WILLIAM ERNEST JOHNSON 1858–1931
(Photograph courtesy of King's College Library Cambridge, England.)

The Annals of Statistics
1982, Vol. 10, No. 4, 1091-1099

W. E. JOHNSON'S "SUFFICIENTNESS" POSTULATE[1]

BY SANDY L. ZABELL

Northwestern University

How do Bayesians justify using conjugate priors on grounds other than mathematical convenience? In the 1920's the Cambridge philosopher William Ernest Johnson in effect characterized symmetric Dirichlet priors for multinomial sampling in terms of a natural and easily assessed subjective condition. Johnson's proof can be generalized to include asymmetric Dirichlet priors and those finitely exchangeable sequences with linear posterior expectation of success. Some interesting open problems that Johnson's result raises, and its historical and philosophical background, are also discussed.

1. Introduction. In 1932 a posthumously published article by the Cambridge philosopher W. E. Johnson showed how symmetric Dirichlet priors for infinitely exchangeable multinomial sequences could be characterized by a simple property termed "Johnson's sufficiency postulate" by I. J. Good (1965). (Good (1967) later shifted to the term "sufficientness" to avoid confusion with the usual statistical meaning of sufficiency.) Johnson could prove such a result, prior to the appearance of de Finetti's work on exchangeability and the representation theorem, for Johnson had himself already invented the concept of exchangeability, dubbed by him the "permutation postulate" (see Johnson, 1924, page 183). Johnson's contributions were largely overlooked by philosophers and statisticians alike until the publication of Good's 1965 monograph, which discussed and made serious use of Johnson's result.

Due perhaps in part to the posthumous nature of its publication, Johnson's proof was only sketched and contains several gaps and ambiguities; the major purpose of this paper is to present a complete version of Johnson's proof. This seems of interest both because of the result's intrinsic importance for Bayesian statistics and because the proof itself is a simple and elegant argument which requires little technical apparatus. Furthermore, it can be easily generalized to characterize both asymmetric Dirichlet priors and finitely exchangeable sequences with posterior expectation of success linear in the frequency count, and the proof below is given in this generality.

After sketching the background to Johnson's result in Section 1, the generalization of his proof mentioned above is given in Section 2. Section 3 discusses a number of complements to the result and some open problems it raises, and Section 4 concludes with a historical note on Johnson and the reception of his work in the philosophical literature.

1. The Bayesian background. Let $X_1, X_2, \cdots$ be an infinite exchangeable sequence of 0's and 1's (to be thought of as indicators of some event E), and let $S_N = X_1 + \cdots + X_N$. Then, as first shown by de Finetti, it follows from exchangeability that the limiting frequency

$$Z = \lim_{N\to\infty} S_N/N \tag{1.1}$$

exists almost surely, and that

Received March 1982; revised June 1982.

[1] This research was supported by Office of Naval Research Contract N00014-76-C-0475 (NR-042-267).

AMS 1980 *subject classification.* Primary 62A15; secondary 62-03, 01A60.

Key words and phrases. W. E. Johnson, sufficientness postulate, exchangeability, Dirichlet prior, Rudolph Carnap.

(1.2) $$P(S_N = k) = \binom{N}{k} \int_0^1 p^k(1-p)^{N-k}\, dF(p)$$

for every $N \geq 1$ and $0 \leq k \leq N$, where $F(p) = P(Z \leq p)$ is the cumulative distribution function of Z. If the parameter p is thought of as a propensity or "objective probability," then dF may be regarded as the degree of belief about or "subjective probability" of the true value of p.

Traditionally, the "flat" prior $dF(p) = dp$ was taken to express "complete ignorance" about p, or the likelihood of the event E (for which the X_i serve as indicators). Bayes's own justification for this was to take $P(S_N = k) = (N+1)^{-1}$ as quantifying complete ignorance about E, observe that (1.2) gave precisely this result (for all k and N) when $dF(p) = dp$, and then conclude that $dF(p)$ is dp. (The argument can be made rigorous by noting that dF is uniquely determined by its moments; see, e.g., Murray (1930); Edwards (1974, 1978). Stigler (1982) traces how Bayes's argument was systematically distorted by later statisticians to fit their own foundational preconceptions.) Laplace justified the choice somewhat more directly by invoking the so-called principle of insufficient reason.

