importance not only in logic but also in statistical and physical science, and we cannot be sure beforehand that the most useful interpretation of it in logic will be appropriate in physics also. Indeed the general difference of opinion between statisticians who for the most part adopt the frequency theory of probability and logicians who mostly reject it renders it likely that the two schools are really discussing different things, and that the word 'probability' is used by logicians in one sense and by statisticians in another.

This is as clear a statement of Carnap's distinction as one might imagine. (It can also be found clearly stated in a number of other places such as [Polya, 1941; Good, 1950].)

Thus, although the clear recognition of the fundamentally dual nature of probability did not originate with Carnap, the importance of his contribution is this: despite clear statements by Poisson in the 19th century, Ramsey in the 20th, and others both before and after, the lesson had not been learned; and even those who recognized the duality implicit in the *usage* of the word for the most part believed this to reflect a confusion of thought, only one of the two senses being truly legitimate. By carefully, forcefully, and in sustained fashion arguing for the legitimacy of both, Carnap enabled the distinction to at last become an entrenched philosophical commonplace. "The duality of probability has long been known to philosophers. The present generation may have learnt it from Carnap's weighty *Logical Foundations*" [Hacking, 1975, p. 13].

13.2 The later Carnap

Just as there is an early and later Wittgenstein, there is an early and later Carnap in inductive logic. Some of these changes were technical, but others reflected substantial shifts in Carnap's underlying views.

The appearance of Carnap's book generated considerable discussion and debate in the philosophical community. A second volume was promised, but never appeared. Like many before him, who found themselves enmeshed in the intellectual quicksand of the problem of induction (such as Bernoulli and Bayes), Carnap continued to grapple with the problem, refining and extending his results, but found that new advances and insights (on the part of himself, his collaborators, and others) were coming so quickly that he eventually abandoned as impractical the project of a definitive and systematic book-length treatment in favor of publishing from time to time compilations of progress reports. Two such installments eventually appeared [Carnap and Jeffrey, 1971; Jeffrey, 1980], although even these were delayed far past their initially anticipated date of publication.

Because no true successor to his *Logical Foundations of Probability* ever appeared, it is not always appreciated just how much of an evolution in Carnap's views about probability took place over the last two decades of his life. This change reflected in part a changing environment: the increasing appreciation of the prewar contributions of Ramsey and de Finetti, and the publication of such books as [Good, 1950; Savage, 1954; Raiffa and Schlaifer, 1961]. Important materials in documenting this shift include the introduction to the second [1962] edition of [Carnap, 1950], his paper "The aim of inductive logic" ([Carnap, 1962], reprinted in revised form in [Carnap and Jeffrey, 1971, Chapter 1]), Carnap's contributions to the Schilpp [1963] volume, and his posthumous "Basic system of inductive logic" ([Carnap and Jeffrey, 1971, Chapter 2; Jeffrey, 1980, Chapter 6]).

Technical shifts

Some of these shifts, although technical in nature, were quite important. First, there was a shift from *sentences* in a formal language to (effectively) *subsets* of a sample space. This reflected in part a desire to use the technical apparatus of modern mathematical probability, and in part

> a desire to formulate inductive logic in terms that had come to be standard in mathematical probability theory and theoretical statistics, where probabilities are attributed to "events" or ("propositions") which are construed as sets of entities which can handily be taken to be *models*, in the sense in which that term is used in logic. [Carnap and Jeffrey, 1971, p.1]

Second, as discussed at the beginning of this chapter, Carnap accepted the Ramsey–de Finetti–Savage link of probability to utility and decision making, its betting odds interpretation, the use of coherence and the Dutch book to derive the basic axioms of probability, and the central role of Bayes's theorem in belief revision. This placed Carnap squarely in the Bayesian camp, the differences coming down to ones of the existence or status of further epistemic constraints. This change came fairly quickly; it is already evident in Carnap's 1955 lecture notes [Carnap, 1973]. It is carefully stated in Carnap [1962] and then systematically elaborated in his *Basic System*.

