S. L. ZABELL

PREDICTING THE UNPREDICTABLE

ABSTRACT. A major difficulty for currently existing theories of inductive inference involves the question of what to do when novel, unknown, or previously unsuspected phenomena occur. In this paper one particular instance of this difficulty is considered, the so-called *sampling of species problem.*

The classical probabilistic theories of inductive inference due to Laplace, Johnson, de Finetti, and Carnap adopt a model of simple enumerative induction in which there are a prespecified number of types or species which may be observed. But, realistically, this is often not the case. In 1838 the English mathematician Augustus De Morgan proposed a modification of the Laplacian model to accommodate situations where the possible types or species to be observed are not assumed to be known in advance; but he did not advance a justification for his solution.

In this paper a general philosophical approach to such problems is suggested, drawing on work of the English mathematician J. F. C. Kingman. It then emerges that the solution advanced by De Morgan has a very deep, if not totally unexpected, justification. The key idea is that although 'exchangeable' random *sequences* are the right objects to consider when all possible outcome-types are known in advance, exchangeable random *partitions* are the right objects to consider when they are not. The result turns out to be very satisfying. The classical theory has several basic elements: a representation theorem for the general exchangeable sequence (the de Finetti representation theorem), a distinguished class of sequences (those employing Dirichlet priors), and a corresponding rule of succession (the continuum of inductive methods). The new theory has parallel basic elements: a representation theorem for the general exchangeable random partition (the Kingman representation theorem), a distinguished class of random partitions (the Poisson-Dirichlet process), and a rule of succession which corresponds to De Morgan's rule.

1. INTRODUCTION

An important question rarely discussed in accounts of inductive inference is what to do when the utterly unexpected occurs, an outcome for which no slot has been provided. Alternatively – since we know this *will* happen on occasion – *how* can we coherently incorporate such new information into the body of our old beliefs? The very attempt to do so seems paradoxical within the framework of Bayesian inference, a theory of consistency between old and new information.

This is not the problem of observing the 'impossible', that is, an event whose possibility we have considered but whose probability we judge to be 0. Rather, the problem arises when we observe an event

Synthese **90**: 205–232, 1992.

whose existence we did not even previously suspect; this is the so-called problem of 'unanticipated knowledge'. This is a very different problem from the one just mentioned: it is not that we judge such events impossible – indeed, after the fact we may view them as quite plausible – it is just that beforehand we did not even consider their possibility. On the surface there would appear to be no way of incorporating such new information into our system of beliefs, other than starting over from scratch and completely reassessing our subjective probabilities. Coherence of old and new makes no sense here; there are no old beliefs for the new to cohere with.

A special instance of this phenomenon is the so-called *sampling of species problem*. Imagine that we are in a new terrain, and observe the different species present. Based on our past experience, we may anticipate seeing certain old friends – black crows, for example – but stumbling across a giant panda may be a complete surprise. And, yet, all such information will be grist to our mill: if the region is found rich in the variety of species present, the chance of seeing a particular species again may be judged small, while if there are only a few present, the chances of another sighting will be judged quite high. The unanticipated has its uses.

Thus, the problem arises: How can the theory of inductive inference deal with the potential existence of unanticipated knowledge, and, how can such knowledge be rationally incorporated into the corpus of our previous beliefs? How can we predict the occurrence of something we neither know, nor even suspect, exists? Subjective probability and Bayesian inference, despite their many impressive successes, would seem at a loss to handle such a problem given their structure and content. Nevertheless, in 1838 the English mathematician Augustus De Morgan proposed a method for dealing with precisely this difficulty. This paper describes De Morgan's proposal and sets it within the context of other attempts to explain induction in probabilistic terms.

The organization of the paper is as follows. The second section gives some historical background and briefly describes De Morgan's rule. As will be seen, although the statement of the rule is unambiguous, its justification – at least, as described by De Morgan – is unclear, and our goal will be to understand why De Morgan's rule makes sense. We begin this task by briefly reviewing, in the third section of the paper, the classical analysis of the inductive process in probabilistic terms. This is very well-known material, and our goal here is simply to set

up a framework in which to place De Morgan's rule. This is then done in the fourth and fifth sections of the paper: the key point is that while 'exchangeable' random *sequences* are the right objects to consider when all possible outcomes are known in advance, exchangeable random *partitions* are the right objects to consider when they are not.

The result turns out to be very satisfying. The classical theory has several basic elements: a representation theorem for the general exchangeable sequence (the 'de Finetti representation theorem'), a distinguished class of sequences (those arising from the so-called 'Dirichlet priors'), a 'rule of succession', specifiying the probability of a future outcome (Carnap's 'continuum of inductive methods'), and an urn-model interpretation ('Polya's urn'). The new theory, developed by the English mathematician J. F. C. Kingman for another purpose but ideally suited for this, has parallel basic elements: a representation theorem for the general exchangeable random partition (the Kingman representation theorem), a distinguished class of random partitions (the 'Poisson-Dirichlet process'), an urn-model representation (sometimes called the 'Chinese restaurant process'), and a rule of succession which corresponds to . . . De Morgan's rule!

The problem considered by De Morgan is closely related to a statistical problem, mentioned earlier, termed 'the sampling of species' problem. There have been a number of attempts to analyze such questions, beginning with the distinguished English statistician R. A. Fisher. This literature is briefly summarized in the final section of the paper, together with some concluding remarks concerning the original inductive problem.

2. THE DE MORGAN PROCESS AND ITS ANTECEDENTS

Hume's problem of induction asks why we expect the future to resemble the past. One of the most common methods of attempting to answer Hume's question invokes probability theory; and *Laplace's rule of succession* is the classical form of this type of explanation. It states that if an event has occurred n times out of N in the past, then the probability that it will occur the next time is $(n + 1)/(N + 2)$. This version of the rule implicitly assumes that possible outcomes are dichotomous; that is, an event of a specified type either did or did not occur. A more complex form of the rule, which can also be found in Laplace's writings, posits instead a multiplicity of possible outcomes. In this setting, the

rule becomes: if there are t possible outcomes (labelled $c_1, c_2, \ldots, c_t$), if X_k denotes the outcome occurring on the k-th trial, and if the vector $\mathbf{n} = (n_1, n_2, \ldots, n_t)$ records the number of instances in which each of the t possible outcomes occur in a total of N trials, then the probability that an outcome of the j-th type will occur again on the next trial is

LAPLACE'S RULE:

$$P[X_{N+1} = c_j | \mathbf{n}] = \frac{n_j + 1}{N + t}$$

But as the English mathematician and logician De Morgan noted,

> [t]here remains, however, an important case not yet considered; suppose that having obtained t sorts in N drawings, and t sorts only, we do not yet take it for granted that these are all the possible cases, but allow ourselves to imagine there may be sorts not yet come out. (De Morgan 1845, p. 414)

The problem of how to deal with the observation of novel phenomena in Bayesian inference is as old as Bayes's theorem itself. In Price's appendix to Bayes's essay (Bayes 1764, pp. 149–53), Price supposes "a solid or die of whose number of sides and constitution we know nothing; and that we are to judge of these from experiments made in throwing it". Price argues that "the first throw only shows that *it has* the side then thrown", and that it is only "*after* the first throw and not before, [that] we should be in the circumstances required" for the application of Bayes's theorem. Price's subsequent analysis, however, is confined to those cases where our experience is *uniform*, that is, where "the same event has followed without interruption in any one or more subsequent experiments" (e.g., the rising of the sun); or where it is known in advance that there are only two categories (e.g., the drawing of a lottery with *Blanks* and *Prizes*).