This principle came under strong criticism during the latter part of the 19th century (most notably by Boole, Venn, and Chrystal; unfortunately, Fisher's account (1956, Chapter 2) of their reservations is seriously flawed; see Zabell, 1982). Some advocates of the principle's use (Edgeworth, 1884, page 230; Pearson, 1907) adopted the position that taking $dF(p) = dp$ was often approximately justifiable on the basis of experience and background information; a position which suggests that other priors might equally well express and quantify states of knowledge previous to the receipt of sampling data. It was against this background that the actuary G. F. Hardy (1889) and the mathematician W. A. Whitworth (1897, pages 224–225) both proposed the class of beta priors

$$B(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \quad \alpha, \beta > 0,$$

as suitable for quantification of prior knowledge.

In 1778 Laplace proposed the obvious multinomial generalization of the Bayes-Laplace prior (Laplace, 1781, Section 33; cf. De Morgan, 1845, Sections 48–49, Bachelier, 1912, page 503, Lidstone, 1920): if $X_1, X_2, \cdots, X_N$ are the outcomes of a t-category multinomial with unknown sampling vector $\mathbf{p} = (p_1, \cdots, p_t)$, and frequency counts $\mathbf{n} = (n_1, n_2, \cdots, n_t)$, then

(1.3) $$P\{n_1, \cdots, n_t\} = \frac{N!}{\prod n_k!} \int_{\sum p_i = 1} \prod_{i=1}^{t} p_i^{n_i}\, dF(\mathbf{p})$$

with $dF(\mathbf{p}) = dp_1 dp_2 \cdots dp_{t-1}$, which implies that

(1.4) $$P\{X_{N+1} \in i\text{th category} \mid \mathbf{n}\} = \frac{n_i + 1}{N + t}.$$

In 1924, W. E. Johnson gave a justification for (1.4) parallel to that of Bayes: if all ordered t-partitions $n_1 + n_2 + \cdots + n_t$ of N are assumed to be *a priori* equally likely, then (1.4) must hold; it follows, as observed by Good (1965, page 25), that the moments of dF, and hence dF itself, are uniquely determined.

It was against this background that Johnson, not entirely satisfied with his equiprobability (or "combination") postulate, proposed another, more general one (his "sufficientness" postulate), which had the consequence of forcing dF to be a member of the Dirichlet family

(1.5) $$\mathrm{Dir}(k_1, \cdots, k_t) = p_1^{k_1-1} \cdot \ldots \cdot p_t^{k_t-1}\, dp_1 \cdot \ldots \cdot dp_{t-1},$$

$k_i > 0$ for all i.

2. Finite exchangeable sequences. Let $X_1, X_2, \cdots, X_{N+1}$ be a sequence of random variables, each taking values in the set $\mathbf{t} = \{1, 2, \cdots, t\}$, $N \geq 1$ and $t \leq \infty$, such that

$$P\{X_1 = i_1, \cdots, X_N = i_N\} > 0, \qquad \text{for all} \quad (i_1, \cdots, i_N) \in \mathbf{t}^N. \tag{2.1}$$

Let $\mathbf{n} = \mathbf{n}(X_1, \cdots, X_N)$ denote the t-vector of frequency counts, i.e., $\mathbf{n} = (n_1, n_2, \cdots, n_t)$, where $n_i = n_i(X_1, \cdots, X_N) = \#\{X_j = i\}$. Johnson's sufficientness postulate assumes that

$$P\{X_{N+1} = i \mid X_1, \cdots, X_N\} = f_i(n_i), \tag{2.2}$$

that is, the conditional probability of an outcome in the ith cell given $X_1, \cdots, X_N$ only depends on n_i, the number of outcomes in that cell previously. (Note that (2.2) is well-defined because of (2.1).) If $X_1, \cdots, X_{N+1}$ is exchangeable, $f_i(n_i) = P\{X_{N+1} = i \mid \mathbf{n}\} = P\{X_{N+1} = i \mid n_i\}$.

LEMMA 2.1. *If $t > 2$ and* (2.1), (2.2) *hold, then there exist constants $a_i \geq 0$ and b such that for all i,*

$$f_i(n_i) = a_i + bn_i. \tag{2.3}$$

PROOF. First assume $N \geq 2$. Let

$$\mathbf{n}_1 = (n_1, \cdots, n_i, \cdots, n_j, \cdots, n_k, \cdots, n_t)$$

be a fixed ordered partition of N, with i, j, k three fixed distinct indices such that $0 < n_i$, n_j and $n_i, n_k < N$, and let

$$\mathbf{n}_2 = (n_1, \cdots, n_i + 1, \cdots, n_j - 1, \cdots, n_k, \cdots, n_t)$$

$$\mathbf{n}_3 = (n_1, \cdots, n_i, \cdots, n_j - 1, \cdots, n_k + 1, \cdots, n_t)$$

$$\mathbf{n}_4 = (n_1, \cdots, n_i - 1, \cdots, n_j, \cdots, n_k + 1, \cdots, n_t).$$