Carnap also announced in the preface to his second edition of *Logical Foundations* the abandonment of his requirements of *logical independence* (replacing it by Kemeny's "meaning postulates"), and *completeness* for primitive predicates (replacing it by axioms relevant to language extensions). These are of less interest to us here.

The emerging Bayesian majority

Carnap's shift to the subjective was certainly noted by others. I. J. Good, for example, remarks "Between 1950 and 1961 Carnap moved close to my position in that he showed a much increased respect for the practical use of subjective probabilities" [Good, 1975, p. 41; see also p. 40, Figure 1]. But for the best evidence of this convergence of view between Carnap and the subjectivists, however, one can summon Carnap himself as a witness. In his *Basic System* (his last, posthumously published work on inductive inference), Carnap tells us

> I think there need not be a controversy between the objectivist point of view and the subjectivist or personalist point of view. Both have a legitimate place in the context of our work, that is, the construction of a set of rules for determining probability values with respect to possible evidence. At each step in the construction, a choice is to be made; the choice is not completely free but is restricted by certain boundaries. Basically, there is merely a difference in attitude or emphasis between the subjectivist tendency to emphasize the existing freedom of choice, and the objectivist tendency to stress the existence of limitations. [Jeffrey, 1980, p. 119]

The ultimate difference between Carnap and subjectivists of the de Finetti–Savage–Good stripe, then, appears to be how they view the logical status of these additional constraints. Carnap seems to have thought of them as forming in some sense a sequence or hierarchy (thus his "at each step in the construction"); modern Bayesians, in contrast, view these more as auxiliary tools. They do not deny the utility of the symmetry arguments that underly much of the Carnapian approach but, as Savage remarks, they "typically do not find the contexts in which such agreement obtains sufficiently definable to admit of expression in a postulate" [Savage, 1954, p. 66]. Such arguments fall instead under the rubric of what I. J. Good terms "suggestions for using the theory, these suggestions belonging to the technique rather than the theory" itself [Good, 1952, p. 107].

Let us take this a little further. Is what is at stake really just a "difference in attitude or emphasis" between choice and limitation? Here is how W. E. Johnson himself saw the enterprise (as he notes in his paper deriving the continuum of inductive methods):

> the postulate adopted in a controversial kind of theorem cannot be generalized to cover all sorts of working problems; so it is the logician's business, having once formulated a specific postulate, to indicate very carefully the factual and epistemic conditions under which it has practical value. [Johnson, 1932, pp. 418–419]

This is surely right. There are no universally applicable postulates: different symmetry assumptions are appropriate under different circumstances, none is logically compulsory. The best one can do is identify symmetry assumptions that seem natural, have identifiable consequences, and may be a natural reflection of one's beliefs under some reasonable set of circumstances. In judging the appropriate use of the sufficientness postulate, for example, the issue is not one of favoring "limitation" versus "choice"; it is one of *whether or not you think the postulate accurately captures the epistemic situation at hand*. This is the mission of partial exchangeability: to find different possible *qualitative* descriptions of the "the factual and epistemic conditions" that obtain in actual situations, descriptions that then turn out to have useful and satisfying *quantitative* implications.

From credence to credibility

Nevertheless Carnap did argue for additional symmetry requirements such as exchangeability; his explanation of this is perhaps most clearly presented in his 1962 paper "The aim of inductive logic". It will be apparent that Carnap and the subjectivists part company at this point because they had radically different goals.

Let Cr_t denote the subjective probability of an individual at time n, termed by Carnap *credence*. Using Bayes's rule, Carnap imagines a sequence of steps in which one obtains discrete quanta of data $E_j, j = 1, 2, ...$, giving rise in turn to a sequence of credences $Cr_{t+j}, j = 1, 2, ...$

> In the case of a human being we would hesitate to ascribe to him a credence function at a very early time point, before his abilities of reason and deliberate action are sufficiently developed. But again we disregard this difficulty by thinking either of an idealized human baby or of a robot. ... [L]et us acribe to him an *inital credence function* Cr_0 for the time point T_0 before he obtains his first datum E_1.