Laplace considered the multinomial case where there are three or more categories (Laplace 1781, Section 33), but his analysis is limited to those instances where the number of categories is fixed in advance. De Morgan, in contrast, proposed a simple way of dealing with the possibility of an unknown number of categories (De Morgan 1838, pp. 66–67; 1845, pp. 414–15). If initially there are t possible outcomes known, then De Morgan gives as the probability of seeing the outcome on trial N + 1 fall into the j-th category:

DE MORGAN'S RULE:

$$P[X_{N+1} = c_j | \mathbf{n}] = \frac{n_j + 1}{N + t + 1}.$$

That is, one creates an additional category: "new species not yet observed", which has a probability of $1/(N + t + 1)$ of occurring.

How can one make sense of De Morgan's idea? First, it is unclear what one should do after observing a new 'species'. De Morgan (1845, p. 415) takes t to be the number of species present in the sample at any given instant; so that it increases over time. But if De Morgan's rule is thought of as a generalization of Laplace's, then it is more appropriate to view t as fixed, the number of species known to exist prior to sampling. (This second convention is the one employed below.) Nor is it clear whether De Morgan's prescription is even consistent, in the sense that one can find a probability function on sequences which agrees with his rule. So, the first item of business is to see that this is indeed the case.

2.1. *The De Morgan Process*

It turns out that there is a simple urn model which generates the sequence of probabilities suggested by De Morgan. Consider an urn with one black ball (the 'mutator'), and t additional balls, each of a different color, say, $c_1, c_2, \ldots, c_t$. We reach into the urn, pick a ball at random, and return it to the urn *together with a new ball*, according to the following rule:

- If a colored ball is drawn, then it is replaced together with another of the *same* color.
- If the mutator is drawn, then it is replaced together with another ball of an *entirely new* color.

The colored balls correspond to species known to exist; selecting a ball of a given color corresponds to observing the species represented by that color; selecting the mutator to observing a hitherto unknown species.

Clearly this sequence of operations generates the probabilities De Morgan suggests. After N drawings, there are $N + t + 1$ balls in the urn, because we started out with t (the colored balls) + 1 (the mutator) and have added N since. Because we are choosing balls at random,

each has a probability of 1 in $N + t + 1$ of being selected. The number of colors is gradually changing, but if there are n_j+1 balls of a specific type, then the probability of observing that type at the next draw is the one given by De Morgan. On the other hand, since there is always only one mutator, the probability of it being selected (the probability that a new species is observed) is $1/(N + t + 1)$. This process generates the probabilities specified by De Morgan, so we shall call it the *De Morgan process*.

More generally, we might imagine that the mutator has a 'weight' θ accorded to it, $0 < \theta < \infty$, so that it is either more or less likely to be selected than the colored balls in the urn, which are accorded a weight of 1. That is, each colored ball has a probability of $(N + t + \theta)^{-1}$ of being selected, while the mutator has probability $1 - (N + t)/(N + t + \theta) = \theta/(N + t + \theta)$. This will also be called a De Morgan process (with parameter θ).

So, De Morgan's prescription is consistent. But does it make sense? Isn't it simply arbitrary, no better or worse than any of a broad spectrum of rules we could invent? The answer, surprisingly, is '*No*': it turns out to be a very special process, with many distinctive and attractive features. But, in order to appreciate this, we need to briefly review the classical probabilistic account of induction for a fixed number of categories, and then leap forward nearly a century and a half, when the De Morgan process mysteriously reappears in the 1970s.

3. EXCHANGEABLE RANDOM SEQUENCES

Attempts to explain enumerative induction in probabilistic terms go back to Bayes and Laplace, but this program was perfected at the hands of the twentieth-century Italian mathematician and philosopher Bruno de Finetti. De Finetti's crucial insight was that those situations in which the simplest forms of enumerative induction are appropriate are captured by the mathematical concept of 'exchangeability', and that the mathematical structure of such sequences is readily described.

3.1. *The De Finetti Representation Theorem*

Let $X_1, X_2, \ldots X_N, \ldots$ be an infinite sequence of random variables taking on any of a finite number of values, say $c_1, c_2, \ldots, c_t$: these are the possible *categories* or *cells* into which the outcomes of the sequence

are classified, and might denote different species in an ecosystem, or words in a language. The sequence is said to be *exchangeable* if for every N the 'cylinder set' probabilities

$$P[X_1 = e_1, X_2 = e_2, \ldots, X_N = e_N] = P[e_1, e_2, \ldots, e_N]$$

are invariant under all possible permutations of the time index. Put another way, two sequences have the same probability if one is a *rearrangement* of the other. If the outcomes are thought of as letters in an alphabet, then this means that all words of the same length having the same letters have the same probability.

Given a sequence of possible outcomes $e_1, e_2, \ldots, e_N$, let n_j denote the number of times the j-th type occurs in the sequence. The frequency vector $\mathbf{n} = (n_1, n_2, \ldots, n_t)$ plays a key role in exchangeability (in Carnap's terminology, it is the "structure-description"). First, it provides an equivalent characterization of exchangeability, since given any two sequences, say $\mathbf{e} = (e_1, e_2, \ldots, e_N)$ and $\mathbf{e}^* = (e_1^*, e_2^*, \ldots, e_N^*)$, one can be obtained from the other by rearrangement if and only if the two have the same frequency vector. Thus, P is exchangeable if and only if two sequences having the same frequency vector have the same probability.

In the language of theoretical statistics, the observed frequency counts $n_j = n_j(X_1, X_2, \ldots, X_N)$ are *sufficient statistics* for the sequence $\{X_1, X_2, \ldots, X_N\}$, in the sense that probabilities conditional on the frequency counts depend only on **n**, and are independent of the choice of exchangeable P: given a particular value of the frequency vector, the only sequences possible are those having this frequency vector, and each of these, by exchangeability, is assumed equally likely. The number of such sequences is given by the *multinomial coefficient* $N!/(n_1!\, n_2! \ldots n_t!)$; and, thus, the probability of such a sequence is

$$P[X_1, X_2, \ldots, X_N | \mathbf{n}] = \frac{n_1!\, n_2! \ldots n_t!}{N!}.$$

The structure of exchangeable sequences is actually quite simple. Let

$$\Delta_t =: \{(p_1, p_2, \ldots, p_t): p_i \geq 0 \text{ and } p_1 + p_2 + \cdots + p_t = 1\}$$

denote the t-simplex of probabilities on t elements. Every element of the simplex determines a *multinomial probability*, and the general exchangeable probability is a mixture of these. This is the content of a

celebrated theorem due to de Finetti: if an infinite sequence of t-valued random variables $X_1, X_2, X_3, \ldots$ is exchangeable, and $(n_1, n_2, \ldots, n_t)$ is the vector of frequencies for $\{X_1, X_2 \ldots, X_N\}$, then the infinite limiting frequency

$$Z =: \lim_{N \to \infty} \left(\frac{n_1}{N}, \frac{n_2}{N}, \ldots, \frac{n_t}{N} \right)$$

exists almost surely; and, if $\mu(A) = P[Z \in A]$ denotes the distribution of this limiting frequency, then

$$P[X_1 = e_1, X_2 = e_2, \ldots, X_N = e_N] = \int_{\Delta_t} p^{n_1} p^{n_2} \ldots p^{n_t} d\mu(p_1, p_2, \ldots, p_{t-1}).$$

The use of such integral representations of course predates de Finetti; de Finetti's contribution was to give a philosophical justification for their use, based on the concept of exchangeability, one not appealing to objective chances or second-order probabilities to explain the nature of the multinomial probabilities appearing in the mixture (see, e.g., Zabell 1988, 1989).

