Recent Progress on de Finetti's Notions of Exchangeability

P. DIACONIS
Stanford University

SUMMARY

This review covers the following topics: exchangeability, partial exchangeability with finitely many types, Markov exchangeability, random walk with reinforcement—an application of the Markov theory, mixtures of exponential families and the Küchler-Lauritzen theorem, Gibbs states, Ressell's work, and finite forms of de Finetti's theorem. It concludes by introducing some "new" results that date back to de Finetti's original paper on partial exchangeability.

Keywords: EXCHANGEABLE; PARTIALLY EXCHANGEABLE; MIXTURES; MARKOV CHAINS, DE FINETTI; CONDITIONED LIMIT THEOREMS.

1. INTRODUCTION

This paper gives a survey of work on exchangeability, between the second and third Valencia meetings. To keep things self-contained, brief reviews of previous work on exchangeability and partial exchangeability have been added. A full survey of this classical material can be found in Diaconis and Freedman (1984), Aldous (1987) or Lauritzen (1982).

Diaconis and Freedman present the material in the language of Bayesian statistics. Aldous describes developments and applications to probability. Lauritzen uses the language of extreme point models pionnered by Martin-Löf.

The review sections also present some new work—an application of de Finetti's theorem for Markov chains to a novel problem of random walk with reinforcement and the Küchler-Lauritzen theorem for characterizing mixtures of exponential families.

Sections on newer results cover Ressel's developments using semi-groups, joint work with Freedman on finite forms of the basic theorems, and a battery of recent solutions of special problems.

The final section attempts to explore some untapped streams in de Finetti's original papers on partial exchangeability using modern notation.

2. EXCHANGEABILITY

Consider a process $X_1, X_2, X_3, \ldots$, taking two values. A probability distribution P for the processes is *exchangeable* if it is invariant under permutations:

$$P\{X_1 = e_1, \ldots, X_n = e_n\} = P\{X_{\pi(1)} = e_1, \ldots, X_{\pi(n)} = e_n\} \qquad (2.1)$$

where $e_1, e_2, \ldots$, is any sequence of possible values, and π is any permutation.

de Finetti's (1931) basic theorem supposes a potentially infinite exchangeable process. He shows that there is a unique representing measure μ on the unit interval such that for any n, and any sequence $e_1, \ldots, e_n$,

$$P\{X_1 = e_1, \ldots, X_n = e_n\} = \int x^a (1-x)^b \mu(dx), \qquad (2.2)$$

111

with a the number of e_i of type 1 among $e_1, e_2, \ldots, e_n$ and b the number of e_i of type 2.

Expressions like the right hand side of (2.2) have been used since Bayes' original paper. In modern language μ is called the prior measure $x^a(1-x)^b$ the likelihood. This μ may be thought of as the prior opinion about the limiting proportion of type one events. Subjectivists prefer not to speak about unobservable events (like limiting frequencies). They are perfectly willing to assign prior opinions to observable events like three successes out of the next ten trials. The theorem shows these two ways of working are equivalent.

More generally, X_i can take values in any nice space $\mathcal{X}$ (such as a complete, separable metric space). Then, de Finetti (1938), Hewitt and Savage (1955), Diaconis and Freedman (1980c), show in varying degree of generality that for an infinite exchangeable process

$$P\{X_1 \in A_1, \ldots, X_n \in A_n\} = \int_{\mathcal{P}} \prod_{i=1}^{n} F(A_i)\mu(dF). \qquad (2.3)$$

On the left, A_i are arbitrary Borel sets in $\mathcal{X}$. On the right $\mathcal{P}$ is the set of all probabilities on the Borel sets of $\mathcal{X}$ (itself given the weak star topology) and μ is a unique probability on the sets of $\mathcal{P}$. The same μ works for any n and A_i.

Curiously, results like (2.3) require some sort of topological restriction (Dubins and Freedman (1979)). Forms for finite n, or forms involving finitely additive μ are available with only the measure structure. See Diaconis and Freedman (1980c).

Most subjectivists find (2.2) a very satisfactory theorem. It seems like a reasonable task to specify a prior distribution on $[0,1]$ if only approximately. As to (2.3), it seems like an essentially impossible task to meaningfully specify a prior on all probabilities on the real line. One searches for additional restrictions, like symmetry and smoothness to cut the problem down to manageable size. See Section 6 below.

Diaconis and Freedman (1986) show how conventional, automated methods of putting a prior on such infinite dimensional spaces can lead to silly procedures in quite practical problems such as the basic measurement error model.

Lester Dubins points out that the unit interval and the space of all probabilities on the real line have the same cardinality c (from a recursively enumerable view both are effectively countable) so it may be only a lack of experience and suitable language that makes (2.3) seem less useful than (2.2).

3. PARTIAL EXCHANGEABILITY WITH FINITELY MANY TYPES

In 1938, de Finetti broadened the concept of exchangeability. Consider first the special case with two types of observations: $X_1, X_2, \ldots$; $Y_1, Y_2, \ldots$. The X_i might represent binary outcomes for a group of men and the Y_i might represent binary outcomes for a group of women. *If* it were judged that the observable covariate men/women did not matter, all of the variables would be judged exchangeable. Often, the covariate is judged as potentially meaningful, the X_i's are judged exchangeable between themselves and the Y_i's are judged exchangeable between themselves. Mathematically, the joint law must be invariant under permutations within the X's and Y's:

$$\mathcal{L}(X_1, \ldots, X_n; Y_1, Y_2, \ldots, Y_m) = \mathcal{L}(X_{\pi(1)}, \ldots, X_{\pi(n)}; Y_{\sigma(1)}, \cdots, Y_{\sigma(m)}). \qquad (3.1)$$

This must hold for all n and m, and permutations π and σ.

de Finetti proved that for an infinite process $\{X_i, Y_j\}$ partially exchangeable as in (3.1) implies

$$\frac{X_1 + \cdots + X_n}{n}, \frac{Y_1 + \cdots + Y_m}{m} \to (p_1, p_2) \qquad \text{almost surely,} \qquad (3.2)$$

$$P\{X_1 = e_1, \ldots, X_n = e_n; Y_1 = f_1, \ldots, Y_m = f_m\} =$$

$$= \int \int p_1^a (1-p_1)^b p_2^c (1-p_2)^d \mu(dp_1, dp_2), \tag{3.3}$$

for a unique measure μ on the Borel sets of the unit square $[0,1]^2$. Of course, a and b are the number of ones and zeros among the e_i and c and d are the number of ones and zeros among the f_i. The same measure μ works for all n, m and all binary sequences $e_1, \ldots, e_n$; $f_1, \ldots, f_m$.