(This curiously echos Price's analysis of inductive inference in his appendix to Bayes's essay; see [Zabell, 1997, Section 3].)

The subsequent conditional credences based on this initial credence Cr_0 Carnap terms a *credibility*; and contrasts these with the "adult credence functions" of Ramsey, Savage, and de Finetti:

> When I propose to take as a basic concept, not adult credence, but either initial credence or credibility, I must admit that these concepts are less realistic and remoter from overt behavior and may therefore appear as elusive and dubious. On the other hand, when we are interested in *rational* decision theory, these concepts have great methodological advantages. Only for these concepts, not for credence, can we find a sufficient number of requirements of rationality as a basis for the construction of a system of inductive logic.

Thus Carnap asserts there are additional rationality requirements for Cr_0, ones having "no analogue for credence functions"; for example, symmetry of individuals (i.e., exchangeability). The assertion is that absent identifiable differences between individuals at the initial time T_0 (and since we are at the initial time T_0 we have not yet learned of any), the probability of any proposition involving two or more individuals should remain unchanged if the individuals are permuted (see [Carnap 1962, pp. 313–314; 1971, p. 118]). Carnap regards this as "the valid core of the old principle of indifference ... the basic idea of the principle is sound. Our task is to restate it by specific restricted axioms" [Carnap, 1962, p. 316; 1973, p. 277].

No wonder this part of Carnap's program never gained traction! It focuses on the credences of an "idealized human baby" rather than an adult; appeals to a state of complete ignorance; and presents itself as a rehabilitated version of the principle of indifference. And what does it mean to talk about individuals about what we know nothing except that they are different? In the end one exchanges one problem for another, replacing the task of finding a probability function by the (in fact much more daunting and questionable) task of establishing the existence of an underlying ideal language, one in which the description of sense experiences can be broken down into atomic interchangeable elements.

Such ideal languages are a seductive dream that in one form or another go back centuries, as in John Wilkins's philosophical language, or Leibniz's "characteristica universalis", which Leibniz thought could be used as the basis of a logical probability [Hacking, 1975, Chapter 15]. If Wittgenstein's early program of logical atomism had been successful, then logical probability might be possible, but the failure of the former dooms the latter. Lacking an ultimate language in one-to-one correspondence with reality, Carnapian programs retain an irreducible element of subjectivism.

Despite the ultimate futility of Carnap's program to justify induction in quantitative terms, the subjective Bayesian does provide a number of qualitative explicata. Inductive rationality in a single individual is not so much a matter of present opinion as the ability to be persuaded by further facts; and for two or more individuals by their ultimate arrival at consensus. To this end a number of results regarding convergence and merging of opinion have been discovered. For convergence of opinion see Skyrms [2006], and the earlier literature cited there; for merging of opinion see the classic paper of Blackwell and Dubins [1962] and the discussion in [Earman, 1992], as well as [Kalai and Lehrer, 1994] and [Miller and Sanchirico, 1999].

For further discussion of Carnap's program for inductive logic in its final form, see [Jeffrey, 1973].

14 CONCLUSION

Like his distinguished predecessors Bernoulli and Bayes, Rudolph Carnap continued to grapple with the elusive riddle of induction for the rest of his life. Throughout he was an effective spokesman for his point of view. But although the technical contributions of Carnap and his invisible college (such as Kemeny, Bar-Hillel, Jeffrey, Gaifman, Hintikka, Niiniluoto, Kuipers, Costantini, di Maio, and others) remain of considerable interest even today, Carnap's most lasting influence was more subtle but also more important: he largely shaped the way current philosophy views the nature and role of probability, in particular its widespread acceptance of the Bayesian paradigm (as, for example, in [Horwich, 1982; Earman, 1992; Mayer, 1993; Jaynes, 2003; Boven and Hartman, 2004; Jeffrey, 2004; Howson and Urbach, 2006]).