3.2. *Determining the Prior Measure $d\mu$*

In order to apply the de Finetti representation theorem, it is necessary to decide on a specific 'prior' $d\mu$. In principle $d\mu$ can be anything, but it is natural to single out classes of priors thought to represent situations of limited knowledge or 'ignorance'. Such ideas go back to Bayes himself, who considered "an event concerning the probability of which we absolutely know nothing antecedently to any trials made concerning it" (Bayes 1764). The earliest and best-known prior is the so-called 'Bayes–Laplace prior', which assumes that there are two categories, say 'success' and 'failure' (so that $t = 2$), and takes $d\mu(p) = dp$. Although Laplace made direct use of this prior, Bayes deduced it by a more circuitous route, assuming that S_N, the number of successes in N trials, is equally likely to assume any value between 0 and N: $P[S_N = k] = 1/(N + 1)$. This assumption in fact uniquely determines $d\mu$ (see Zabell 1988, pp. 159–60).

There is an obvious generalization of Bayes's postulate, employing

the frequency vector, which was proposed by the English logician, philosopher, and economic theorist William Ernest Johnson. This is Johnson's "combination postulate" (Johnson 1924): *All ordered t-partitions of N are equally likely*. That is, all possible frequency vectors $\mathbf{n} = (n_1, n_2, \ldots, n_t)$ are assumed to have equal probability of occurring. (Note that if $t = 2$, then $(n_1, n_2) = (k, N - k)$ and Johnson's postulate reduces to Bayes's.) Since there are

$$A_{N,t} =: \binom{N + t - 1}{t}$$

ordered t-partitions of N (see, e.g., Feller 1968, p. 38), each of these, assuming the combination postulate, has probability $1/A_{N,t}$ of occurring. In mathematical probability the frequency counts are often referred to as *occupancy numbers*, and the probability distribution arising from the combination postulate as *Bose–Einstein statistics* (see, generally, Feller 1968, chapter 2, Section 5). The force of Johnson's combination postulate is that, just as in the binomial case, it uniquely determines the mixing measure $d\mu$; here the uniform or 'flat' prior $d\mu(p_1, p_2, \ldots, p_t) = dp_1 dp_2 \ldots dp_{t-1}$, first introduced by Laplace in 1778.

3.3. *The Rule of Succession*

Once the prior $d\mu$ has been implicitly or explicitly specified, one can immediately calculate the *predictive probabilities* that it gives rise to:

$$P[X_{N+1} = c_i | X_1, X_2, \ldots, X_N] = P[X_{N+1} = c_i | \mathbf{n}].$$

Such a conditional probability is sometimes called a 'rule of succession' (the terminology is due to the English logician John Venn). For example, in the case of the Bayes–Laplace prior (where $t = 2$), a simple integration immediately yields Laplace's rule of succession, $(n_1 + 1)/(N + 1)$; and for Johnson's combination postulate the corresponding rule of succession is $(n_j + 1)/(N + t)$ (Johnson 1924, Appendix). A rule of succession uniquely determines the probability of any possible sequence; and the probability specification on sequences corresponding to the combination postulate is, in Carnap's terminology, the c* function.

There is an air of arbitrariness about the combination postulate, and both Johnson (and later Carnap) ultimately replaced it with one less

stringent, Johnson's 'sufficientness' postulate (the terminology is due to I. J. Good):

$$P[X_{N+1} = i|\mathbf{n}] = f(n_i, N).$$

That is, the only relevant information conveyed by the sample, vis-à-vis predicting whether the next outcome will fall into a given category, is the number of outcomes observed in that category to date; any knowledge of how outcomes not in that category distribute themselves among the remainder is posited to be irrelevant.

As a consequence of the sufficientness postulate, Johnson was able to derive, just as in the case of the combination postulate, the corresponding rule of succession: if $X_1, X_2, \ldots$ is an exchangeable sequence satisfying the sufficientness postulate, and $t \geqslant 3$, then (assuming that all cylinder set probabilities are positive so that the relevant conditional probabilities exist)

$$P[X_{N+1} = i|\mathbf{n}] = \frac{n_i + \alpha}{N + t\alpha}$$

(see, generally, Zabell 1982). The corresponding measure in the de Finetti representation in this case is the *symmetrical Dirichlet prior* with parameter α:

$$d\mu(p_1, p_2, \ldots, p_t) = \frac{\Gamma(t\alpha)}{\Gamma(\alpha)^t} p_1^{\alpha-1} p_2^{\alpha-1} \ldots p_t^{\alpha-1} dp_1 dp_2 \ldots dp_{t-1}.$$

3.4. *Polya's Urn Model*

There is a simple urn model which generates Laplace's rule of succession. It is usually referred to as the Polya urn model (see, e.g., Feller 1968, pp. 119–21), after the mathematician George Polya, who proposed its use as a model for the spread of contagious diseases, although a description of it (in the case of all successes) can be found in Quetelet's *Lettres sur la théorie des probabilités* (Quetelet 1846, p. 367).

4. PARTITION EXCHANGEABILITY

Johnson's sufficientness postulate, or its later equivalent formulation, Carnap's 'continuum of inductive methods', attempts to capture the

concept of prior ignorance about individual categories. Despite its attractiveness, however, it is far from clear that Johnson's sufficientness postulate is a necessary condition for such a state of ignorance. Is it possible to further weaken the notion of absence of information about the categories? A natural idea is that ignorance about individual *categories* should result in a symmetry of beliefs similar to that captured by de Finetti's notion of exchangeability with respect to *times*. This suggests the following definition.

DEFINITION: A probability function P is *partition exchangeable* if the cylinder set probabilities $P[X_1 = e_1, X_2 = e_2, \ldots, X_N = e_N]$ are invariant under permutations of the time index *and* the category index.

For example, if we are rolling a die (so that t = 6), and our subjective probabilities for the various outcomes are partition exchangeable, then

$$P[6, 4, 6, 4, 4, 5, 1, 2, 5] = P[1, 1, 1, 2, 2, 3, 3, 4, 5].$$

This can be seen by first arranging the sequence

$$\{6, 4, 6, 4, 4, 5, 1, 2, 5\}$$

into 'regular position':

$$\{4, 4, 4, 5, 5, 6, 6, 1, 2\},$$

(i.e., descending order of observed frequency for each face); and then follow this up by the category permutation

$$1 \to 4 \to 1,\ 2 \to 5 \to 2,\ 3 \to 6 \to 3,$$

which can be more compactly written as (1, 4)(2, 5)(3, 6).

The 'sufficient statistics' for a partition exchangeable sequence are the *frequencies of the frequencies* (or '*abundances*'):

$$\mathbf{a_r} =: \text{number of } \mathbf{n_j} \text{ equal to } \mathbf{r}$$

Example: Suppose one observes the sequence 5, 2, 6, 1, 2, 3, 5, 1, 1, 2. Then:

$N = 10$; $t = 6$.
$n_1 = 3$, $n_2 = 3$, $n_3 = 1$, $n_4 = 0$, $n_5 = 2$, $n_6 = 1$.
$\mathbf{n} = (3, 3, 1, 0, 2, 1)$ "=" $0^1 1^2 2^1 3^2$

$$a_0 = 1,\ a_1 = 2,\ a_2 = 1,\ a_3 = 2,\ a_4 = \ldots a_{10} = 0.$$
$$\mathbf{a} = (1, 2, 1, 2, 0, \ldots, 0)$$

A useful bit of terminology will be to call the **a**-vector the *partition vector*. (Kingman (1980, p. 36) calls it the "allelic partition".) Note that in a partition exchangeable sequence, $P[X_1 = 1/t]$, so the number of categories that appear in such a sequence must be finite.