Note that for any $\mathbf{n}$,

$$\textstyle\sum_{n_\ell \in n} f_\ell(n_\ell) = 1, \tag{2.4}$$

hence taking $\mathbf{n} = \mathbf{n}_1, \mathbf{n}_2, \mathbf{n}_3, \mathbf{n}_4$, we obtain

$$f_i(n_i + 1) - f_i(n_i) = f_j(n_j) - f_j(n_j - 1) = f_k(n_k + 1) - f_k(n_k) = f_i(n_i) - f_i(n_i - 1). \tag{2.5}$$

Thus

$$f_i(n_i) = a_i + bn_i,$$

where we define $a_i = f_i(0) \geq 0$ and $b = \Delta f_i(n_i)$ is independent of i (because of (2.5)).

If $N = 1$, let $c_i = f_i(1)$; it then follows from (2.4) that for any i and j, $a_i + c_j = a_j + c_i$, hence $c_i - a_i = c_j - a_j = b$. □

REMARK 1. If $t = 2$, Johnson's sufficientness postulate is vacuous and (2.3) need not hold; see Good (1965, page 26). Thus, in the binomial case, it is necessary to make the additional assumption of linearity. In either case ($t = 2$ or $t > 2$), Johnson's argument requires that $a_i > 0$; the next two remarks address this point and are both applied in Lemma 2.2 below.

REMARK 2. If (2.1) holds for $N + 1$ as well as N, then $a_i > 0$. The reader can, if he wishes, simply replace (2.1) by this strengthened version in the sequel, and ignore the following remark on a first reading.

REMARK 3. If $X_1, \cdots, X_{N+1}$ is exchangeable and (2.1) holds for N, then $a_i > 0$ if $N \geq 2$. (If $a_i = 0$ for some i, then $f_i(1) > 0$, hence $b > 0$. But if $a_i = 0$, then $f_j(N - 1) = 0$ for $j \neq i$, hence $b \leq 0$, a contradiction.) This need not hold when $N = 1$; for example, let $t = 2$ and $P(1, 1) = P(2, 2) = 1/2$. This is the reason for assuming $N_0 \geq 3$ in Theorem 2.1 below: if N_0

$= 1$ the statement is vacuous, while, if $N_0 = 2$, $k_i = 0$ can occur (unless the strengthened version of (2.1) is assumed).

Let $A = \sum_i a_i$. It follows from (2.3), (2.4) that

$$A + bN = 1, \tag{2.6}$$

hence A is finite and

$$b = (1 - A)/N. \tag{2.7}$$

Suppose $b \neq 0$. Then letting $k_i = a_i/b$ and $K = \sum k_i$, we see from (2.6) that

$$b^{-1} = N + A/b = N + K,$$

hence

$$f_i(n_i) = a_i + bn_i = \frac{k_i + n_i}{b^{-1}} = \frac{n_i + k_i}{N + K}.$$

EXAMPLE 2.1 (Sampling without replacement.) Let $X_1 = x_1, \cdots, X_{N+1} = x_{N+1}$ denote a random sample drawn from a finite population with $m_i \geq 1$ members in each category i. Let $M = m_1 + \cdots + m_t$ and let $N \leq m_i$, all i. Then

$$P\{X_{N+1} \in \text{category } i \,|\, \mathbf{n}\} = \frac{m_i - n_i}{M - N} = \left(\frac{m_i}{M - N}\right) + \left(\frac{1}{N - M}\right) n_i. \tag{2.8}$$

Thus $a_i = m_i/(M - N)$ and $b = (N - M)^{-1} < 0$. Note that $k_i = -m_i$; thus k_i (and hence K) is independent of N, although a_i, A, and b are not. The next lemma states that this is always the case if, as here, the X_i are exchangeable and $b \neq 0$.

Let $a_i^{(N)}$, $b^{(N)}$, $k_i^{(N)}$, and $f_i(n_i, N)$ denote the dependence of a_i, b, k_i, and $f_i(n_i)$ on N. Thus, if (2.1) and (2.2) are satisfied for a fixed $N \geq 1$, then there exist $a_i^{(N)}$ and $b^{(N)}$ such that for all i, $f_i(n_i, N) = a_i^{(N)} + b^{(N)} n_i$. Note that $b^{(N)} = 0$ if and only if $\{X_1, \cdots, X_N\}$ and X_{N+1} are independent.