BIBLIOGRAPHY

[Armendt, 1993] Brad Armendt. Dutch books, additivity, and utility theory. *Philosophical Topics*, 21:1–20, 1993.

[Ayer, 1972] A. J. Ayer. *Probability and Evidence*. Macmillan, London, 1972.

[Barker, 1957] S. F. Barker. *Induction and Hypothesis*. Cornell University Press, Ithaca, 1957.

[Bayes, 1764] Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1764.

[Blackwell and Dubins, 1962] David Blackwell and Lester Dubins. Merging of opinions with increasing information. *Annals of Mathematical Statistics*, 33:882-886, 1962.

[Boole, 1854] George Boole. *An Investigation of the Laws of Thought*. Macmillan, London, 1854. Reprinted 1958, Dover Publications, New York.

[Bovens and Hartmann, 2004] Luc Bovens and Stephan Hartmann. *Bayesian Epistemologyy*. Oxford University Press, Oxford, 2004.

[Broad, 1918] C. D. Broad. The relation between induction and probability. *Mind*, 27:389–404; 29:11–45, 1918.

[Broad, 1924] C. D. Broad. Mr. Johnson on the logical foundations of science. *Mind*, 33:242–261, 369–384, 1924.

[Carnap, 1945a] Rudolph Carnap. On inductive logic. *Philosophy of Science*, 12:72–97, 1945.
[Carnap, 1945b] Rudolph Carnap. The two concepts of probability. *Philosophy and Phenomenological Research*, 5:513–532, 1945.
[Carnap, 1950] Rudolph Carnap. *Logical Foundations of Probability*. University of Chicago Press, Chicago, 1950. Second edition, 1962.
[Carnap, 1952] Rudolph Carnap. *The Continuum of Inductive Methods*. University of Chicago Press, Chicago, 1952.
[Carnap, 1962] Rudolph Carnap. The aim of inductive logic. In E. Nagel, P. Suppes, and A. Tarski, editors, *Logic, Methodology and Philosophy of Science*, pages 303–318. Stanford University Press, Stanford, 1962.
[Carnap, 1973] Rudolph Carnap. Notes on probability and induction. *Synthese*, 25:269–298, 1973.
[Carnap and Jeffrey, 1971] Rudolph Carnap and Richard C. Jeffrey, editors. *Studies in Inductive Logic and Probability*, volume I. University of California Press, Berkeley and Los Angeles, 1971.
[Carnap and Stegmüller, 1959] Rudolph Carnap and W. Stegmüller. *Inductive Logik und Wahrscheinlichkeit*. Springer-Verlag, Vienna, 1959.
[Cifarelli and Regazzini, 1996] Donato Michele Cifarelli and Eugenio Regazzini. De Finetti's contribution to probability and statistics. *Statistical Science*, 11:2253–282, 1996.
[de Finetti, 1937] Bruno de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7:1–68, 1937. Translated in H. E. Kyburg, Jr. and H. E. Smokler (eds.), *Studies in Subjective Probability*, Wiley, New York, 1964, pp. 93-158.
[de Finetti, 1938] Bruno de Finetti. Sur la condition de "équivalence partielle". In *Actualites Scientifiques et Industrielles*, volume 739, pages 5–18. Hermann, Paris, 1938.
[de Finetti, 1972] Bruno de Finetti. *Probability, Induction, and Statistics*. Wiley, New York, 1972.
[De Morgan, 1838] Augustus De Morgan. *An Essay on Probabilities, and their Application to Life Contingencies and Insurance Offices*. Longman, Orme, Brown, Green, and Longmans, London, 1838.
[Diaconis and Freedman, 1980a] Persi Diaconis and David Freedman. De Finetti's theorem for Markov chains. *Annals of Probability*, 8:115–130, 1980.
[Diaconis and Freedman, 1980b] Persi Diaconis and David Freedman. Finite exchangeable sequences. *Annals of Probability*, 8:745–764, 1980.
[Diaconis and Freedman, 1984] Persi Diaconis and David Freedman. Partial exchangeability and sufficiency. In J. K. Ghosh and J. Roy, editors, *Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference*, pages 205–236. Indian Statistical Institute, Calcutta, 1984.
[Diaconis and Ylvisaker, 1980] Persi Diaconis and Donald Ylvisaker. Conjugate priors for exponential families. *The Annals of Statistics*, 7:269-281, 1979.
[Earman, 1992] John Earman. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. M. I. T. Press, 1992.
[Edgeworth, 1922] Francis Ysidro Edgeworth. The philosophy of chance. *Mind*, 31:257–283, 1922.
[Ellis, 1854] Robert Leslie Ellis. Remarks on the fundamental principle of the theory of probabilities. *Transactions of the Cambridge Philosophical Society*, 9:605–607, 1854.
[Falk, 1992] Ruma Falk. A closer look at the probabilities of the notorious three prisoners. *Cognition*, 43:197–223, 1992.
[Feller, 1968] William Feller. *An Introduction to Mathematical Probability*. Wiley, New York, 3rd edition, 1968.
[Fitelson, 2008] Brandon Fitelson. Goodman's "new riddle". *Journal of Philosophical Logic*, 37:613–643, 2008.
[Garber, 1983] Daniel Garber. Old evidence and logical omniscience in Bayesian confirmation theory. In J. Earman, editor, *Testing Scientific Theories*, volume 10, pages 99–131. University of Minnesota Press, Minneapolis, 1983.
[Good, 1950] I. J. Good. *Probability and the Weighing of Evidence*. Hafner Press, New York, 1950.
[Good, 1952] I. J. Good. Rational decisions. *Journal of the Royal Statistical Society B*, 14:107–114, 1952.
[Good, 1959] I. J. Good. Kinds of probability. *Science*, 129:443–447, 1959.

