PERSI DIACONIS

FINITE FORMS OF de FINETTI'S THEOREM ON EXCHANGEABILITY

ABSTRACT. A geometrical interpretation of independence and exchangeability leads to understanding the failure of de Finetti's theorem for a finite exchangeable sequence. In particular an exchangeable sequence of length r which can be extended to an exchangeable sequence of length k is almost a mixture of independent experiments, the error going to zero like $1/k$.

1. THE GEOMETRY OF INDEPENDENCE

One important difference between the subjective and objective schools is seen in the use of independence. Typically the objective statistician speaks about independent tosses of a coin with unknown probability p. For a subjectivist, probability represents degree of belief, and after viewing 20 tosses of a coin with unknown probability and observing 17 heads, the probability of heads on the 21st toss is likely to be assessed as quite different from the probability on the 1st toss. To say that the probability of heads on the 21st toss is the same as on the 1st toss is possible for a subjectivist; it simply means a refusal to change one's belief from experience, a rather extreme point of view.

The crucial idea of exchangeability allows the subjective statistician to make contact with classical models and techniques. Readable over-views of the use of exchangeability can be found in de Finetti (1964, 1972). A mathematical treatment and extensive bibliography of early work is in Hewitt and Savage (1955). The idea is that before we observe the 20 flips of the coin, if we think of 17 heads and 3 tails as a possible outcome, we don't think of the positions that the 3 heads can occupy as being special. More formally, if we think of heads as 1, tails as 0, and write $P(\ldots)$ to indicate probability, then if we let X_i take values 0 or 1 we say $\{X_i\}_{i=1}^n$ are exchangeable if for every fixed sequence of zeros and ones $\{e_i\}_{i=1}^n$, and for all π,

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

where π denotes a permutation of the set $\{1, 2, \ldots n\}$.

A striking connection between exchangeable sequences and the standard Bayesian treatment of coin tossing is provided by the following theorem.

THEOREM (de Finetti). Let $\{X_i\}_{i=1}^\infty$ be an infinite sequence of random variables with $\{X_i\}_{i=1}^n$ exchangeable for each n; then there is a unique

Synthese **36** (1977) 271–281.

probability measure μ on $[0, 1]$ such that for each fixed sequence of zeros and ones $\{e_i\}_{i=1}^{n}$, we have

$$(1) \qquad P(X_1 = e_1, \ldots, X_n = e_n) = \int_0^1 p^k(1-p)^{n-k}\,\mathrm{d}\mu(p)$$

where

$$k = \sum_{i=1}^{n} e_i.$$

Thus if we envision an infinite exchangeable sequence of coin tosses we may as well behave as if we had a prior measure on the parameter space [0,1] which represents our degree of belief in the probability of heads. The integrated binomial probabilities on the right-hand side of (1) will be referred to as mixtures of independent and identically distributed trials in what follows.

One serious problem with de Finetti's theorem is that it requires an infinite exchangeable sequence and is easily seen to be false if the sequence is finite. For example, let

$$P(X_1 = 1, X_2 = 0) = P(X_1 = 0, X_2 = 1) = \tfrac{1}{2}.$$
$$P(X_1 = 0, X_2 = 0) = P(X_1 = 1, X_2 = 1) = 0.$$

The pair X_1, X_2 is exchangeable, but if there were a prior probability measure μ such that

$$0 = P(X_1 = 1, X_2 = 1) = \int_0^1 p^2\,\mathrm{d}\mu(p),$$

then necessarily μ puts mass 1 at the point 0, so we cannot have

$$0 = P(X_1 = 0, X_2 = 0) = \int_0^1 (1-p)^2\,\mathrm{d}\mu(p).$$

Far from being pathological, the above pair of random variables represents a sample of size 2 drawn without replacement from an urn containing two balls, one marked 0 and the other marked 1.

To get a clearer idea of the extent to which de Finetti's theorem can fail, consider the space $\mathscr{P}_2$ of all possible assignments of probabilities to the four events $P(X_1 = e_1, X_2 = e_2)$. One way to represent these probabilities is as the set of all points (p_1, p_2, p_3, p_4) such that $p_i \geqslant 0$ and $p_1 + p_2 + p_3 + p_4 = 1$.