The partition vector plays the same role relative to partition exchangeable sequences that the frequency vector plays for ordinary exchangeable sequences; that is, two sequences are equivalent, in the sense that one can be obtained from the other by a permutation of the time set and a permutation of the category set, if and only if the two sequences have the same partition vector. Thus, an alternative characterization of partition exchangeability is that: *all sequences having the same partition vector have the same probability*. The frequencies of the frequencies, furthermore, are the sufficient statistics for a partition exchangeable sequence, since probabilities conditional on the partition vector $\mathbf{a} = (a_0, a_1, \ldots, a_t)$ are independent of P: and, given a partition vector **a**, the only possible sequences have **a** as partition vector and each of these is equally likely. (Note that this refers only to the cylinder set probabilities involving $X_1, X_2, \ldots, X_N$. The predictive probabilities for $X_{N+1}, X_{N+2}, \ldots$ will still depend on the a_r.)

According to the de Finetti representation theorem, a partition exchangeable sequence, being exchangeable, can be represented by a mixing measure $d\mu$ on the t-simplex Δ_t. An important subset of the t-simplex in the partition exchangeable case is the subsimplex of ordered probabilities:

$$\Delta_t^* = \{(p_1^*, p_2^*, \ldots, p_t^*): p_1^* \geq p_2^* \geq \cdots \geq p_t^* \geq 0, \Sigma_j p_j^* = 1\}$$

In the partition exchangeable case, once the prior $d\mu$ is known on the ordered t-simplex Δ_t^*, it is automatically determined on all of Δ_t by symmetry. It is not really difficult to prove this, but it is perhaps best seen by considering a few simple examples.

Consider, first, the case of a coin which we know to biased 2:1 in favor of one side, but where we don't know which side it is – it could be either with equal probability. Then, $p_1^* = 2/3$, $p_2^* = 1/3$. In terms of the original, unordered probabilities, this corresponds to either $p_1 = 2/3$, $p_2 = 1/3$ or $p_1 = 1/3$, $p_2 = 2/3$ and, since we are indifferent be-

tween categories, these two possibilities are equally likely; thus, we have as the mixing measure on the simplex Δ_2 the measure

$$d\mu(p) = \frac{1}{2}\delta_{2/3} + \frac{1}{2}\delta_{1/3},$$

where δ_x is the Dirac measure which assigns probability 1 to x. This is a partition exchangeable probability, since it is invariant under the interchange $H \rightarrow T$, $T \rightarrow H$.

Consider next a die with six faces. The most general exchangeable probability is obtained by mixing multinomial probabilities over the simplex Δ_6. The partition exchangeable probabilities are those which are invariant with respect to interchange of the faces. This would be equivalent to specifying a probability over

$$\Delta_6^* = \{(p_1^*, p_2^*, \ldots, p_6^*): p_1^* \geqslant \cdots \geqslant p_6^* \geqslant 0, \sum_{j=1}^{6} p_j^* = 1\}$$

Specifying such a probability would be to say we have opinions about the bias of the die (the most likely face has probability p_1^*, the second most likely p_2^*, and so on), but not about *which* face has the bias, since our probability function is symmetric with respect to faces.

A little thought should make it clear that the frequencies of frequencies can provide information relevant to the prior $d\mu$ on Δ_t^* (in the partition exchangeable case). For example, suppose that we know the die is biased in favor of one face, and that the other faces are equally likely. Then, the unknown vector of ordered probabilities satisfies $p_1^* > p_2^* = p_3^* = \cdots = p_6^*$. Suppose now that in 100 trials we observe the frequency vector (20, 16, 16, 16, 16, 16). Then, we would guess that $p_1 = p_1^* = .2$ (approximately), and $p_2 = p_2^* = p_3 \cdots = p_6^* = .16$. But, if we observed the frequency vector (20, 40, 10, 15, 10, 5), we would guess $p_2 = p_1^* = .4$, and $p_1 = p_2^* = (20 + 10 + 15 + 10 + 5)/\{(100)(5)\} = .12$. Our estimate for p_1 differs in the two cases (.16 vs. .12) despite the fact that the frequency count for the first category is the same in both cases.

This is clearly then an objection to Johnson's sufficientness postulate (and, thus, also Carnap's continuum of inductive methods): although on the surface it appears to be a reasonable quantification of a state of ignorance about individual categories, it asserts that the frequencies of the frequencies lack relevant information about the probabilities of

those categories. Nevertheless, as the example demonstrates, it is certainly possible to have degrees of belief which are category symmetric, and yet for which the frequencies of frequencies provide very real information. This far from obvious fact was apparently first noted by the brilliant English mathematician Alan Turing during World War II (see Good 1965, p. 68; 1979).

In general, the predictive probabilities for partition exchangeable probabilities will have the form

$$P[X_{N+1} = c_i | X_1, X_2, \ldots, X_N] = f(n_i, a_0, a_1, \ldots, a_N).$$

Johnson's sufficientness postulate thus makes the very strong supposition that the predictive probabilities reduce to a function $f(n_i, N)$. In a very interesting paper, Hintikka and Niiniluoto (1980) explore the consequences of the weaker assumption that the predictive probabilities are functions $f(n_i, a_0, N)$; that is, these may also depend on the number of categories which are thus far unobserved. This generalization of Johnson's postulate seems very natural within the context of partition exchangeability, but it is unclear why the dependence on the partition vector should be limited to only its first component. Ultimately it is only partition exchangeability which exactly captures the notion of complete ignorance about categories; any further restriction on a probability beyond that of category symmetry necessarily involves, at least implicitly, some assumption about the categories. The temptation to do so, of course, is understandable; unlike the continuum of inductive methods, the partition exchangeable probabilities do not form a finite-dimensional family, which can be described by a finite number of parameters.

NOTE: In general there are t! permutations of the set of integers $\{1, 2, \ldots, t\}$; and to every such permutation there corresponds a subsimplex $\Delta_{t,\sigma}$ of Δ_t, namely, $\Delta_{t,\sigma} = \{(p_1, p_2, \ldots, p_t) \in \Delta_t: p_{\sigma(1)} \geq p_{\sigma(2)} \geq \cdots \geq p_{\sigma(t)}\}$. The map $(p_1, p_2, \ldots, p_t) \rightarrow (p_{\sigma(1)}, p_{\sigma(2)}, \ldots, p_{\sigma(t)})$ defines a homeomorphism of $\Delta_{t,\sigma}$ onto Δ_t^*, and this map permits one to transfer the values of a prior $d\mu$ on Δ_t^* to the subsimplex $\Delta_{t,\sigma}$.

5. EXCHANGEABLE RANDOM PARTITIONS

Now we come to the major point of this paper. How can a Bayesian allow for (1) infinite categories, or (2) unknown species?

If the number of categories is infinite, then no prior can be category symmetric, for such a prior would have to assign equal weight to each category, which is impossible; that is, if there are an infinite number of colors (say), $c_1, c_2, \ldots$, then $P[X_1 = c_1] = P[X_2 = c_2] = \cdots 1/t$, which is impossible, since $t = \infty$. We are thus compelled to consider probability assignments which contain some element of asymmetry between the different categories.

But, more seriously, *what* does it even mean to assign probabilities in a situation where we are encountering previously unknown species, continuously observing new and possibly unsuspected kinds? According to (at least one naive version of) the classical Bayesian picture, one assigns probabilities in advance to all possible outcomes and, then, updates via Bayes's theorem as new information comes in. How can one introspect and assign probabilities when the possible outcomes are unknown beforehand?