[Good, 1960] I. J. Good. The paradoxes of confirmation. *British Journal for the Philosophy of Science*, 11:145–149, 1960. 12, 63–64.

[Good, 1965] I. J. Good. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. M. I. T. Press, Cambridge, Mass., 1965.

[Good, 1967] I. J. Good. On the principle of total evidence. *British Journal for the Philosophy of Science*, 17:319–321, 1967.

[Good, 1971] I. J. Good. 46656 varieties of Bayesians. *American Statistician*, 25:62–63, 1971.

[Good, 1975] I. J. Good. Explicativity, corroboration, and the relative odds of hypotheses. *Synthese*, 30:39–73, 1975.

[Good, 1983] I. J. Good. *Good Thinking*. University of Minnesota Press, Minneapolis, 1983.

[Good, 1986] I. J. Good. Some statistical applications of poisson's work. *Statistical Science*, 1:157–170, 1986.

[Goodman, 1946] Nelson Goodman. A query on confirmation. *Journal of Philosophy*, 43:383–385, 1946.

[Goodman, 1954] Nelson Goodman. *Fact, Fiction, and Forecast*. Hackett, Indianopolis, 1954.

[Hacking, 1967] Ian Hacking. Slightly more realistic personal probability. *Philosophy of Science*, 34:311–325, 1967.

[Hacking, 1975] Ian Hacking. *The Emergence of Probability*. Cambridge University Press, Cambridge, 1975.

[Heath, 1949] Sir Thomas Heath. *Mathematics in Aristotle*. Clarendon Press, Oxford, 1949.

[Hempel, 1945] C. G. Hempel. Studies in the logic of confirmation. *Mind*, 54:1–26, 97–121, 1945.

[Hintikka and Niiniluoto, 1980] J. Hintikka and I. Niiniluoto. An axiomatic foundation for the logic of inductive generalization. In R. C. Jeffrey, editor, *Studies in Inductive Logic and Probability*, volume 2, pages 157–181? University of California Press, Berkeley, 1980.

[Hintikka, 1966] J. Hintikka. A two-dimensional continuum of inductive methods. In J. Hintikka and P. Suppes, editors, *Aspects of Inductive Logic*, pages 113–132. North-Holland, Amsterdam, 1966.