Return to the case of two types and consider situations where one is unsure if the covariate matters: e.g., if the outcome is passing a written test or not, it may well be that the covariate has only a negligible influence. de Finetti's theorem represents the joint law of the processes as a mixture over the unit square. If the X_i and Y_j were *all* exchangeable, the mixing measure μ would be supported on the diagonal $p_1 = p_2$.

de Finetti (1972, Chap. 9) has shown how to build natural prior distributions on the unit square which are supported near the diagonal as a way of allowing that a difference might show up in the data but allowing expression of the belief that most probably the covariate does not matter. The paper by Bruno translated as Chapter 10 of de Finetti (1972) works out some numerical examples with three types that are fascinating: as a constant stream of data comes in one can oscillate between the types. Diaconis and Freedman (1988c) construct examples with infinitely many types in which convergence never occurs. Alas, this seems like a model of the way things work in practical inference—as more data comes in, one admits a richer and richer variety of explanatory hypothesis. Without care, a simple message can become unrecognizable.

4. MARKOV CHAINS

de Finetti (1959) mentions the possibility of a subjective treatment of Markov chains using partial exchangeability. The idea is that the type of the i-th observable depends on the outcome of the previous observation. With binary processes, there are three types: the first observation, observations following a zero, and observations following a one. A precise mathematical formulation becomes tricky and it is simpler to proceed along the following lines developed by Freedman (1962a).

Consider two binary sequences of zeros and ones. Call them *equivalent* if they begin with the same symbol and have the same number of transitions from 0 to 0, 0 to 1, 1 to 0 and 1 to 1. Thus

$$0101101011 \quad \text{and} \quad 0110101011 \quad \text{and} \quad 0101011011$$

are all equivalent having transitions

$$\text{from} \quad \begin{matrix} & & \text{to} \\ & & 1 \\ 0 \\ 1 \end{matrix} \begin{pmatrix} 0 & 4 \\ 3 & 2 \end{pmatrix}.$$

Note that the first string is obtained from the second by switching the first one and the block of 2 ones. Switching such blocks does not change the overall transitions and it is easy to show that equivalent sequences can be obtained by switching blocks.

A probability on binary sequences is called *partially exchangeable* if it assigns equal probability to equivalent strings. Freedman (1962a) showed that a stationary partially exchangeable process is a mixture of Markov chains. Diaconis and Freedman (1980a) eliminated the stationary assumption. To get a mixture of Markov chains, infinitely many returns to the starting state are needed. This is guaranteed by a recurrence assumption

$$P\{X_n = X_0 \text{ infinitely often}\} = 1. \tag{4.1}$$

Theorem. *Let $X = (X_0, X_1, X_2, \ldots)$ be a partially exchangeable process satisfying the recurrence condition (4.1). Then X is a mixture of Markov chains: $P\{X_0 = e_0, X_1 = e_1, \ldots, X_n = e_n\} = \int \int \prod p_{ij}^{t_{ij}} \mu(dp_{ij})$ with t_{ij} being the number of i to j transitions in the sequence $e_0, e_1, \ldots, e_n, 0 \le i, j \le 1$. The measure μ is unique and does not depend on n.*

The extension to countable state space is straightforward. On the other hand, the extension to general state spaces require sophisticated machinery, see Freedman (1963, 1984) and Kallenberg (1975, 1981). These last papers also discuss the extension to continuous time.

More generally, the extension of de Finetti's theorem to cover naturally occuring stochastic processes is a challenging research area. This was started by Freedman (1962, 1963) who characterized Poisson processes and Brownian motion with unknown parameters. Kallenberg (1976) contains more recent results.

One recently solved problem deserves special mention. Consider a pure birth process (Galton-Watson process) with one unknown parameter —the rate of births λ. This description, in terms of an unknown parameter λ, is like the right side of (2.2) in Section 1. What is the characterization in terms of observables? Roughly this: For each time t, two paths with the same number of births and area (the area under the path and above the x axis) should be assigned the same probability. The details are very tricky. See Oleg and Hamler (1988) for a careful statement and proof.

There is much further work to be done in this area.

5. RANDOM WALK WITH REINFORCEMENT – AN APPLICATION

Consider a triangle
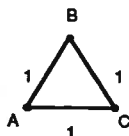
A random walk starts at A and chooses B or C with probability $\frac{1}{2}$. Each time the walk travels over an edge, 1 is added to the edge-weight. At a new vertex, the walk chooses the next vertex with probability proportional to edge-weights leading out of the present vertex.

Thus if the first choice is C, the walk moves to C and the 1 on the AC edge is changed to a 2. The walk next chooses to go back to A (probability $\frac{2}{3}$) or move to B (probability $\frac{1}{3}$). The process continues in this way.

Random walk with re-enforcement is a simple version of models for neural networks. It was introduced as a simple model of exploring a new city. At first all routes are equally unfamiliar and one chooses at random between them. As time goes on, routes that have been traveled more in the past are more likely to be traveled.

Of course, such a walk can be performed on any graph. For now, let us stick to the triangle and ask what happens as time goes on? People often guess that the walk eventually dies on an edge, or else winds up visiting each vertex about a third of the time. The answer is not so simple.

Let W_{AB}^n, W_{AC}^n, W_{BC}^n be the edge-weights at time n. These add up to $n + 3$. The theory to be described shows that the W's divided by n tend to a limit

$$\frac{1}{n}(W_{AB}^n, W_{AC}^n, W_{BC}^n) \to (L_{AB}, L_{AC}, L_{BC}) \quad \text{almost surely.}$$

Here $L_{AB} + L_{AC} + L_{BC} = 1$, and the limits are random, with an absolutely continuous distribution on the simplex $x + y + z = 1$. The limiting density is proportional to

$$(xy + xz + yz)^{1/2}(x + y)^{-1}(x + z)^{-3/2}(y + z)^{-3/2}. \tag{5.1}$$

A similar result holds for any finite graph and starting edge-weights: the edge-weights, divided by the number of steps converge almost surely. The limit is random with an absolutely continuous density which can be explicitly described in terms of the homology group of the graph.