The earlier discussion of partition exchangeable sequences suggests a solution to this second difficulty: rather than focus on the probability of a sequence of outcomes $(e_1, e_2, \ldots, e_N)$, or the probability of a frequency vector $(n_1, n_2, \ldots, n_t)$ (the elements of which refer to specific species), focus instead on the partition vector $(a_1, a_2, \ldots, a_N)$ and its probability. Even if one does not know *which* species are present prior to sampling, one can still have beliefs as to the *relative abundances* in which those species, as yet unobserved, will occur. (Note that in this setting a_0 is excluded from the partition vector: lacking prior knowledge as to the totality of species present, it is impossible to specify at any given stage how many species present do not yet appear in the sample.)

One could, in fact, now proceed exclusively at the level of partition vectors, and construct a theory of the type we are seeking (although it is far from obvious at this stage how to cope in a category symmetric fashion with the $t = \infty$ case discussed above). But there would appear to be a substantial cost: the rich theoretical structure of exchangeability, the representation theorem, ignorance priors, and the like. One need not despair, however. All this can be obtained, *provided* one looks at the matter in a new, if initially somewhat unorthodox manner.

5.1. *Exchangeable Random Partitions*

The key point is to recognize that in the sampling of species scenario, the relevant information being received is an *exchangeable random partition*. Because the individual species do not, in effect, have an

individuality – we simply observe the first species, then at some subsequent time the second, at a still later time the third, and so on – the relevant information being received is a *partition* of the integers.

In other words, the first species occurs at some set of times

$$A_1 =: \{t_1^1, t_1^2, t_1^3, \ldots : t_1^1 < t_1^2 < t_1^3 < \cdots\}$$

where necessarily $t_1^1 = 1$, and in general the set A_1 may only contain a finite number of times even if an infinite number of observations is made (this will happen if the first species is only observed a finite number of times, possibly even only once, in which case $A_1 = \{t_1^1\}$). Likewise, the second species occurs at some set of times

$$A_2 =: \{t_2^1, t_2^2, t_2^3, \ldots : t_2^1 < t_2^2 < t_2^3 < \cdots\}$$

where necessarily t_2^1 is the first positive integer not in A_1, and A_2 may again be finite. In general, the i-th species to be observed occurs at some set of times $A_i = \{t_i^j : j = 1, 2, 3, \ldots\}$ and the collection of sets $A_1, A_2, A_3, \ldots$ forms a *partition* of the positive integers $\mathbf{N}$ in the sense that

$$\mathbf{N} = A_1 \cup A_2 \cup A_3 \cup \cdots \quad \text{and} \quad A_i \cap A_j = \emptyset,\ i \neq j.$$

In the example considered before, the partition of $\{1, 2, 3, \ldots 10\}$ generated is

$$\{1, 7\} \cup \{2, 5, 10\} \cup \{3\} \cup \{4, 8, 9\} \cup \{6\}.$$

Note a new interpretation we can now give the partition vector $\mathbf{a} = (a_1, a_2, \ldots, a_{10})$: it records the sizes of the sets in the partition and the number of species observed. Thus, in our example, given the partition vector is $(2, 1, 2, 0, \ldots, 0)$, two sets in the resulting partition have a single element (since $a_1 = 2$), one set in the partition has two elements (since $a_2 = 1$), two sets in the partition have three elements (since $a_3 = 2$), and the total number of species observed is 5 (since $a_1 + a_2 + \cdots + a_{10} = 5$). Although originally defined in terms of the underlying sequence, the partition vector is a function solely of the resulting partition of the time set; and one can therefore refer to the partition vector of a partition.

Thus, observing the successive species in our sample generates a random partition of the positive integers. Now let us consider in what sense such a partition could be 'exchangeable'. An obvious idea is to

examine the structure of random partitions arising from exchangeable sequences and see if we can characterize them in some way.

This turns out to be relatively simple: *if a random sequence is exchangeable, then the partition structures for two possible sequences have the same probability whenever they have the same partition vector* **a**.

In order to see this, let's think about what happens to a partition when we permute the categories or times of the underlying sequence which generates it. Consider our earlier example of the sequence $\{5, 2, 6, 1, 2, 3, 5, 1, 1, 2\}$, and suppose we permute the category index in some way, say, the cyclic permutation

$$1 \to 2 \to 3 \to 4 \to 5 \to 6 \to 1.$$

Then, our original sequence becomes transformed into {6, 3, 1, 2, 3, 4, 6, 2, 2, 3}, and the resulting partition of the time set from 1 to 10 is the same as before: species 6 occurs at times 1 and 7, hence, we get $A_1 = \{1, 7\}$, and so on. *Permuting the category index results in a new sequence but leaves the resulting partition unchanged.*

Next, suppose we were to permute the times, say, by the cyclic permutation

$$1 \to 2 \to 3 \to 4 \to 5 \to 6 \to 7 \to 8 \to 9 \to 10 \to 1.$$

(That is, what happened at time 1 is observed to occur at time 2 instead; at time 2, at time 3 instead; and so on.)

Then, our original sequence becomes transformed into {2, 5, 2, 6, 1, 2, 3, 5, 1, 1}, and we get a new partition of the time set, namely,

$$\{1, 3, 6\} \cup \{2, 8\} \cup \{4\} \cup \{5, 9, 10\} \cup \{7\}.$$

Because of the exchangeability of the underlying sequence, this new partition has the same probability of occurring as the original one. Note that it has the same frequency vector **n** and, therefore, partition vector **a**. This observation is the one underlying the idea of an exchangeable random partition:

DEFINITION: A random partition is *exchangeable* if any two partitions π_1 and π_2 having the same partition vector have the same probability; i.e., if

$$\mathbf{a}(\pi_1) = \mathbf{a}(\pi_2) \Rightarrow P[\pi_1] = P[\pi_2].$$

5.2. *The Kingman Representation Theorem*

In the case of sequences, the de Finetti representation theorem states that the general exchangeable sequence can be constructed out of elementary building blocks: Bernoulli trials (coin-tossing sequences) in the case of 0,1-valued random variables; multinomial trials in the case of t-valued random variables; and in general sequences of independent- and identically-distributed random variables. The corresponding building blocks of the general exchangeable random partition are the *paintbox processes*.

In order to construct a paintbox process, consider an ordered 'defective' probability vector

$$\mathbf{p} = (p_1, p_2, p_3, \ldots), \quad \text{where} \quad p_1 \geq p_2 \geq p_3 \cdots \geq 0 \quad \text{and} \quad p_1 + p_2 + p_3 + \cdots \leq 1,$$

and let ∇ denote the infinite simplex of all such vectors.

Given such a defective probability vector $\mathbf{p} = (p_1, p_2, p_3, \ldots) \in \nabla$, let $p_0 =: 1 - \Sigma_i p_i$; and let $\mu_{\mathbf{p}}$ be a probability measure on the unit interval $[0, 1]$ having point masses p_j at some set of distinct points x_j, $j \geq 1$ (which points are selected doesn't matter), and a continuous component assigning mass p_0 to $[0, 1]$. Call such a probability measure a *representing probability measure* for $\mathbf{p}$. Let $X_1, X_2, X_3, \ldots$ be a sequence of independently- and identically-distributed random variables with common distribution $\mu_{\mathbf{p}}$, and consider the exchangeable random partition generated by the rule:

$$A_j = \{i: X_i = x_j\} \text{ where } A_1 \cup A_2 \cup \cdots = \{1, 2, \ldots, N\}.$$

That is, partition the integers $1, 2, \ldots, N$ by grouping together those times i at which the random variables X_j have a common value x_j.