[Hoppe, 1984] Fred Hoppe. Polya-like urns and the Ewens samapling formula. *Journal of Mathematical Biology*, 20:91–94, 1984.

[Horwich, 1982] Paul Horwich. *Probability and Evidence*. Cambridge University Press, Cambridge, 1982.

[Hosiasson-Lindenbaum, 1940] Janina Hosiasson-Lindenbaum. On confirmation. *Journal of Symbolic Logic*, 5:133–148, 1940.

[Howson and Urbach, 2006] Colin Howson and Peter Urbach. *Scientific Reasoning: The Bayesian Approach*. Open Court Press, Chicago and La Salle, IL, 3rd edition, 2006.

[Howson, 1973] Colin Howson. Must the logical probability of laws be zero? *British Journal for Philosophy of Science*, 24:153–163, 1973.

[Howson, 1987] Colin Howson. Popper, prior probabilities, and inductive inference. *British Journal for Philosophy of Science*, 38:207–224, 1987.

[Huzurbazar, 1955] V. S. Huzurbazar. On the certainty of an inductive inference. *Proceedings of the Cambridge Philosophical Society*, 51:761–762, 1955.

[Jaynes, 2003] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, 2003.

[Jeffrey, 1973] Richard C. Jeffrey. Carnap's inductive logic. *Synthese*, 25:299–306, 1973.

[Jeffrey, 1975] Richard C. Jeffrey. Probability and falsification: critique of the Popper program. *Synthese*, 30:95–117, 1975.

[Jeffrey, 1977] Richard C. Jeffrey. Mises redux. In R. E. Butts and J. Hintikka, editors, *Basic Problems in Methodology and Linguistics*, pages 213–222. Reidel, Dordrecht, 1977.

[Jeffrey, 1980] Richard C. Jeffrey, editor. *Studies in Inductive Logic and Probability*, volume II. University of California Press, Berkeley and Los Angeles, 1980.

[Jeffrey, 1983] Richard C. Jeffrey. *The Logic of Decision*. University of Chicago Press, Chicago, 2nd ed, 1983.

[Jeffrey, 1988] Richard C. Jeffrey. Conditioning, kinematics, and exchangeability. In W. L. Harper and B. Skyrms, editors, *Causation, Chance, and Credence*, volume 1, pages 221–255. Kluwer, Dordrecht, 1988.

[Jeffrey, 1992] Richard C. Jeffrey. *Probability and the Art of Judgment*. Cambridge University Press, Cambridge, 1992.

[Jeffrey, 2004] Richard C. Jeffrey. *Subjective Probability: The Real Thing*. Cambridge University Press, Cambridge, 2004.

[Jeffreys, 1961] Harold Jeffreys. *Theory of Probability*. Clarendon Press, Oxford, 3rd edition, 1961.

[Jevons, 1874] Wliiam Stanley Jevons. *The Principles of Scienc: A Treatise on Logic and Scientific Method*. Macmillan, London, 1st edition, 1874. (2nd e. 1877; reprinted 1958, Dover, New York).

[Johnson, 1924] William Ernest Johnson. *Logic, Part III: The Logical Foundations of Science*. Cambridge University Press, 1924.

[Johnson, 1932] William Ernest Johnson. Probability: The deductive and inductive problems. *Mind*, 41:409–423, 1932.

[Kalai and Lehrer, 1994] Ehud Kalai and Ehud Lehrer. Weak and strong merging of opinions. *Journal of Mathematical Economics*, 23:73-86, 1994.

[Katz and Olin, 2007] Bernard D. Katz and Doris Olin. A tale of two envelopes. *Mind*, 116:903–926, 2007.

[Kemeny, 1955] John Kemeny. Fair bets and inductive probabilities. *Journal of Symbolic Logic*, 20:263–273, 1955.

[Keynes, 1921] John Maynard Keynes. *A Treatise on Probability*. Macmillan, London, 1921.

[Kneale, 1949] William Kneale. *Probability and Induction*. The Clarendon Press, Oxford, 1949.