LEMMA 2.2. *Let $X_1, X_2, \cdots, X_{N+1}, X_{N+2}$ be an exchangeable sequence of* $\mathbf{t}$-*valued random variables, $N \geq 1$ and $t \geq 2$, satisfying* (2.1) *and* (2.3) *for both N and $N + 1$.*

(i) *If $b^{(N)} \cdot b^{(N+1)} = 0$, then $b^{(N)} = b^{(N+1)} = 0$.*

(ii) *If $b^{(N)} \cdot b^{(N+1)} \neq 0$, then $b^{(N)} \cdot b^{(N+1)} > 0$ and $k_i^{(N)} = k_i^{(N+1)}$, all i.*

PROOF. (i) Choose and fix two distinct indices $i \neq j$. Let $a_i = a_i^{(N)}$, $a_i' = a_i^{(N+1)}$, $b = b^{(N)}$, $b' = b^{(N+1)}$, etc. Suppose $b = 0$. It follows from exchangeability that for any partition $\mathbf{n}$ of N,

$$P\{X_{N+1} = i, X_{N+2} = j \,|\, \mathbf{n}\} = P\{X_{N+1} = j, X_{N+2} = i \,|\, \mathbf{n}\}, \tag{2.9}$$

hence

$$(a_i)(a_j' + b'n_j) = (a_j)(a_i' + b'n_i). \tag{2.10}$$

First taking $\mathbf{n}$ in (2.10) with $n_i = 0$, $n_j = N$, then with $n_i = N$, $n_j = 0$ and subtracting, we obtain $a_i b'N = -a_j b'N$, hence $b' = 0$ (since $a_i, a_j > 0$). Similarly, if $b' = 0$ then $b = 0$.

(ii) Suppose $b \cdot b' \neq 0$. Then it follows from (2.9) that for any partition $\mathbf{n}$ of N,

$$\left(\frac{n_i + k_i}{N + K}\right)\left(\frac{n_j + k_j'}{N + 1 + K'}\right) = \left(\frac{n_j + k_j}{N + K}\right)\left(\frac{n_i + k_k'}{N + 1 + K'}\right), \tag{2.11}$$

hence

$$k_i n_j + k_j' n_i + k_i k_j' = k_i' n_j + k_j n_i + k_i' k_j. \tag{2.12}$$

Letting $n_i = 0$, $n_j = N$ in (2.12), then $n_i = N$, $n_j = 0$ and subtracting, we obtain $k_i + k_j = k_i'$

$+ k_j'$; since i and j were arbitrary, this implies $K = K'$ and, if $t > 2$, $k_i = k_i'$ for all i. Since $a_i, a_i' > 0$, clearly b and b' must have the same sign.

Suppose $t = 2$ (so that $i = 1, j = 2$, say, and $K = k_1 + k_2$). Taking $n_i = 0, n_j = N$ in (2.12), we obtain $k_1(N + k_2') = k_1'(N + k_2)$, hence

$$k_1(N + K) = k_1'(N + K)$$

from which it follows (since $N + K = b^{-1} \neq 0$) that $k_1 = k_1'$ hence $k_2 = k_2'$. □

Together, Lemmas 2.1 and 2.2 immediately imply the following.

THEOREM 2.1. *Let $X_1, X_2, \cdots, X_{N_0}$ ($N_0 \geq 3$) be an exchangeable sequence of t-valued random variables such that for every $N < N_0$,* (i) (2.1) *holds,* (ii) (2.2) *holds if $t > 2$ or* (2.3) *holds if $t = 2$. If the $\{X_j\}$ are not independent ($\Leftrightarrow b^{(1)} \neq 0$), then there exist constants $k_i \neq 0$, either all positive or all negative, such that $N + \sum k_i \neq 0$ and*

$$P\{X_{N+1} = i \mid \mathbf{n}\} = \frac{n_i + k_i}{N + \sum k_i} \tag{2.13}$$

for every $N < N_0$, partition $\mathbf{n}$ of N, and $i \in \mathbf{t}$.

COROLLARY 2.1. *If $X_1, X_2, X_3, \cdots$ is an infinitely exchangeable sequence which for every $N \geq 1$, satisfies both* (i) (2.1), *and* (ii) *either* (2.2), *if $t > 2$ or* (2.3), *if $t = 2$, then $b^{(1)} \geq 0$.*

PROOF. Suppose $b^{(1)} < 0$. But then $N + K = 1/b^{(N)} < 0$ for all N, which is clearly impossible. □

COROLLARY 2.2. *For all $N \leq N_0$, under the conditions of Theorem* 2.1,

$$\begin{aligned} P\{X_1 = i_1, X_2 = i_2, \cdots, X_N = i_N\} &= \frac{\prod_{i=1}^{t} \{\prod_{j=0}^{n_i - 1} (j + k_i)\}}{\prod_{j=0}^{N-1} (j + K)} \\ &= \frac{\Gamma(K)}{\Gamma(N + K)} \prod_{i=1}^{t} \left\{\frac{\Gamma(n_i + k_i)}{\Gamma(k_i)}\right\}. \end{aligned} \tag{2.14}$$