These results are closely liked to de Finetti's theorem and indeed contribute to statistical inference for Markov chains. To describe the link, consider a simple graph, with starting edge-weights 2

$$
\begin{array}{ccc}
2 & & 2 \\
\bullet\!\!-\!\!\!-\!\!\bullet\!\!-\!\!\!-\!\!\bullet \\
A & B & C
\end{array}
$$

A walk starts at B. At first it chooses between A and C with probability $1/2$. Say it goes to A. According to the rules, it next moves back to B and the edge-weights are

$$
\begin{array}{ccc}
4 & & 2 \\
\bullet\!\!-\!\!\!-\!\!\bullet\!\!-\!\!\!-\!\!\bullet \\
A & B & C
\end{array}
$$

Thus the next step is twice as likely to be back to A, etc.

On reflection, one easily sees that these edge-weights evolve exactly like the balls in Polya's urn. Here one begins with an urn containing one red and one white ball. Balls are choosen from the urn at random and replaced each time along with another of the same color. Polya showed that the proportion of red balls has a uniform limit.

Starting with a star graph with d external vertices

gives a Polya urn with d colors. Starting with a star having countably many extremal vertices, with the i-th edge having initial edge weight $\alpha(i)$ gives a Dirichlet random measure on the integers as limiting value.

As is well known, successive draws from a Polya urn form an exchangeable process and the limiting Dirichlet distribution is the representing measure in de Finetti's theorem.

Random walk with reinforcement on a general graph has a similar connection with the partially exchangeable prcesses discussed in Section 4 above. To describe the connection, consider a graph (V, E) with V a set of vertices and E a set of edges. There are starting weights $\alpha(e)$ for each $v \in V$, $\Sigma_{v \in e}\alpha(e) < \infty$. Start at vertix v_0 and run random walk with reinforcement. This generates a process $V : V_0 = v_0, V_1, V_2, \ldots$; the successive vertices visited.

A direct computation, similar to the proof of exchangeability for Polya's urn, show that the process V is partially exchangeable as in Section 4.

By the results of Diaconis and Freedman (1980a) a de Finetti type representation obtains: the process V can be represented as a mixture of Markov chains. Further, the empirical transition count matrix after n steps, divided by n, has an almost sure limit. This is how we know that the edge weights for the triangle converge.

Calculation of the representing measure is a far trickier business. At present, the only method involves a difficult combinatorics calculation. This is carried out by Coppersmith and Diaconis (1986). We merely state the result:

Theorem. *For random walk with reinforcement on a finite graph (V, E) let $\{W_e^n\}_e \in E$ be the edge-weights. Then*

(a) $\frac{W_e^n}{n}$ *converges almost surely to a limit on the $|E|$ simplex.*

(b) *the limit is random with absolutely continuous distribution having density*

$$C\Pi_e x_e^{(\alpha_e - \frac{1}{2})}\Pi_v x_v^{-(s_v+1)/2} x_{v_0}^{\frac{1}{2}}|A|^{\frac{1}{2}}. \tag{5.2}$$

In (5.2), x_e are variables on the simplex, $x_v = \sum_{v \in e} x_e$, α_e are the starting edge-weights, $s_v = \sum_{v \in e} \alpha_e$, v_0 is the starting vertix. The matrix A has dimension $|E| - |V| + 1$. This is the dimension of the first homology group(see Giblin (1977)). This is the group of "loops" c_i given with an arbitrary orientation. A has entries:

$$A_{ii} = \sum_{e \in c_i} 1/x_e$$

$$A_{ij} = \sum_{e \in c_i \cap c_j} \pm 1/x_e \quad \text{The sign being } \pm \text{ as edge } e \text{ has the same or}$$

opposite orientation in the i-th and j-th loop.

Finally C is a normalizing constant making the density integrate to 1 over the simplex.

As as example, the triangle only has one loop. If the starting edge-weights are taken as 1, the limiting density is proportional to

$$(x_{ab}x_{ac}x_{bc})^{1/2}(x_{ab}+x_{ac})^{-1}(x_{ab}+x_{bc})^{-3/2}(x_{ac}+x_{bc})^{-3/2}\left(\frac{1}{x_{ab}} + \frac{1}{x_{ac}} + \frac{1}{x_{bc}}\right)^{1/2}. \tag{5.3}$$

This is a density on $x_{ab} + x_{ac} + x_{bc} = 1$ with respect to the uniform density.

The theorem above gives the limiting distribution for the edge-weights. The de Finetti representation is in terms of a mixture of Markov chains for the vertix process. In probabilistic language this means there is a measure on the set of 3×3 transition matrices of form

$$\text{from } \begin{array}{c} A \\ B \\ C \end{array}\begin{array}{c} \text{to} \\ \begin{pmatrix} A & B & C \\ 0 & a & 1-a \\ b & 0 & 1-b \\ c & 1-c & 0 \end{pmatrix} \end{array}.$$

One picks a matrix (and so a, b, and c) and then runs the Markov chain determined by this matrix from starting state A. The representation theorem says that this is an equivalent description of the process originally described by reinforcement.

The law of a, b, and c can be described as

$$a = \frac{L_{AB}}{L_{AB} + L_{AC}} \qquad b = \frac{L_{AB}}{L_{AB} + L_{BC}} \qquad c = \frac{L_{AC}}{L_{AC} + L_{BC}}$$

where the law of (L_{AB}, L_{AC}, L_{BC}) is given by (5.3). Observe that $abc = (1-a)(1-b)(1-c)$ so the mixing measure is singular, even on matrices with zeros on the diagonal.

Here is an application of the previous work to statistical inference for Markov chains. The Dirichlet prior is a basic ingredient for inference about multinomial parameters. Unfortunately, the conjugate priors for Markov chains have the rows of the unknown transition matrix as independent Dirichlet variables. See, e.g. Martin (1969). This seems to be quite a severe limitation. It seems fair to say that even in the 2×2 case

$$\begin{pmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{pmatrix}.$$

There is not a simple tractable family of prior distributions rich enough to capture believable prior information.

The mixing measures for random walk with reinforcement offer some variety *and* mathematical tractability. This is not evident from the density; indeed, it is not even immediately clear the density has a finite integral, and far less clear what the normalizing constant is.