Let us agree to write

$$p_1 = P(X_1 = 0, X_2 = 0),$$
$$p_2 = P(X_1 = 0, X_2 = 1),$$
$$p_3 = P(X_1 = 1, X_2 = 0),$$
$$p_4 = P(X_1 = 1, X_2 = 1).$$

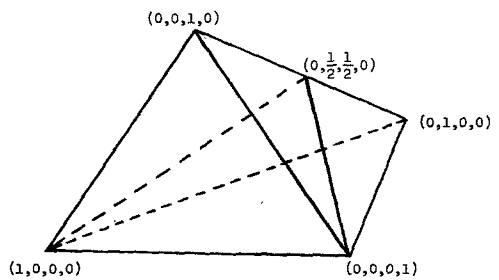

Fig. 1.

Figure 1 shows a picture of $\mathscr{P}_2$ as a tetrahedron. Here we have also indicated the exchangeable point described above, namely $(0, \frac{1}{2}, \frac{1}{2}, 0)$. With this representation, the subclass, $\mathscr{E}_2 \subset \mathscr{P}_2$, of exchangeable probabilities is the set of (p_1, p_2, p_3, p_4) where $p_2 = p_3$. This is a plane connecting the three points $(1, 0, 0, 0)$, $(0, 0, 0, 1)$ and $(0, \frac{1}{2}, \frac{1}{2}, 0)$. An interesting subclass of the exchangeable measures is the class, $\mathscr{I}_2 \subset \mathscr{E}_2$, of independent and identically distributed (i.i.d.) probabilities in $\mathscr{P}_2$. This class may be parametrized as the quadratic, $(p^2, p(1-p),\ p(1-p),\ (1-p)^2)$, lying in $\mathscr{E}_2$ as depicted in Figure 2.

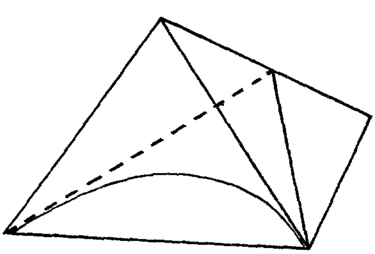

Fig. 2.

The class of mixtures of i.i.d. measures is the convex set beneath the curve. Thus all of the points in the shaded portion in Figure 3 are exchangeable measures for which de Finetti's characterization fails.

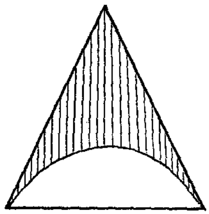

Fig. 3.

de Finetti's theorem yields a unique mixture of i.i.d. measures in the case of an infinite sequence. A glance at Figure 3 shows that points below the shaded region can be represented as a mixture of i.i.d. measures in uncountably many ways.

It is also interesting to draw a picture of the surface of independent probabilities in $\mathscr{P}_2$ which do not necessarily have identical marginal distributions. The doubly ruled surface thus formed is called a hyperbolic paraboloid in older books on geometry.

A similar picture can be contemplated for the space $\mathscr{P}_3$ of all probability assignments to 3 flips of a coin. $\mathscr{P}_3$ is a 7-dimensional space, and the set of exchangeable probabilities, $\mathscr{E}_3$, is a 3-dimensional tetrahedron (which may thus be drawn). The set of i.i.d. probabilities, $\mathscr{I}_3$, is a curve which twists through $\mathscr{E}_3$.

If X_1, X_2 and X_3 are exchangeable, then X_1 and X_2 are exchangeable, and it seems natural to consider the set of all exchangeable pairs (X_1, X_2) which can arise from an exchangeable sequence of length 3. In the next section we will show that this is the portion of $\mathscr{E}_2$ below the line connecting the points $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0)$ and $(0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Thus the shaded portion of Figure 4 represents

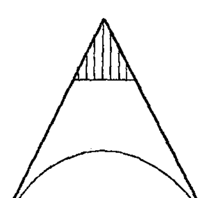

Fig. 4.

exchangeable probabilities which cannot be imbedded in a sequence of length 3. If we continue, asking for points in $\mathscr{E}_2$ which can be imbedded in an exchangeable sequence of length k, we specify a decreasing sequence of curves which closes in on the parabola $\mathscr{I}_2$ (see Figure 5 in Section 2). This will lead to results which say that if an exchangeable sequence $\{X_i\}_{i=1}^{l}$ can be imbedded into an exchangeable sequence of length $k > l$, then the original sequence is almost a mixture of i.i.d. variables with an error going to 0 as k goes to ∞. In the course of proving this, we give a new geometric proof of another finite form of de Finetti's theorem.