PROOF. It follows from the product rule for conditional probabilities that it suffices to prove $P\{X_1 = i\} = k_i/K$ for all $i \in \mathbf{t}$. But

$$P\{X_1 = i, X_2 = j\} = \{a_j^{(1)} + b^{(1)}\delta_j(i)\}P\{X_1 = i\}, \tag{2.15}$$

where $\delta_j(i)$ is the indicator function of $\{i = j\}$. Summing over i in (2.5) gives $P\{X_2 = j\} = a_j^{(1)} + b^{(1)}P\{X_1 = j\}$, hence by exchangeability $P\{X_1 = j\} = a_j^{(1)}/(1 - b^{(1)}) = k_i/K$, since $a_j^{(1)} = k_i b^{(1)}$, $1 - b^{(1)} = A^{(1)}$ (cf. (2.6)), and $K = A^{(1)}/b^{(1)}$. □

It follows from Corollary 2.2 that $\{k_i : i \in \mathbf{t}\}$ uniquely determines $P = \mathscr{L}(X_1, X_2, \cdots, X_{N_0})$. Conversely, for every summable sequence of constants $\{k_i\}$, all of the same sign, there exists a maximal sequence of $\mathbf{t}$-valued random variables $X_1, X_2, \cdots, X_{N_0}$ ($N_0 \leq \infty$) such that (2.1) and (2.13) hold. The length of this sequence is determined by N^*, the largest value of N such that

$$p_{i,N} = \frac{n_i + k_i}{N + K}$$

determines a probability measure on $\mathbf{t}$, i.e., $N_0 = N^* + 1$, where

(i) if $k_i > 0$, all i, and $\sum k_i < \infty$, then $N^* = \infty$, or

(ii) if $k_i < 0$, all i, and $\sum |k_i| < \infty$, then

$$N^* = \max\{N \geq 0 : N + K < 0; N + k_i \leq 0, \text{ all } i\}.$$

Thus, if $K < 0$, N^* is the integer part of $\min\{|k_i| : i \in \mathbf{t}\}$. Hence, if $N_0 > 1$, then $t < \infty$ (since $\sum |k_i| < \infty$ implies $k_i \to 0$).

When $k_i > 0$ and $t < \infty$, the cylinder set probabilities in (2.14) coincide with those arising from the Dirichlet distribution in (1.5), and the characterization referred to at the end of Section 1 follows.

3. Complements and Extensions.

3.1. *The Symmetric Dirichlet.* Johnson considered the special case where (i) f_i is independent of i, i.e., for each N, there exists a *single* function f such that

$$P\{X_{N+1} = i \mid \mathbf{n}\} = f(n_i, N) \quad \text{for all} \quad i; \tag{3.1}$$

(ii) b is positive (This is the major gap in Johnson's proof. If $\{X_1, X_2, \ldots\}$ is infinitely exchangeable, but not independent, the assumption that b is positive is superfluous; see Corollary 2.1 above.)

Under these conditions $t < \infty$, $a_i \equiv a$, $k_i \equiv k > 0$, $P\{X_1 = i\} = \dfrac{1}{t}$,

$$P\{X_{N+1} = i \mid \mathbf{n}\} = \frac{n_i + k}{N + kt}, \tag{3.2}$$

and $X_1, \cdots, X_N$ can be extended to an infinitely exchangeable sequence, whose mixing measure dF in the de Finetti representation is the symmetric Dirichlet distribution with parameter k. Good (1965, page 25) suggests that Johnson was "unaware of the connection between the use of a flattening constant k and the symmetrical Dirichlet distribution." However, Johnson was at least aware of the connection when $k = 1$, for he wrote of his derivation of (1.4) via the combination postulate,

> ... I substitute for the mathematician's use of Gamma functions and the 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, page 418); Johnson's α corresponds to our t.]

3.2 *Alternate Approaches.* Let Δ_t be the probability simplex $\{p_i \geq 0,\ i = 1, \cdots, t: \sum p_i = 1\}$. Doksum (1974, Corollary 2.1) states in the present setting that a probability measure dF on Δ_t has a posterior distribution $dF(p_i \mid X_1, \cdots, X_n)$, which depends on the sample only through the values of n_i and N, if and only if dF is Dirichlet or

(i) dF is degenerate at a point (i.e., $X_1, X_2, \cdots$ is independent);

(ii) dF concentrates on a random point (i.e., dF is supported on the extreme points $\{\delta_i(j) : i = 1, \cdots, t\}$ of Δ_t, so that (2.1) would not hold);

(iii) dF concentrates on two nonrandom points (i.e., $t = 2$ or can be taken to be so).