The idea is to use the process representation: given the past of the process up to time n, the law of the future is given by the reinforcement description. For example, the chance of the next value being any of the vertices is proportional to the edge-weight leading to that vertix. This allows easy computation of posterior predictions and so, of posterior means of parameters. This seems to be a potentially fruitful direction for future development.

I will not take the time here to mention the interesting probability developments arising from random walk with reinforcement on infinite graphs. Pemantle(1988a, 1988b) contains much of interest here.

6. MIXTURES OF EXPONENTIAL FAMILIES

de Finetti's theorem presents a real-valued exchangeable process as a mixture over the set of all measures. Most of us find it hard to meaningfully quantify a prior distribution over so large a space. There has been a search for additional restrictions that get things down to a mixture of familiar families parametrized by low dimensional Euclidean parameter spaces.

The first result of this type was given by Freedman (1963) who characterized orthogonally invariant probabilities as scale mixtures of mean zero normals. Freedman also gave characterizations of the familiar 1 parameter exponential families such as Poisson, geometric, and gamma.

Dawid (1977) characterized covariance mixtures of multivariate normals. Diaconis, Eaton, and Lauritzen (1988) characterize the usual normal models for regression and analysis of variance in terms of symmetry or sufficiency.

By now it is clear that the correct version of these characterization results involves sufficiency: Start with a standard family on a space $P_\theta(dx)$. Let P_θ^n be product measure on $\mathcal{X}^n$. Let $T_n : \mathcal{X}^n \to \mathcal{Y}_n$ be a sufficient statistic. Let $Q_{n,t_n}(dx)$ be a regular conditional probability on $\mathcal{X}^n$ given $T_n = t_n$. By sufficiency, Q_{n,t_n} does not depend on θ.

One can define a general extension of exchangeability given T_n and Q_n by declaring a measure P on $\mathcal{X}^\infty$ to be *partially exchangeable* if for every n, Q_{n,t_n} is a regular conditional probability for the marginal law of P on $\mathcal{X}^n$ given $T_n = t_n$.

The class of partially exchangeable probabilities is a convex set and its extreme points can be identified in an indirect way as measures having a trivial partially exchangeable tail field. Then an abstract version of de Finetti's theorem follows: every partially exchangeable probability is a unique mixture of extreme points.

Myriad details have been left out of the above brief description. Careful versions have been given by Diaconis and Freedman (1984) in Bayesian language. Closely related results are given by Lauritzen (1974) in the language of extreme point models and by Dynkin (1978) or Reulle (1978) in the language of Gibbs states and statistical mechanics. The main results are essentially the same, but the language and examples vary widely between presentations.

The general theorem gives an abstract result which may require a fair amount of work to understand in specific cases. In particular, if the original family $P_\theta(dx)$ is a standard exponential family with $T_n = X_1 + \cdots + X_n$ the usual sufficient statistic, it is not at all clear if the extreme point representation says that partially exchangeable probabilities are mixtures of the exponential families that were started with. In other words, are the extreme points the P_θ^∞ and nothing else?

Following work of Martin-Löf (1970), Diaconis and Freedman (1984) solved the problem for 1-dimensional discrete exponential families. The problem for continuous families has

recently been solved by Küchler and Lauritzen (1986). They assume the base measure has a continuous and everywhere positive denstity. Here is a refinement of their result from Diaconis and Freedman (1988).

Let h be a nonnegative, finite, locally integrable Borel function on $\mathcal{R}$: Let $c(\theta) = \int_{-\infty}^{\infty} e^{\theta x} h(x)dx$ and let Θ be the set of θ with $c(\theta) < \infty$. Assume Θ is non-empty. The exponential family through h is defined as $P_\theta(dx) = e^{\theta x}h(x)dx/c(\theta)$.

Let P_θ^∞ be product measure on $\mathfrak{R}^\infty$. Define Q_{n_s} as the regular conditional probability for the coordinate functions $X_1, X_2, \ldots, X_n$ given $X_1 + \cdots + X_n = s$. This is only defined for $s \in \mathfrak{R}$ with $0 < h^{(n)}(s) < \infty$ with $h^{(n)}$ the n-fold convolution of h with itself. Let D_n be this set of s-values.

Define M_Q—the Q-exchangeable probabilities—as the set of probabilities P on $\mathcal{R}$ such that for every n

(a) $P\{X_1 + \cdots + X_n \in D_n\} = 1$

(b) Q_{n_s} is a regular conditional P-distribution for $X_1, \ldots, X_n$ given $X_1 + \cdots + X_n = s$.

Clearly $P_\theta^\infty \in M_Q$. So is P_μ, defined as $\int_\Theta P_\theta^\infty \mu(d\theta)$ with μ a probability on the Borel sets of θ. The following version of the Küchler-Lauritzen theorem is proved in Diaconis and Freedman (1988c).

Theorem. *P is Q-exchangeable if and only if P has the unique integral representation*

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

For example, if $P_\theta(dx)$ is a mean zero normal scale family, then $S_n = X_1^2 + \cdots + X_n^2$ is the sufficient statistic and given $S_n = s, X_1, X_2, \ldots, X_n$ is uniformly distributed on the sphere of radius $\sqrt{s}$. This specifies $Q_{n,s}$. Now a measure is conditionally uniform if and only if it is orthogonally invariant. The theorem specializes to Freedman's original result, P on $\mathfrak{R}^\infty$ is orthogonally invariant if and only if it is a scale mixture of mean zero normal variables.

It is natural to try to extend this theorem to other natural sufficient statistics such as products and maxima. Vector valued versions are also natural requests. Lauritzen (1975) set things up in the language of semi-groups: one looks at the conditional law of $x_1, \ldots, x_n$ given $x_1 * x_2 * \cdots * x_n$. This idea has been put into definitive form by Ressel (1985) who showed that the extreme points of the partially exchangeable probabilities are in $1 - 1$ correspondence with the positive part of the dual group. de Finetti's theorem for these cases then becomes Bochner's theorem which represents a positive definite function as an integral of characters. The details make extensive use of the modern theory of Abelian semi-groups as developed in Berg, Christiansen and Ressel (1984). One nice bonus is that the duals of many semi-groups have been classified so that new theorems result. One disadvantage: the results often wind up in highly analytic form, e.g. as conditions on the characteristic function of the process instead of on observables. It is a worthwhile project to try to systematically translate these results to conditions on observables.