It is then not difficult to see that if $\mathbf{p} \in \nabla$, and $\mu_{\mathbf{p}}$ and $\nu_{\mathbf{p}}$ are two different representing probability measures for $\mathbf{p}$, then $\mu_{\mathbf{p}}$ and $\nu_{\mathbf{p}}$ generate the same exchangeable random partition Π, in the sense that the two random partitions have the same stochastic structure. Thus, we have a well-defined rule for associating exchangeable random partitions with vectors in ∇: given $\mathbf{p}$, select $\mu_{\mathbf{p}}$, and use $\mu_{\mathbf{p}}$ to generate Π. Let's call this resulting exchangeable random partition $\Pi_{\mathbf{p}}$.

Now we are ready to state the Kingman representation theorem:

THEOREM (Kingman 1978): The general exchangeable random partition is a mixture of paintbox processes.

Let us make this precise. Suppose that $Z_1, Z_2, Z_3, \ldots$ is a sequence of random partitions; specifically, for each $N \geq 1$, Z_N is an exchangeable random partition of $\{1, 2, \ldots, N\}$. There is an obvious sense in which such a sequence is consistent. Namely, any partition of $\{1, 2, \ldots, N+1\}$ gives rise to a partition of $\{1, 2, \ldots, N\}$ by simply omitting the integer $N+1$ from the subset in which it occurs. Let $T_{N+1,N}$ denote the map which performs this operation. Then, the pair Z_{N+1} and Z_N are *consistent* if $P[Z_N \in A] = P[T_{N+1,N}(Z_{N+1}) \in A]$, where A is a set of partitions of $\{1, 2, \ldots, N\}$; and the sequence is consistent if Z_N and Z_{N+1} are consistent for every $N \geq 1$. Every such consistent sequence gives rise to a probability measure on the partitions of $\mathbf{N} =: \{1, 2, 3, \ldots\}$. If Π is the probability distribution on the partitions of $\mathbf{N}$ arising from such an arbitrary exchangeable random partition, then the Kingman representation theorem states that there exists a (unique) probability measure $d\mu$ on ∇, the infinite simplex of all ordered defective probability vectors, such that for every (measurable) set A of partitions,

$$\Pi(A) = \int_{\nabla} \Pi_{\mathbf{p}}(A)\, d\mu(\mathbf{p}).$$

Note that instead of integrating over the probability simplex, one integrates over the ordered defective probability simplex ∇ consisting of all possible defective probability vectors $\mathbf{p}$. Moreover, as proven by Kingman, the ordered sample frequencies arising from the random partition converge in joint distribution to the mixing measure $d\mu$. (Just as in de Finetti's theorem the *unordered* sample frequencies $(n_1/N, \ldots, n_t/N)$ converge to the mixing measure $d\mu$, in the de Finetti representation, here the *ordered* sample frequencies converge to the mixing measure $d\mu$ in the Kingman representation.)

The distinctive role that the *continuous* component p_0 of a paintbox process plays in the theorem deserves some comment. When Kingman first investigated exchangeable random partitions, he was puzzled by the fact that mixtures over the *discrete* nondefective ordered probabilities $(p_1^*, p_2^*, p_3^*, \ldots)$ generated many, but by no means all possible exchangeable random partitions. The key to this puzzle is the far from

obvious observation that when a new species appears, it must always suffer one of two fates: either it never appears again, or it is subsequently seen an infinite number of times. No intermediate fate is possible. The species that arise once and only once are precisely those that arise from the continuous component.

The Reverend Dr. Richard Price would not have found this surprising. As he states (Bayes 1764, p. 312), the first appearance of an event only informs us of its *possibility*, but would not "give us the least reason to apprehend that it was, in that instance or in any other, regular rather than irregular in its operations"; that is, we are given no reason to think that its probability of recurring is positive (read "regular") rather than 0 (read "irregular"). In effect, Price is saying that the first observation tells us that the outcome lies in the *support* of the unknown representing probability μ_p, while the second observation tells us that it lies in the *discrete component* of this probability.

5.3. *The Poisson–Dirichlet Process*

Thus far we have managed to capture a notion of exchangeable random outcome suitable to the sampling of species setting, and have a representation theorem as well. But the classical theories of induction that employ probability theory usually attempt to go further and identify classes of possible priors $d\mu$ thought to represent situations of limited information. In the de Finetti representation discussed earlier, this was easy: the so-called flat priors dp or $dp_1 dp_2 \ldots dp_{t-1}$ immediately suggested themselves, and the game was to come up with characterizations of these priors in terms of symmetry assumptions about the underlying cylinder set probabilities. Here, however, it is far from apparent what a 'flat' prior would be.

At this point we encounter a deep and truly ingenious idea of Kingman's. Let $\alpha > 0$. Suppose we took a symmetric Dirichlet prior $\mathbf{D}(\alpha)$ on the t-simplex Δ_t and let the number of categories tend to infinity (i.e., let $t \to \infty$). The resulting probabilities would then 'wash out'. For any fixed $t_0 < \infty$ and $(x_1, x_2, \ldots x_{t_0}) \in \Delta_t$, the cylinder set probabilities are:

$$P_{\alpha,t}[p_1 \leq x_1, p_2 \leq x_2, \ldots, p_{t_0} \leq x_{t_0}] \to 0 \quad \text{as } t \to \infty.$$

But, suppose instead that we consider the vector of ordered probabil-

ities. Then, something truly remarkable occurs. Since we can map the t-simplex Δ_t onto the ordered t-simplex Δ_t^* (by associating to any vector $(p_1, p_2, \ldots, p_t)$ its ordered rearrangement $(p_1^*, p_2^*, \ldots, p_t^*)$), the symmetric Dirichlet prior on Δ_t induces a probability distribution on Δ_t^*: for any fixed $t_0 \leqslant t < \infty$ and sequence $(x_1^*, x_2^*, \ldots, x_{t_0}^*) \in \Delta_t^*$, there is a corresponding cylinder set probability

$$P_{\alpha,t}[p_1^* \leqslant x_1^*, p_2^* \leqslant x_2^*, \ldots, p_{t_0}^* \leqslant x_{t_0}^*].$$

Then, as Kingman shows, if $t \to \infty$ *and* $\alpha \to 0$ in such a way that $t\alpha \to \theta > 0$, for some positive number θ, then the resulting sequence of probabilities does not 'wash out': instead, it has a proper limiting distribution. And, since this is so for each t, the result is a probability measure on ∇. (A 'consistent' set of probabilities on the finite cylinder sets always corresponds to a unique probability on infinite sequence space.) This is called the *Poisson–Dirichlet distribution* (with parameter θ). (The terminology is intended to suggest an analogy with the classical Poisson-binomial limit theorem in probability theory.)

A simple example will illustrate the phenomenon. Suppose you pick a point $\mathbf{p}$ at random from Δ_t according the symmetric Dirichlet distribution $P_{\alpha,t}$ and ask for the probability $P_{\alpha,t}[p_1 \geqslant x_1]$. As $t \to \infty$, this probability tends to 0 (since a typical coordinate of $\mathbf{p}$ will be small if t is large). But suppose, instead, you ask for the probability that the *maximum* coordinate of $\mathbf{p}$ exceeds x_1: that is, $P_{\alpha,t}[p_1^* \geqslant x_1]$. Then, Kingman's theorem states that this probability has a *nonzero* limit as $t \to \infty$. Such a result, although hardly obvious, is evidently neither counterintuitive nor paradoxical.