This is a slightly weaker result than Johnson's, which only makes the corresponding assumption about the posterior *expectations* of the p_i.

Diaconis and Ylvisaker (1979, pages 279–280) prove (using Ericson's theorem, 1969, page 323) that the beta family is the unique one allowing linear posterior expectation of success in exchangeable binomial sampling, i.e., $t = 2$ and $\{X_n\}$ infinitely exchangeable, and remark that their method may be extended to similarly characterize the Dirichlet priors in multinomial sampling. Ericson's results can even be applied in the finitely exchangeable case and permit the derivation of alternate expressions for the coefficients a_i and b of (2.3).

3.3. *When is Johnson's postulate inadequate*? In practical applications Johnson's sufficientness postulate, like exchangeability, may or may not be an adequate description of our state of knowledge. Johnson himself did not review his postulate as universely applicable:

> 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, pages 418–419).]

Jeffreys (1939, Section 3.23) briefly discusses when such conditions may hold. Good (1953, page 241; 1965, pages 26–27) remarks that the use of Johnson's postulate fails to take advantage of information contained in the "frequencies of frequencies" (often useful in sampling of species problems), and elsewhere (Good, 1967) advocates mixtures of symmetric Dirichlets as frequently providing more satisfactory initial distributions in practice.

3.4. *Partition exchangeability.* If the cylinder sets $\{X_1 = i_1, \cdots, X_n = i_N\}$ are identified with the functions $g: \{1, \cdots, N\} \to \{1, \cdots, t\}$, then the exchangeable probability measures P are precisely those P such that

$$P\{g \circ \pi\} = P\{g\}$$

for all g and all permutations π of $\mathbf{N} = \{1, 2, \cdots, N\}$. Equivalently, the exchangeable P's are those such that the frequencies $\mathbf{n}$ are sufficient statistics with $P\{\cdot \mid \mathbf{n}\}$ uniform.

The rationale for exchangeability is the assumption that the index set $\mathbf{N}$ conveys no information other than serving to distinguish one element of a sample from another. In the situation envisaged by Johnson, Carnap (see Section 4 below), and others, a similar state of knowledge obtains *vis-a-vis* the index set $\mathbf{t}$ (think of the categories as colors). Then it would be reasonable to require of P that

$$P\{\pi_2 \circ g \circ \pi_1\} = P\{g\}$$

for all functions $g: \mathbf{N} \to \mathbf{t}$, and permutations π_1 of $\mathbf{N}$, π_2 of $\mathbf{t}$. Call such P's *partition-exchangeable.* The motivation for the name is the following. Let $\mathbf{a}(\mathbf{n}) = \{a_r : 0 \le r \le N\}$ denote the frequencies of the frequencies $\mathbf{n}$, i.e., $a_r = \#\{n_i = r\}$. Then P is partition-exchangeable if and only if the a_r are sufficient with $P\{\cdot \mid \mathbf{a}(\mathbf{n})\}$ uniform, i.e. $P\{g_1\} = P\{g_2\}$ whenever $\mathbf{a}(\mathbf{n}(g_1)) = \mathbf{a}(\mathbf{n}(g_2))$. The set of partition-exchangeable probabilities is a convex set containing the symmetric Dirichlets. From this perspective the frequencies of frequencies emerge as maximally informative statistics and the mixtures of symmetric Dirichlets as partition-exchangeable.

It would be of interest to have extensions of Johnson's results to "representative functions" of the functional form $f = f(n_i, \mathbf{a}(\mathbf{n}))$; for partial results in this direction ($f = f(n_i, a_0)$), see Hintikka and Niiniluoto (1976), Kuipers (1978). It would also be of interest to have Johnson type results for Markov exchangeable and other classes of partially exchangeable sequences of random variables; cf. Diaconis and Freedman (1980) for the definition and further references; Niiniluoto (1980) for an initial attempt.