[Knopp, 1947] Konrad Knopp. *Theory and Application of Infinite Series*. Hafner Press, New York, 1947.

[Kuipers, 1978] Theo A. F. Kuipers. *Studies in Inductive Probability and Rational Expectation*. D. Reidel, Dordrecht, 1978.

[Maher, 1993] Patrick Maher. *Betting on Theories*. Cambridge Studies in Probability, Induction and Decision Theory. Cambridge University Press, Cambridge, 1993.

[Miller, 1997] David Miller. Sir Karl Raimund Popper, CH, FBA. *Biographical Memoirs of Fellows of The Royal Society of London*, 43:367–409, 1997.

[Miller and Sanchirico, 1999] Ronald I. Miller and Chris William Sanchirico. The role of absolute continuity in "merging of opinions" and "rational learning. *Games and Economic Behavior*, 29:170-190, 1999.

[Niiniluoto, 1973] Ilkka Niiniluoto. Review: Alex C. Michalos' *The Popper-Carnap Controversy*. *Synthese*, 25:417–436, 1973.

[Niiniluoto, 2009] Ilkka Niiniluoto. The development of the Hintikka program. In Dov Gabbay, Stephan Hartmann, and John Woods, editors, *Handbook of the History of Logic*, volume 10. Elsevier, London, 2009.

[Poincaré, 1896] Henri Poincaré. Calcul des probabilités. Paris, Gauthier-Villars, 1896 (2nd ed. 1912).

[Polya, 1941] George Polya. Heuristic reasoning and the theory of probability. *American Mathematical Monthly*, 48:450–465, 1941.

[Popper, 1959] Karl Popper. *The Logic of Scientific Discovery*. Basic Books, New York, 1959. Second ed. 1968, New York: Harper and Row.

[Prevost and L'Huilier, 1799a] Pierre Prevost and S. A. L'Huilier. Sur les probabilités. *Mémoires de l'Academie Royale de Berlin 1796*, pages 117–142, 1799.

[Prevost and L'Huilier, 1799b] Pierre Prevost and S. AS. L. L'Huilier. Mémoire sur l'arte d'estimer la probabilité des cause par les effets. *Mémoires de l'Academie Royale de Berlin*, 1796:3–24, 1799.

[Ramsey, 1931] Frank Plumpton Ramsey. Truth and probability. In R. B. Braithwaite, editor, *The Foundations of Mathematics and Other Logical Essays*, pages 156–198. Routledge and Kegan Paul, London, 1931. Read before the Cambridge Moral Sciences Club, 1926.

[Romeijn, 2006] Jan Willem Romeijn. Analogical predictions for explicit similarity. *Erkenntnis*, 64:253–280, 2006.

[Savage, 1954] Leonard J. Savage. *The Foundations of Statistics*. John Wiley, New York, 1954. Reprinted 1972, New York: Dover.

[Schillp, 1963] P. A. Schillp, editor. *The Philosophy of Rudolph Carnap*. Open Court, La Salle, IL, 1963.

[Schwartz, 2009] Robert Schwartz. Goodman and the demise of the syntactic model. In Dov Gabbay, Stephan Hartmann, and John Woods, editors, *Handbook of the History of Logic*, volume 10. Elsevier, London, 2009.

[Shafer, 1985] Glenn Shafer. Conditional probability. *International Statistical Review*, 53:261–275, 1985.

[Skyrms, 1987] Brian Skyrms. On the principle of total evidence with and without observation sentences. In *Logic, Philosophy of Science and Epistemology: Proceedings of the 11th International Wittgenstein Symposium*, pages 187–195. Hölder–Pichler–Tempsky, 1987.

[Skyrms, 1990] Brian Skyrms, editor. *The Dynamics of Rational Deliberation*. Harvard University Press, Cambridge, MA, 1990.

[Skyrms, 1991] Brian Skyrms. Carnapian inductive logic for Markov chains. *Erkenntnis*, 35:439–460, 1991.