7. FINITE THEOREMS

The classical results on exchangeability involve infinite exchangeable processes and constructions like tail fields which are known to have no finite content. It is also known that there are finite exchangeable sequences which cannot be represented as a mixture of i.i.d. processes.

Reasonable finite versions of de Finetti theorems have been evolving.

Theorem. (*Diaconis and Freedman, 1980c*). *Let* $X_1, X_2, \ldots, X_n$ *be a binary exchangeable sequence. Then there is a* μ *on* $[0, 1]$ *such that for* $k \leq n$

$$\|\mathcal{L}(X_1, \ldots, X_k) - P_\mu^k\| \leq \frac{2k}{n}. \tag{7.1}$$

In (7.1), $\mathcal{L}(X_1, \ldots, X_k)$ *stands for the marginal distribution of the first* k *coordinates as a measure on binary* k-*tuples,* $p_\mu^k(e_1, \ldots, e_n) = \int x^a (1 - x)^{k-a} \mu(dx)$ *with* $a = e_1 + \cdots + e_n$. *The norm is total variation distance:*

$$\|P - Q\| = \sup_A |P(A) - Q(A)|.$$

The theorem says that if one is considering a binary exchangeable process of length k which can be extended to an exchangeable process of length n (so that the experiment can be repeated in principle) then, de Finetti's theorem almost holds. We showed that the k/n rate cannot be improved.

If the process takes c values instead of 2, the rate ck/n obtains. For processes taking infinitely many values, the rate $2k/\sqrt{n}$ obtains. For infinite state spaces, there are no topological restrictions (such as Polish spaces). Taking limits leads to the most general known version of the infinite form of de Finetti's theorem.

These theorems are all proved by using an exact finite form of de Finetti's theorem representing the processes as a mixture of urn processes—samples without replacement from an urn of fixed composition. Then, sharp bounds between sampling with and without replacement are derived to show that the urn processes are approximately binomial processes. Thus the original processes are approximately mixtures of binomial processes.

A crucial point: sampling with and without replacement are close in variation distance *uniformly* in the contents of the urn.

Diaconis and Freedman (1987) give similar theorems for particular versions of de Finettis' theorem like normal location, or scale parameters, mixtures of Poisson, geometric and gamma (shape or scale parameters). In each case, there is a finite theorem with rate k/n in variation distance. Alas, all of the arguments are different, and it is not at all clear where the k/n comes from.

We embarked on a systematic investigation of exponential families in (1988a) by working directly on the conditional density of $X_1, X_2, \ldots, X_k$ given $X_1 + \cdots + X_n$. This is a ratio of convolutions and Edgeworth expansions allow careful bounds to be obtained. To guarantee tht the bounds hold uniformly in the conditioning variable $X_1 + \cdots + X_n$, strong conditions are imposed on the underlying exponential family (things like uniformly bounded standardized fourth moments). These rule out some natural examples (where the uniform form of the theorem is known to hold) but they show where the k/n rate comes from (terms of form $k/\sqrt{n}$ cancel out of numerator and denominator) and allow construction of counterexamples where the k/n rate fails.

Finite versions of de Finetti's theorem for Markov chains are given by Diaconis and Freedman (1980c) and Zaman (1986). Here one of the important steps was to find an exact representation of a finite partially exchangeable process as a mixture of urn processes. Zaman's (1984) elegant form of this has been used by geneticists to simulate the law of a random finite string with fixed transition counts as a way of calibrating DNA string matching algorithms. Zaman's result has an annoying extra factor of log n. It is not clear if this is really there or just an artifact of the proof.

Lest the reader think that all the the interesting problems have been solved, I suggest two open problems: Find a finite version of Aldous' (1981) basic theorem on random binary arrays invariant under permuting rows and columns. Diaconis and Freedman (1981) apply Aldous' theorem to a problem in human perception.

A second problem: The finite version of Freedman's basic theorem on orthogonally invariant measures required a sharp version of the following: pick a point at random on a high dimension sphere; the first k-coordinates are approximately independent normal. Diaconis and Freedman (1987) prove this for $k = o(n)$. This is a very special case of the following : let Γ be a uniformly distributed random orthogonal n by n matrix. In some sense all of the entries of Γ are approximately independent normal variables. Perhaps any $o(n^2)$ can be proved to be variation distance close to independent normals. Diaconis and Shahshahani (1987) and Diaconis, Eaton and Lauritzen (1988) contain background and references.

8. ON READING DE FINETTI

de Finetti wrote about partial exchangeability in his article of 1938 (translated in 1979) and 1959 section 9.6.2 (translated in 1972). Both treatment are rich sources of ideas which take many readings to digest. His basic examples involve several types (e.g., men and women) with exchangeability within type. He derives parametric representations, with one parameter per type. He emphasizes situations of almost exchangeability where the mixing measure concentrates near the diagonal.

de Finetti (1938) briefly describes two unusual examples that I want to present here.

Example 1. Suppose we are considering families and $X_i = 1$ or 0 as the i-th family reports an accident in the coming year or not. As a covariate, we currently know the number of people in each family.

If this is all we know, it is natural to assume that families with the same size are exchangeable, so there is one type for each family size. Then, de Finetti's theorem yields one parameter per family size and a mixture over the "cube" $0 \leq P_j \leq 1; j = 1, 2, \ldots$.

Thinking further, we may, as a first approximation, judge that *all* of the people involved are exchangeable. Then, the prior on the cube will be concentrated near a curve: If θ is the proportion of people in a single person family having an accident, the proportion of i member families reporting an accident should be $1 - (1 - \theta)^i$ approximately. Such a prior is clustered about the curve $(1 - \theta, 1 - \theta^2, 1 - \theta^3, \ldots)$ $0 \leq \theta \leq 1$.

Of course, one wants to allow mass off the curve. As de Finetti suggests, priors based on a fair amount of background data will be approximately normal. A prior with the features above has density proportional to $\exp -A\{(P_1^2 - P_2)^2 + (P_1^3 - P_3) + \cdots\}$. This would have to be truncated to the cube.

In carrying out approximations to Bayes' theorem the curvature of the curve would presumably appear in the expansions of the posterior as in Efron (1975).