5.4. *The Ewens Sampling Formula*

Since the Poisson–Dirichlet distribution with parameter θ is a probability measure on ∇, and each paintbox process in ∇ gives rise to an exchangeable random partition, for every sample size N the Poisson–Dirichlet distribution induces a probability distribution $P[a_1, a_2, \ldots, a_N]$ on the set of possible partition vectors. Kingman shows that these probabilities are given by the so-called

EWENS SAMPLING FORMULA:

$$\frac{n!}{\theta(\theta+1)\ldots(\theta+n-1)}\prod_{r=1}^{n}\frac{\theta^{a_r}}{r^{a_r}a_r!}$$

This little formula turns out to be remarkably ubiquitous: it is called the Ewens sampling formula, because it was first discovered by the geneticist Warren Ewens in the course of his work in theoretical population genetics (Ewens 1972). It crops up in a large number of seemingly unrelated contexts. One example of many is: if one picks a random permutation of the integers $\{1, 2, \ldots, N\}$, and lets a_j denote the number of j-cycles, then the probability distribution for $a_1, a_2, \ldots, a_N$ is provided by the Ewens formula.

Given the Ewens formula for the cylinder set probabilities, it is a simple calculation to derive the corresponding predictive probabilities or rules of succession. It is important, however, to be clear what this means, so let's back up for a moment. Suppose we are performing a sequence of observations $X_1, X_2, \ldots, X_N, \ldots$, noting at each stage either the species of an animal, the next word used by Shakespeare, or whatever. At each point, we observe either a species previously observed or an entirely new species. Before these are observed, it doesn't make sense to refer to these outcomes as exchangeable; in fact, it doesn't even make sense to refer to the probabilities of such outcomes, because ahead of time we don't know what a complete list of possible outcomes is. We're learning as we go along. But at time N we can construct a partition of $\{1, 2, \ldots, N\}$ on the basis of what we've seen thus far, and it *does* make sense to talk prospectively about the probability of seeing a particular partition. It is then natural to assume that the resulting random partition is exchangeable; it is necessary to tutor one's intuition, but this is the end result. (As Diaconis and Freedman (1980, p. 248) observe about the concept of Markov exchangeability, "the notion of symmetry seems strange at first . . . . A feeling of naturalness only appears after experience and reflection".) Having arrived at this epistemic state, we can then invoke the Kingman representation theorem, and write our exchangeable random partition as a mixture of paintbox processes. Although we do not, indeed cannot, have prior beliefs about the probabilities of the species we observe, since we didn't know they existed until we saw them, we can certainly have opinions

about their *abundances*: that is, what is the frequency of occurrence of the most abundant species, the second most abundant, and so on, and this is what our prior on ∇ summarizes.

Now, given that we make a series of N observations, it is clear that our exchangeable probability assignment will predict whether a new species will be observed on the next trial. And, if we don't observe a new species, whether we see a member of the same species as the very first animal observed. (That is, whether the new partition resulting after time N + 1 will add the integer N + 1 to the member of the partition containing 1.) Or, whether a member of the second species observed. (That is, whether the new partition adds N + 1 to that member of the partition containing the first integer not in the member of the partition containing 1.) And so on.

Given that we have observed a number of species so far – with n_1 of the first type, n_2 of the second, and so on – *what* are the resulting succession probabilities for observing one of the known species or an unknown one? The answer, given the Poisson–Dirichlet prior (and letting s_j denote the j-th species observed to date) is:

$$P[X_{N+1} = s_j \,|\, \mathbf{n}] = \frac{n_j}{(N + \theta)}$$

That is, with $\theta = 1$ and $t = 0$ *the answer is identical to De Morgan's*!

Thus, De Morgan's answer emerges as far from arbitrary. It arises from the canonical 'ignorance prior' for exchangeable random partitions.

5.5. *The Chinese Restaurant Process*

Completing our analogy with the case of exchangeable sequences, what is the generating urn process for this 'benchmark' process? We already know the answer to this: it is a classical urn model with the added facet of a black ball representing the 'mutator'.

This process has in fact been independently noted several times during the last two decades. Perhaps the most attractive version is the *Chinese restaurant process*: on any given evening in Berkeley, a large number of people go to some Chinese restaurant in the downtown area. As each person arrives, he looks in the window of each restaurant to

decide whether or not to go inside. His chance of going in increases with the number of people already seen to be inside, since he takes that as a good sign. But there's always some probability that he goes to an empty restaurant. In a second (and, in fact, the original) version of the process, people enter a single restaurant and sit down at random at one of several circular tables (see Aldous 1985, p. 92). (The main point of this version is that the groups around the tables define the cycles of a random permutation.)

6. SOME FURTHER LITERATURE

The problem discussed above is often referred to in the statistical literature as the *sampling of species problem*. One of the earliest references is a short but important paper by Fisher (Fisher et al., 1943). The sampling of species problem has since been considered by several people from a Bayesian perspective. As noted earlier, Turing seems to have been the first to realize the potential informativeness of the frequencies of the frequencies, a discovery he made during the course of his cryptanalytic work at Bletchley Park during World War II. The noted Bayesian statistician I. J. Good was Turing's statistical assistant at the time, and after the war he published a series of interesting papers in this area (see, e.g., Good 1953; Good and Toulmin 1956; and Good 1965, chapter 8). These methods have recently been employed to estimate the total number of words known to Shakespeare (Efron and Thisted 1976), and to test whether a poem attributed to Shakespeare was in fact authored by him (Thisted and Efron 1987). During the last two decades the American statistician Bruce Hill has also investigated the sampling of species problem (see, e.g., Hill 1968, 1979). *Zipf's law* is an empirical relationship that the elements of a partition vector are often found to follow (see Hill 1970). Hill (1988) discusses some relationships between his own methods and those of Kingman.

Kingman's beautiful work is summarized in his monograph, *The Mathematics of Genetic Diversity* (1980; see also Kingman 1975). Kingman's theory was originally stated in terms of "partition structures" (Kingman 1978a), as was his original proof of the representation theorem for exchangeable random partitions (Kingman 1978b). The account given above draws heavily on Aldous (1985, pp. 85–92). The Ewens sampling formula was of course discovered by Ewens (1972); it thus provides a counterexample to Stigler's law of eponomy, but it was

also independently discovered shortly after by Charles Antoniak in a Bayesian setting (Antoniak 1974). The urn model discussed in Section 2 is implicit in De Morgan (1838, 1845), but was never formally stated by him. During the 1970s the model surfaced in Berkeley, first as a special case of a class of urn models discussed by Blackwell and MacQueen (1973) and, then, in the guise of the Chinese restaurant process (fathered by Lester Dubins and Jim Pitman). The CRP remained 'folklore', however, until it was described in Aldous's 1985 monograph. The urn model itself became more widely known after 1984, when Fred Hoppe drew attention to it as a simple method of generating the Ewens sampling formula (see Hoppe 1984, 1987; and Donnelly 1986).

An axiom corresponding to the assumption of partition exchangeability is briefly mentioned by Carnap at the beginning of his book (Carnap 1950), but not pursued further by him. Good has studied priors for multinomial probabilities which are mixtures of symmetric Dirichlet priors (and therefore partition exchangeable); there is a close relationship between some of his work (Good 1953) and recent efforts by Theo Kuipers (1986) to estimate the λ-parameter in Carnap's continuum of inductive methods (equivalently, the α-parameter of the corresponding symmetric Dirichlet prior). Kuipers had earlier discussed a mutation model similar to De Morgan's, but in his system the mutation rate does not tend to zero (see Kuipers 1973).

The concept of exchangeability was introduced into the philosophical literature by Johnson, who termed it the "permutation postulate", and analyzed its consequences assuming first the combination postulate (Johnson 1924) and then the less restrictive sufficientness postulate (Johnson 1932). Exchangeability was soon after independently discovered by de Finetti, who skillfully employed his representation theorem to analyze the structure of the general exchangeable sequence, making no appeal to additional, restrictive postulates. After World War II, Carnap investigated exchangeability as part of a broad attack on the problem of inductive inference, rediscovering many of Johnson's results and carrying his investigations into new territory (see, especially, Carnap 1980).