[Skyrms, 1993] Brian Skyrms. Analogy by similarity in hypercarnapian inductive logic. In G. J. Massey J. Earman, A. I. Janis and N. Rescher, editors, *Philosophical Problems of the Internal and External Worlds: Essays Concerning the Philosophy of Adolf Grünbaum*, pages 273–282. Pittsburgh University Press, Pittsburgh, 1993.

[Skyrms, 1996] Brian Skyrms. Inductive logic and Bayesian statistics. In *Statistics, Probability, and Game Theory: Papers in Honor of David Blackwell*, volume 30 of *IMS Lecture Notes — Monograph Series*, pages 321–336. Institute of Mathematical Statistics, 1996.

[Skryms, 2006] Brian Skyrms. Diachronic coherence and radical probabilism. *Philosophy of Science*, 73:959-968, 2006.

[Sprenger, 2009] Jan Sprenger. Hempel and the paradoxes of confirmation. In Dov Gabbay, Stephan Hartmann, and John Woods, editors, *Handbook of the History of Logic*, volume 10. Elsevier, London, 2009.

[Stalker, 1994] Douglas Stalker. *Grue! The New Riddle of Induction*. Open Court, Chicago, 1994.

[Stigler, 1982] Stephen M. Stigler. Thomas Bayes's Bayesian inference. *Journal of the Royal Statistical Society Series A*, 145:250–258, 1982.

[Venn, 1866] John Venn. *The Logic of Chance*. Macmillan, London, 1866. (2nd ed. 1876, 3rd ed. 1888; reprinted 1962, Chelsea, New York).

[Vranas, 2004a] P. B. M. Vranas. Hempel's raven paradox: A lacuna in the standard Bayesian solution. *The British Journal for the Philosophy of Science*, 55:545–560, 2004.

[Vranas, 2004b] P. B. M. Vranas. Have your cake and eat it too: the old principal principle reconciled with the new. *Philosophy and Phenomenological Research*, 69:368-382, 2004.

[Waismann, 1930] Friedrich Waismann. Logische Analyse des Wahrscheinlichkeitsbegriffs. *Erkenntis*, 1:228–248, 1930.

[Wrinch and Jeffreys, 1919] Dorothy Wrinch and Harold Jeffreys. On certain aspects of the theory of probability. *Philosophical Magazine*, 38:715–731, 1919.

[Wrinch and Jeffreys, 1921] Dorothy Wrinch and Harold Jeffreys. On certain fundamental principles of scientific enquiry. *Philosophical Magazine, Series 6*, 42:369–390, 1921.

[Zabell, 1982] S. L. Zabell. W. E. Johnson's "sufficientness postulate". *Annals of Statistics*, 10:1091–1099, 1982.

[Zabell, 1988] S. L. Zabell. Symmetry and its discontents. In W. L. Harper and B. Skyrms, editors, *Causation, Chance, and Credence*, volume 1, pages 155–190. Kluwer, Dordrecht, 1988.

[Zabell, 1989] S. L. Zabell. The rule of succession. *Erkenntnis*, 31:283–321, 1989.

[Zabell, 1992] S. L. Zabell. Predicting the unpredictable. *Synthese*, 90:205–232, 1992.

[Zabell, 1995] S. L. Zabell. Characterizing Markov exchangeable sequences. *Journal of Theoretical Probability*, 8:175–178, 1995.

[Zabell, 1996] S. L. Zabell. Confirming universal generalizations. *Erkenntnis*, 45:267–283, 1996.

[Zabell, 1997] S. L. Zabell. The continuum of inductive methods revisited. In John Earman and John D. Norton, editors, *The Cosmos of Science: Essays of Exploration*, Pittsburg-Konstanz Series in the Philosophy and History of Science, pages 351–385. University of Pittsburgh Press/Universitätsverlag Konstanz, 1997.

[Zabell, 2007] S. L. Zabell. Carnap on probability and induction. In Michael Friedman and Richard Creath, editors, *The Cambridge Companion to Carnap*, pages 273–294. Cambridge University Press, 2007.