Example 2. This illustrates a common problem: we observe a decline in the mortality rate caused by a certain treatment for animals and expect an analogous decline for humans. A simple set-up involves four types

- untreated lab animals
- treated lab animals
- untreated humans
- treated humans.

Supposing that all observations of type i are exchangeable binary variables, de Finetti's theorem gives a representation with four parameters P_1, P_2, P_3, P_4.

There are probably real situations where the untreated survival rates P_1 and P_3 are quite different, but in which the percent improvement P_2/P_1 will be about the same for humans

and animals. Then, the prior will be taken concentrated near the surface

$$\frac{P_2}{P_1} = \frac{P_4}{P_3} \quad \left(\text{or equivalently} \quad \frac{P_2}{P_1 + P_2} = \frac{P_4}{P_3 + P_4} \right).$$

de Finetti continues: "But this can be used to explain more than the mechanism of this particular reasoning: on looking deeper, the very fact of this belief in a near-proportionality should be interpreted in the framework of these same considerations, since it turns essentially on similar observations made for other medical treatments. Let the treatments be $i = 1, 2 \ldots, c$; under an obvious interpretation of the symbols, the conclusion is established if we allow that $P_{4i+1} : P_{4i+2} - P_{4i+3} : P_{4i+4} = 0 \ (i = 1, 2, \ldots, c)$ entails also $P_1 : P_2 = P_3 : P_4 = 0$. The very belief in the plausibility of extending certain conclusions concerning certain medical treatments to certain others can be explained in turn by the observation of analogies in a broader and vaguer sense, and in the same way one can explain every similar belief, which manifests itself by the formulation of a 'statistical law'".

The two examples suggests fresh avenues of research. Notice that they are stated in the language of parameters. Are there representation theorems characterizing such parametric families in terms of conditions on observables?

Consider the first example with families containing only one or two members. One can consider $2n$ single person families $X_1, X_2, \ldots, X_{2n}$ and m two person families $Y_1, Y_2, \ldots, Y_m$. Pair the X's as $(X_1, X_2), (X_3, X_4), \ldots, (X_{2n-1}, X_{2n})$ and let $Z_1 = X_1 \cdot X_2, Z_2 = X_3 \cdot X_4$, etc. Then the Z_i and Y_j are all exchangeable. This gives a characterization.

For the second example, consider $P_1/P_2 = P_3/P_4 = c$. Suppose first that c is known and equal (say) to $1/2$. Priors with $P_3 = P_4/2$ with P_i unknown can be characterized by saying the observables corresponding to P_3 "thinned" with a fair coin flip. This leads to a rather contorted characterization theorem.

The contortions above underscore the idea that parametric representations can be useful. They also point to a failing in our carrying out of de Finetti's program. Most of the extensions of de Finetti's theorem have been on the lines of taking a classical model and finding a Bayesian version.

de Finetti's alarm at statisticians introducing reams of unobservable parameters has been repeatedly justified in the modern curve fitting exercises of today's big models. These seem to lose all contact with scientific reality focusing attention on details of large programs and fitting instead of observation and understanding of basic mechanism. It is to be hoped that a fresh implementation of de Finetti's program based on observables will lead us out of this mess.

REFERENCES

Aldous, D. (1981). Partial exchangeability and $\bar{d}$- topologies. *Exchangeability Probability and Statistics*, (G. Koch and F. Spizzichino, eds.). Amsterdam: North-Holland, 23–38.

Aldous, D. (1985). Exchangeability and related topics. *Springer Lecture Notes in Mathematics* 1117.

Berg, C., Christensen, J. P. J. and Ressel, P. (1984). *Harmonic Analysis on Semigroups*. Berlin: Springer-Verlag.

Coppersmith, D. and Diaconis, P. (1986). Random walk with reinforcement. (To appear).

Dawid, A. P. (1977). Extendibility of spherical matrix distributions. *J. Multivar. Anal.* **8**, 567–572.

de Finetti, B. (1931). Funcione caratteristica di un fenomeno aleatorio. *Atti della R. Accademia Nazionale dei Lincii Ser. 6, Memorie, classe di Scienze, Fisiche, Matamatiche e Naturali* **4**, 251–299.

de Finetti, B. (1938). Sur la condition d'equivalence partielle. *Actualites Scientifiques et Industrielles* 739. Paris: Herman and Cii. Translated in *Studies in Inductive Logic and Probability* II, (R. Jeffrey, ed.). Berkeley: University of California.

de Finetti, B. (1972). *Probability, Induction and Statistics*. New York: Wiley.

Diaconis, P. Eaton, M. and Lauritzen, S. (1988). de Finetti's theorem for the linear model and analysis of variance. *Tech. Rep.* Department of Statistics, University of Minnesota.

Diaconis, P. and Freedman, D. (1980a). de Finetti's theorem for Markov chains. *Ann. Prob.* **8**, 115–130.

Diaconis, P. and Freedman, D. (1980b). de Finetti's generalizations of exchangeability. *Studies in Inductive Logic and Probability* II, (R. C. Jeffrey, ed.). Berkeley: University of California Press.

Diaconis, P. and Freedman, D. (1980c). Finite exchangeable sequences. *Ann. Prob.* **8**, 745–764.

Diaconis, P. and Freedman, D. (1981). On the statistics of vision: the Julesz conjecture. *Jour. Math. Psychol.* **24** 112, 138.

Diaconis, P. and Freedman, D. (1984). Partial exchangeability and sufficiency. *Proc. Indian Statist. Inst. Golden Jubilee Int'l Conf. on Statistics: Applications and New Directions*, (J. K. Ghosh and J. Roy, eds.). Calcutta: Indian Statistical Institute, 205–236.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14**, 1–67.

Diaconis, P. and Freedman, D. (1987). A dozen de Finetti-style results in search of a theory. *Ann. Inst. Henri Poincaré* **23**, 394–423.

Diaconis, P. and Freedman, D. (1988a). Conditional limit theorems for exponential families and finite versions of de Finetti's theorem on exchangeability. (To appear in *Theoretical Prob.*)

Diaconis, P. and Freedman, D. (1988b). On the problem of types. *Tech. Rep.* 153, Department of Statistics, University of California, Berkeley.

Diaconis, P. and Freedman, D. (1988c). On a theorem of Küchler and Lauritzen. *Tech. Rep.* 152, Department of Statistics, University of California, Berkeley.