It is an important historical footnote that Carnap clearly recognized the importance of studying the case of inductive inference when the number of categories is not fixed in advance, and thought that this could be done by employing 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 any further, however, because he judged that it would introduce further complexities into the analysis, which would have been premature given the relatively primitive state of the subject at that time. (My thanks to Richard Jeffrey, to whom I owe the information in this paragraph.)

As we can now appreciate, Carnap displayed great prescience here: the use of such an equivalence relation would have been tantamount to considering partitions rather than sequences, and the resulting complexities are indeed an order of magnitude greater. That we can now see further today is a tribute to the beautiful and profound work of Kingman discussed above.

ACKNOWLEDGMENT

I thank Domenico Costantini, Persi Diaconis, Ubaldo Garibaldi, Tom Nagylaki, and Jim Pitman for helpful discussions and references, and Richard Jeffrey for his comments on a draft of the paper.

REFERENCES

Aldous, D. J.: 1985, 'Exchangeability and Related Topics', in P. L. Hennequin (ed.), *École d'Été de Probabilités de Saint-Flour XIII – 1983*, *Lecture Notes in Mathematics* **1117**, 1–198.

Antoniak, C. E.: 1974, 'Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems', *Annals of Statistics* **2**, 1152–74.

Bayes, T.: 1764, 'An Essay Towards Solving a Problem in the Doctrine of Chances', *Philosophical Transactions of the Royal Society of London* **53**, 370–418 (reprinted: 1958, *Biometrika* **45**, 293–315 (page citations in the text are to this edition)).

Blackwell, D. and MacQueen, J. B.: 1973, 'Ferguson Distributions via Polya Urn Schemes', *Annals of Statistics* **1**, 353–55.

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

Carnap, R.: 1980, 'A Basic System of Inductive Logic, Part II', in R. C. Jeffrey (ed.), *Studies in Inductive Logic and Probability*, Vol. 2, University of California Press, Berkeley and Los Angeles, pp. 7–155.

De Finetti, B.: 1937, 'La prevision: ses lois logiques, ses sources subjectives', *Annales de l'Institut Henri Poincaré* **7**, 1–68.

De Morgan, Augustus: 1838, *An Essay on Probabilities, and on their Application to Life Contingencies and Insurance Offices*, Longman et al., London.

De Morgan, A.: 1845, 'Theory of Probabilities', in *Encyclopedia Metropolitana, Volume 2: Pure Mathematics*, B. Fellowes et al., London, pp. 393–490.

Diaconis, P. and Freedman, D.: 1980, 'De Finetti's Generalizations of Exchangeability', in R. C. Jeffrey (ed.), *Studies in Inductive Logic and Probability*, Vol. 2, University of California Press, Berkeley and Los Angeles, pp. 233–50.

Donnelly, P.: 1986, 'Partition Structures, Polya Urns, the Ewens Sampling Formula, and the Ages of Alleles', *Theoretical Population Biology* **30**, 271–88.

Efron, B. and Thisted, R.: 1976, 'Estimating the Number of Unseen Species: How Many Words did Shakespeare Know?', *Biometrika* **63**, 435–47.

Ewens, W. J.: 1972, 'The Sampling Theory of Selectively Neutral Alleles', *Theoretical Population Biology* **3**, 87–112.

Feller, William: 1968, *An Introduction to Probability Theory and its Applications*, Vol. 1, 3d ed., Wiley, New York.

Fisher, R. A., Corbet, A. S. and Williams, C. B.: 1943, 'The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population', *Journal of Animal Ecology* **12**, 42–58.

Good, I. J.: 1953, 'On the Population Frequencies of Species and the Estimation of Population Parameters', *Biometrika* **40**, 237–64.

Good, I. J. and Toulmin, G. H.: 1956, 'The Number of New Species, and the Increase in Population Coverage, When a Sample is Increased', *Biometrika* **43**, 45–63.

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

Good, I. J.: 1979, 'Turing's Statistical Work in World War II', *Biometrika* **66**, 393–96.

Hill, B.: 1968, 'Posterior Distribution of Percentiles: Bayes's Theorem for Sampling from a Finite Population', *Journal of the American Statistical Association* **63**, 677–91.

Hill, B.: 1970, 'Zipf's Law and Prior Distributions for the Composition of a Population', *Journal of the American Statistical Association* **65**, 1220–32.

Hill, B.: 1979, 'Posterior Moments of the Number of Species in a Finite Population, and the Posterior Probability of Finding a New Species', *Journal of the American Statistical Association* **74**, 668–73.

Hill, B.: 1988, 'Parametric Models for A_n: Splitting Processes and Mixtures', unpublished manuscript.

Hintikka, J. and Niiniluoto, I.: 1980, 'An Axiomatic Foundation for the Logic of Inductive Generalization', in R. C. Jeffrey (ed.), *Studies in Inductive Logic and Probability*, Vol. 2, University of California Press, Berkeley and Los Angeles, pp. 157–82.

Hoppe, F.: 1984, 'Polya-Like Urns and the Ewens Sampling Formula', *Journal of Mathematical Biology* **20**, 91–94.

Hoppe, F.: 1987, 'The Sampling Theory of Neutral Alleles and an Urn Model in Population Genetics', *Journal of Mathematical Biology* **25**, 123–59.

Jeffrey, R. C. (ed.): 1980, *Studies in Inductive Logic and Probability*, Vol. 2, University of California Press, Berkeley and Los Angeles.

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

Johnson, William Ernest: 1932, 'Probability: the Deductive and Inductive Problems', *Mind* **49**, 409–23.

Kingman, J. F. C.: 1975, 'Random Discrete Distributions', *Journal of the Royal Statistical Society* **B37**, 1–22.

Kingman, J. F. C.: 1978a, 'Random Partitions in Population Genetics', *Proceedings of the Royal Society* **A361**, 1–20.

Kingman, J. F. C.: 1978b, 'The Representation of Partition Structures', *Journal of the London Mathematical Society* **18**, 374–80.

Kingman, J. F. C.: 1980, *The Mathematics of Genetic Diversity*, SIAM, Philadelphia.
Kuipers, T. A. F.: 1973, 'A Generalization of Carnap's Inductive Logic', *Synthese* **25**, 334–36.
Kuipers, T. A. F.: 1986, 'Some Estimates of the Optimum Inductive Method', *Erkenntnis* **24**, 37–46.
Laplace, P. S., Marquis de: 1781, 'Mémoire sur les probabilités', *Mem. Acad. Sci. Paris 1778*, 227–32 (*Oeuvres complètes*, Vol. 9, pp. 383–485).
Quetelet, A.: 1846, *Lettres à S.A.R. le Duc Régnant de Saxe-Cobourg et Gotha, sur la théorie des probabilités, appliquée aux sciences morales et politiques*, Hayez, Brussels.
Thisted, R. and Efron, B.: 1987, 'Did Shakespeare Write a Newly-Discovered Poem?', *Biometrika* **74**, 445–55.
Zabell, S. L.: 1982, 'W. E. Johnson's "Sufficientness" Postulate', *Annals of Statistics* **10**, 1091–99.
Zabell, S. L.: 1988, 'Symmetry and its Discontents', in B. Skyrms and W. L. Harper (eds.), *Causation, Chance, and Credence*, Vol. 1, Kluwer, Dordrecht, pp. 155–90.
Zabell, S. L.: 1989, 'The Rule of Succession', *Erkenntnis* **31**, 283–321.

Department of Mathematics
Northwestern University
Lunt Hall
Evanston, IL 60208
U.S.A.