4. Historical Note. Johnson's results appear to have attracted little interest during his lifetime. C. D. Broad, in his review of Johnson's *Logic* (vol. 3, 1924), while favorable in his overall assessment of the book, was highly critical of the appendix on "eduction" (in which Johnson introduced the concept of exchangeability and characterized the multinomial generalization of the Bayes-Laplace prior!): "About the Appendix all I can do is, with the utmost respect to Mr. Johnson, to parody Mr. Hobbes's remark about the treatises of Milton and Salmasiius: 'Very good mathematics; I have rarely seen better. And very bad probability: I have rarely seen worse.'" (Broad (1924, page 379); see generally pages 377–379.) Other than this, two of the few references to Johnson's work on the multinomial, prior to Good (1965), are passing comments in Harold Jeffreys's *Theory of Probability* (1939, Section 3.23), and Good (1953, pages 238–241). This general neglect is all the more surprising, inasmuch as Johnson could count among his students Keynes, Ramsey, and Dorothy Wrinch (one of Jeffreys's collaborators). (For Keynes's particular indebtedness to Johnson, see the former's *Treatise on Probability* (1921, pages 11 (footnote 1), 68–70, 116, 124 (footnote 2), 150–155; cf. Broad (1922, pages 72, 78–79), Passmore (1968, pages 345–346).)

It is ironical that in the decades after Johnson's death, Rudolph Carnap and his students would, unknowingly, reproduce much of Johnson's work. In 1945 Carnap introduced the function $c^*[= P\{X_{N+1} = i \mid \mathbf{n}\}]$ and proved that it had to have the form (1.4) under the assumption that all "structure-descriptions" [= partitions $\mathbf{n}$] were *a priori* equally likely (see Carnap, 1945; Carnap, 1950, Appendix). And just as Johnson grew uneasy with his combination postulate, so too Carnap would later introduce the family of functions $\{c_\lambda\colon 0 \leq \lambda \leq \infty\}$ [$= (n_i + k)/N + kt$, λ corresponding to our k], the so-called "continuum of inductive methods" (Carnap, 1952). But while Johnson proved that (3.2) followed from the sufficientness postulate (3.1), Carnap initially *assumed* both, although his collaborator John G. Kemeny was soon after able to show their equivalence for $t > 2$. Subsequently Carnap generalized these results, first proving (3.2) follows from a linearity assumption ((2.3)) when $t = 2$ (Carnap and Stegmüller, 1959), and later, in his last and posthumously published work on the subject, dropping the equiprobability assumption (3.1) in favor of (2.2) (Carnap, 1980, Section 19; cf. Kuipers, 1978). For the historical evolution of this aspect of Carnap's work, see Schilpp (1963, pages 74–75, 979–980); Carnap and Jeffrey (1971, pages 1–4, 223); Jeffrey (1980, pages 1–5, 103–104).

For details of Johnson's life, see Broad (1931), Braithwaite (1949); for assessments of his philosophical work, Passmore (1968, pages 135–136, 343–346), Smokler (1967), Prior (1967, page 551). In addition to his work in philosophy, Johnson wrote several papers on economics, one of which, on utility theory, is of considerable importance; all are reprinted, with brief commentary, in Baumol and Goldfeld (1968).

Acknowledgment. I thank Persi Diaconis and Stephen Stigler for a number of helpful comments and references. I am particularly grateful to Dr. Michael A. Halls, of King's College Library, for locating and providing a copy of the photograph of Johnson reproduced here. The photograph may have been taken in 1902, when Johnson became a Fellow of King's College.

REFERENCES

Bachelier, L. (1912). *Calcul des probabilités*, vol. 1. Gauthier-Villars, Paris.

Baumol, W. J. and Goldfeld, S. M. (1968). *Precursors in Mathematical Economics: An Anthology.* Series of Reprints of Scarce Works on Political Economy **19.** London School of Economics and Political Science, London.

Braithwaite, R. B. (1949). Johnson, William Ernest. In *Dictonary of National Biography 1931–1940.* Oxford University Press, 489–490.

Broad, C. D. (1922). Critical notice of J. M. Keynes. *A Treatise on Probability. Mind* **31** 72–85.

Broad, C. D. (1924). Mr. Johnson on the logical foundations of science. *Mind* **33** 242–269, 367–384.

Broad, C. D. (1931). William Ernest Johnson. *Proceedings of the British Academy* **17** 491–514.

Carnap, R. (1945). On inductive logic. *Philosophy of Science* **12** 72–97.

Carnap, R. (1950). *Logical Foundations of Probability.* The University of Chicago Press. Second edition 1962 (original text reprinted with minor corrections and supplementary bibliography added).

Carnap, R. (1952). *The Continuum of Inductive Methods.* The University of Chicago Press.

Carnap, R. (1980). A basic system of inductive logic, part II. In Jeffrey (1980), 7–155.

Carnap, R. and Jeffrey, R. C. (1971). *Studies in Inductive Logic and Probability*, volume I. University of California Press, Berkeley.

Carnap, R. and Stegmüller, W. (1959). *Induktive Logik und Wahrscheinlichkeit.* Springer-Verlag, Vienna.

De Morgan, A. (1845). Theory of probabilities. *Encyclopedia Metropolitana*, Volume 2, London, 393–490.