Diaconis, P. and Shahshahani, M. (1987). The subgroup algorithm for generating uniform random variables. *Prob. In Eng. and Info. Sci.* **1**, 15–32.

Dubins, L. and Freedman, D. (1979). Exchangeable processes need not to be mixtures of independent identically distributed random variables. *Z. Wahr. verw. Geb.* **48**, 115–132.

Dynkin, E. (1978). Sufficient statistics and extreme points. *Ann. Prob.* **6**, 705–730.

Efron, B. (1975). Defining the curvature of a statistical problem. *Ann. Statist.* **6**, 362–376.

Freedman, D. (1962a). Mixtures of Markov processes. *Ann. Math. Statist.* **33**, 114–118.

Freedman, D. (1962b). Invariants under mixing which generalize de Finetti's theorem. *Ann. Math. Statist.* **33**, 916–923.

Freedman, D. (1963). Invariants under mixing which generalize de Finetti's theorem: Continuous time parameter. *Ann. Math. Statist.* **34**, 1194–1216.

Freedman, D. (1984). de Finetti's theorem in continuous time. *Tech. Rep.* 36, Department of Statistics, University of California, Berkeley.

Giblin, P. J. (1977). *Graphs, Surfaces, and Homology: An Introduction to Algebraic Topology*. London: Chapman and Hall.

Hewitt, E. and Savage, L. J. (1955). Symmetric measures on Cartesian products. *Trans. Amer. Math. Soc.* **80**, 470–501.

Kallenberg, O. (1975). Infinitely divisible processes with interchangeable increments and random measures under convolution. *Z. Wahr. verw. Geb.* **32**, 309–321.

Kallenberg, O. (1976). *Random Measures*. Berlin: Academic Press.

Kallenberg, O. (1981). Characterizations and embedding properties in exchangeability. *Tech. Rep.* 10, Department of Mathematics, Chalmers University of Technology, Göteborg, Sweden.

Lauritzen, S. L. (1974). Sufficiency, prediction and extreme models. *Scand. J. Statist.* **2**, 128–134.

Lauritzen, S. L. (1975). General exponential models for discrete observations. *Scand. J. Statist.* **2**, 23–33.

Lauritzen, S. L. (1982). *Statistical Models as Extremal Families*. Aalborg: Aalborg University Press. (To appear in *Springer Lecture Notes in Statistics.*)

Martin, J. J. (1967). *Bayesian Decision Processes and Markov Chains*. New York: Wiley.

Martin-Löf, P. (1970). Statistika Modeller. Notes by Rolf Sundberg. Mimeographed lecture notes.

Pemantle, R. (1988). Random walk with reinforcement on trees. (To appear in *Ann. Prob.*)

Pemantle, R. (1988). Processes with reinforcement. Ph. D. thesis, Department of Mathematics, Massachusetts Institute of Technology.

Ressel, P. (1985). de Finetti-type theorems: an analytic approach. *Ann. Prob.* **13**, 898–922.

Ruelle, D. (1978). *Thermodynamic Formalism*. Reading, MA: Addison-Wesley.

Smith, A. F. M. (1981). On random sequences with centered spherical symmetry. *J. Roy. Statist. Soc. B* 208–209.

Zaman, A. (1984). Urn models for Markov exchangeability. *Ann. Prob.* **12**, 223–229.

Zaman, A. (1986). A finite form of de Finetti's theorem for stationary Markov exchangeability. *Ann. Prob.* **14**, 1418–1427.

DISCUSSION

D. BLACKWELL *(U.C. Berkeley)*

My comments on Professor Diaconis' paper are on three topics: 1) de Finetti's Theorem, 2) finite exchangeability, 3) partial exchangeability. My comments are for $0 - 1$ variables only.

de Finetti's Theorem

If a sequence $X = (X_1, \ldots, X_n)$ is finitely exchangeable then two sequences x, y have the same probability if they have the same number of 1s. Thus we have

$$P(X = x) = p(s) \bigg/ \binom{n}{s},$$

where $s = \Sigma x_i$ and $p(s) = P(\Sigma X_i = s)$. So Diaconis and Freedman (D and F hereafter) represent the distribution P of X by

$$P = \Sigma p(s) H(n, s),$$

where H is the uniform distribution over the $\binom{n}{s}$ sequences of 0s and 1s of length n with s1s.
　　de Finetti's Theorem asserts that every infinite exchangeable sequences $X_1, X_2, \ldots$ is a mixture of i.i.d. Bernoulli processes. D and F's beautiful and constructive approach to this theorem shows that every finitely exchangeable sequence is *nearly* a mixture of i.i.d. Bernoulli sequences. They tell us what mixture, and how near, as follows. With $B(n, s/n)$ the distribution of n i.i.d. Bernoulli variables with parameter s/n, they show that

$$P^* = \Sigma p(s) B(n, s/n)$$

is close to P in the sense that the distributions of $X_1, \ldots, X_k$ under P and P^* are within $4k/n$ of each other. Thus the main step in their proof is a careful assessment of the difference between sampling with replacement: $B(n, s/n)$ and without: $H(n, s)$. I shall return to this assessment at the end.
　　This finite form of de Finetti's Theorem is very welcome. Jimmie Savage once wondered whether the Hewitt-Savage $0 - 1$ law and other $0 - 1$ laws have any finite content. Now that D and F have shown us the finite content de Finetti's Theorem, perhaps they will turn their attention to $0 - 1$ laws.

Finite exchangeability

Finite exchangeability is important on its own, not just as an approach to exchangeability. It already explains much about the relation between frequency and probability. For example suppose that $X_1, \ldots, X_{n+1}$ are finitely exchangeable. We observe $X_1, \ldots, X_n$ and are interested in $P(X_{n+1} = 1 | X_1, \ldots, X_n)$. We have, with $X = (X_1, \ldots, X_n)$, $s = X_1 + \cdots + X_n$, $p(t) = P(X_1 + \cdots + X_{n+1} = t)$.

$$\frac{P(X_{n+1} = 1 | X)}{P(X_{n+1} = 0 | X)} = \frac{p(s+1)/\binom{n+1}{s+1}}{p(s)/\binom{n+1}{s}} = \frac{p(s+1)}{p(s)} \cdot \frac{s+1}{n-s+1}.$$

Thus if, before observing $X_1, \ldots, X_n$, we considered s and $s+1$ about equally likely as values of $X_1 + \cdots + X_{n+1}$, our posterior odds for $X_{n+1} = 1$ are very nearly the frequency odds $s/(n - s)$ for 1s in the first n trials. Bayes may have made this very calculation; according to Steve Stigler [1982], Bayes used the uniform prior distribution precisely because it makes $p(s)$ independent of s.