Diaconis, P. and Freedman, D. (1980). De Finetti's generalizations of exchangeability. In Jeffrey (1980), 233–249.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7** 269–281.

Doksum, K. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Ann. Probab.* **2** 183–201.

EDGEWORTH, F. Y. (1884). Philosophy of chance. *Mind* **9** 222–235.
EDWARDS, A. W. F. (1974). A problem in the doctrine of chances. *Proceedings of Conference on Foundational Questions in Statistical Inference* (Aarhus, 1973), 43–60. *Memoirs* No. 1, Dept. Theoret. Statist., Inst. Math., Univ. Aarhus, Aarhus.
EDWARDS, A. W. F. (1978). Commentary on the arguments of Thomas Bayes. *Scand. J. Statist.* **5** 116–118.
ERICSON, W. A. (1969). Subjective Bayesian models in sampling finite populations. *J. Roy. Statist. Soc. Ser. B* **31** 195–224.
FISHER, R. A. (1956). *Statistical Methods and Scientific Inference.* Hafner, New York. Third edition, 1973.
GOOD, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40** 247–264.
GOOD, I. J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* Research Monograph No. 30, M.I.T. Press.
GOOD, I. J. (1967). A Bayesian significance test for multinomial distributions. *J. Roy. Statist. Soc. Ser. B* **29** 399–431 (with discussion).
HARDY, G. F. (1889). Letter. *Insurance Record* (1889), page 457. [Reprinted in *Trans. Fac. Actuaries* **8** (1920), pages 180–181.]
HINTIKKA, J. and NIINILUOTO, I. (1976). An axiomatic foundation for the logic of inductive generalization. In *Formal Methods in the Methodology of Empirical Sciences* (M. Przelecki, K. Saniawski, and R. Wojcicki, eds.), D. Reidel, Dordrecht. [Reprinted in Jeffrey (1980), pages 157–181.]
JEFFREY, R. C., ed. (1980). *Studies in Inductive Logic and Probability,* volume II. University of California Press, Berkeley.
JEFFREYS, H. (1939). *Theory of Probability.* Clarendon Press, Oxford. Third edition, 1961.
JOHNSON, W. E. (1924). *Logic, Part III. The Logical Foundations of Science.* Cambridge University Press. [Reprinted 1964, Dover, New York]
JOHNSON, W. E. (1932). Probability: the deductive and inductive problems. *Mind* **49** 409–423. [Appendix on pages 421–423 edited by R. B. Braithwaite]
KEYNES, J. M. (1921). *A Treatise on Probability.* Macmillan, London.
KUIPERS, T. (1978). *Studies in Inductive Probability and Rational Expectation.* Synthese Library **123**. D. Reidel, Dordrecht.
LAPLACE, P. S. *(1781).* Mémoire sur les probabilités. *Mem. Acad. Sci. Paris 1778,* 227–232; *Oeuvres complètes* **9** 383–485.
LIDSTONE, G. J. (1920). Note on the general case of the Bayes-Laplace formula for inductive or *a posteriori* probabilities. *Trans. Fac. Actuar.* **8** 182–192.
MURRAY, F. H. (1930). Note on a scholium of Bayes. *Bull. Amer. Math. Soc.* **36** 129–132.
NIINILUOTO, I. (1981). Analogy and inductive logic. *Erkenntnis* **16** 1–34.
PASSMORE, J. (1968). *A Hundred Years of Philosophy,* 2nd ed. Penguin, New York.
PEARSON, K. (1907). On the influence of past experience on future expectation. *Philosophical Magazine,* Series 6 **13** 365–378.
PRIOR, A. N. (1967). Keynes; Johnson. In *The Encyclopedia of Philosophy* (Paul Edwards, ed.) **4** 550–551. Macmillan and Free Press, New York.
SCHILLP, P. A., ed. (1963). *The Philosophy of Rudolph Carnap.* Open Court, La Salle, Ill.
SMOKLER, H. E. (1967). Johnson, William Ernest. In *The Encyclopedia of Philosophy* (Paul Edwards, ed.) **4** 292–293. Macmillan and Free Press, New York.
STIGLER, S. M. (1982). Thomas Bayes's Bayesian inference. *J. Roy. Statist. Soc. Series A* **145**, Part II 250–258.
WHITWORTH, W. A. (1897). *DCC Exercises in Choice and Chance.* [Reprinted 1965, Hafner, New York.]
ZABELL, S. L. (1982). R. A. Fisher on the history of inverse probability. Manuscript in preparation.

DEPARTMENT OF MATHEMATICS
NORTHWESTERN UNIVERSITY
EVANSTON, ILLINOIS 60201