Partial exchangeability

Partial exchangeability may turn out to be of more practical importance than full exchangeability. For instance if $X = (X_1, \ldots, X_m)$ is a sample of men and $Y = (Y_1, \ldots, Y_n)$ is a sample of women, X tells us something about Y even if the Xs and Ys are not exchangeable with each other. Say that X, Y are *partial exchangeable* if given $s = \Sigma X_i$ and $t = \Sigma Y_j$, the $\binom{m}{s}$ $\binom{n}{t}$ possible X, Y sequences are equally likely. Thus the joint distribution of s and t describes the prior relation between X and Y.

D and F show how to define partial exchangeability in a much more general context, and obtain, for infinite partial exchangeable sequences, an extension of de Finetti's Theorem. I'm not sure, though, that they have fully captured what de Finetti had in mind. He specifically mentioned the possibility that no two variables would be exactly exchangeable, for instance with a sequence of measurements made at different temperatures (another example is estimation of a dose-response curve from responses to a sequence of different doses). Persi, does your general concept of partial exchangeability cover this case?

In the D and F formulation of partial exchangeability, sufficiency plays a central role. For instance in our X, Y example, the pair (s, t) is sufficient for X, Y. Perhaps they have revived the concept for us Bayesians. We haven't needed it up to now, since all Bayes estimates, tests, predictions,... will just naturally come out depending on sufficient statistics only.

A D and F inequality

Denote by $W(A, a)$ the chance that, in drawing a random sample of size a with replacement from a population of size A, we get a different individuals. A basic inequality in the D and F estimate of the difference between sampling with and without replacement is that, for every $A, a, B, b,$

$$W(A, a)W(B, b) \leq W(A + B, a + b - 2).$$

(It is Lemma on p. 748 of D and F [1980], specialized to $c = 2$ and slightly rewritten).

It has the following interpretation. Suppose you have a population of size C and must draw a sample of size c. You may split the population into two subpopulations of any sizes A, B with $A + B = C$ and draw random samples with replacement of any sizes a, b with $a + b = c$ from the two subpopulations. How should you choose, A, B, a, b to maximize your chance of getting c different individuals? The inequality says that your chance does not exceed that with all C in one population, but drawing a sample of only $c - 2$. For c small compared to C, calculations seem to indicate that the D and F inequality is sharp, that how you split C into A, B doesn't matter much, but that c should be split in about the same proportions as C. For $C = 10000$, $c = 100$, here is a table showing, as a function of A, the best $a = a^*$ and the corresponding $P = W(A, a^*)W(B, b^*)$.

A	a^*	P
1	1	.61462
100	1	.61177
500	5	.61160
2000	20	.61163
4000	40	.61147
5000	50	.61162.

The D and F upper bound for P is $W(10000, 98) = .62072$.

Their proof is analytic, using for instance the concavity of $-x \log x$. It would be nice to have a combinatorial proof of their combinatorial inequality.

Thank you, Persi, for what you have taught us and for what you have given us to think about.

SIMON FRENCH (*University of Manchester*)

The Bayesian coin has two faces: probability and utility. Exchangeability ideas are a way of structuring probability distributions so that certain qualitative perceptions in the user's beliefs are represented. They imply the forms of probability models without requiring the user to provide any modelling input in the usual quantitative sense. Within utility theory similar qualitative perceptions, generally called independence conditions, have been shown to imply the form of multi-attribute preference models (Fishburn and Farquhar, 1981, 1982; French, 1986).

I have a gut feeling that there must be a close relationship between probability and exchangeability ideas on the one hand and utility and independence ideas on the other. However, exploring that relationship defeats me. Professor Diaconis said in his presentation that some of the exchangeability ideas "require a fair amount of translation". With my pidgin measure theory and group theory I cannot make that translation. Yet I am sure that relating these two areas of probability and utility is important. Apart from any theoretical cross-fertilisation of ideas that may result, there are practical implications. Decision analysis could be provided with a common approach to structuring probability models and utility functions.

REPLY TO THE DISCUSSION

Professors Blackwell and French have suggested new research projects that I find quite interesting. I do not know any finite version of the zero one laws, Hewitt-Savage, Kolmogorov, or others. I have a nice finite setting to think about what they might mean. In joint work with Freedman (1981) we used Aldous' theorem to construct counter-examples to a set of conjectures about visual perception. The results are full of trivial tail, shift, and partially exchangeable fields. Yet in the end we drew some pictures on a 100 by 100 grid which proved convincing to the experimenters involved. The connection between tail fields and reality remains beyond me, but this applied problem seems like a good place to focus.

The most exciting part of Blackwell's contribution is his suggeston about a purely probabilistic proof of the analytic fact that forms the base of my work with Freedman on finite exchangeability. It seems right to me that there is a proof along the lines that he suggests. I haven't found one.

Aldous and I have started to build a theory of sequences that are almost invariant under permutations, and so almost mixtures of i.i.d. variables. I had hoped to include details here, but there is still too much undone.

Professor French's suggestion is very welcome. There has been far too little emphasis on utility in statistical decision theory. Presumably symmetry considerations can be introduced before the decomposition into probability and utility. Perhaps the split can be made differently. I will take the suggestion to heart and report, at the next Valencia meeting. It is a pleasure to thank both discussants.

REFERENCES IN THE DISCUSSION

Diaconis, P. and Freedman, D. (1980). Finite exchangeable sequence. *Ann. of Prob.* **8**, 745–764.

Farquhar, P. H. and Fishburn, P. C. (1981). Equivalences and continuity in multivalent preference structures. *Ops. Res.* **29**, 282–293.

Fishburn, P. C. and Farquhar, P. H. (1982). Finite degree utility independence. *Math. Ops. Res.* **7**, 348–353.

French, S. (1986). *Decision Theory: An introduction to the Mathematics of Rationality.* Chichester: Ellis Horwood.

Stigler, S. M. (1982). Thomas Bayes's Bayesian inference. *J. Roy. Statist. Soc. A* **145**, 250–258.