The pictures in this section have some connection with the articles of Fienberg and Gilbert (1970) and Fienberg (1968) and suggest a connection between de Finetti's theorem and contingency tables which I hope to explore elsewhere.

2. TWO FINITE FORMS OF de FINETTI'S THEOREM

Let $\mathscr{P}_k$ represent all probabilities on $\Pi_{i=1}^{k} S_i$ where $S_i = \{0,1\}$ for each i. $\mathscr{P}_k$ is a $2^k - 1$ dimensional simplex which is naturally embedded in Euclidian 2^k space. $\mathscr{P}_k$ can be coordinatized by writing $\mathbf{p} = (p_0, p_1, \ldots, p_{2^k-1})$ where p_j represents the probability of the outcome j where $0 \leqslant j < 2^k$ is thought of as the binary representation of the integer j, written with k binary digits. Thus if $k = 3$, $j = 1$ refers to the point 001. Let $S(l, k)$ be the set of j, $0 \leqslant j < 2^k$, with exactly l digits 1. The number of elements in $S(l, k)$ is $\binom{k}{l}$.

Let $\mathscr{E}_k$ be the exchangeable measures in $\mathscr{P}_k$; $\mathscr{E}_k$ is clearly convex as a subset of $\mathscr{P}_k$. The next theorem is a finite form of de Finetti's theorem. This has been given several proofs, all different from the one below. See in particular de Finetti (1972, p. 213), Kendall (1967) and Ericson (1973).

THEOREM 1. $\mathscr{E}_k$ has $k + 1$ extreme points $\mathbf{e}_0, \mathbf{e}_1, \ldots, \mathbf{e}_k$ where $\mathbf{e}_l$ is the measure putting mass $1/\binom{k}{l}$ at each of the coordinates $j \in S(l, k)$ and mass 0 on the other coordinates. Each point $x \in \mathscr{E}_k$ has a unique representation as a mixture of the $k + 1$ extreme points.

Proof. Probabilistically $\mathbf{e}_k$ represents the measure associated with k drawings without replacement from an urn containing k balls, l marked with 1 and $k - l$ marked with 0. Each $\mathbf{e}_k$ is plainly exchangeable. Suppose $\mathbf{e}_k$ is a proper mixture of two other exchangeable points: $\mathbf{e}_k = p\mathbf{a} + q\mathbf{b}$ where

$0 < p < 1$, $q = 1 - p$. Clearly **a** and **b** must both assign probability zero to outcomes to which $\mathbf{e}_k$ gives mass zero. By definition of exchangeability the entries at coordinates $j \in S(l, k)$ in both **a** and **b** are all equal. Since all entries sum to one, entries for these coordinates j must be $1/\binom{k}{l}$. This implies $\mathbf{a} = \mathbf{b}$, so that $\mathbf{e}_k$ is in fact an extreme point of $\mathscr{E}_k$. It is a familiar geometrical fact that every point in a simplex has a unique representation as a mixture of extreme points. □

Theorem 1 is completely natural. If only finitely many events are contemplated, one puts a prior on the set of possible outcomes, and exchangeability forces outcomes with the same number of successes to have the same probability. The prior probabilities are given by the unique set of weights given by the theorem. In this way, Theorem 1 is a finite form of de Finetti's theorem.

One natural situation where finite exchangeable sequences arise is in sampling from finite populations. Versions of Theorem 1 is this context are usefully exploited in Ericson (1973).

While the infinite form of de Finetti's theorem can fail, it may be regarded as a useful approximation to the true state of affairs. Often in contemplating a sequence of length r it is reasonable that the experiments could be continued to yield k observations in all. We now show that if k is large, the original assignment must have almost been a mixture of i.i.d. probabilities.

The projection T: $\mathscr{P}_k \to \mathscr{P}_r$, which takes a point $\mathbf{p} \in \mathscr{P}_k$ into $\mathscr{P}_r$ by using the first r marginal probabilities determined by **p**, is a linear map since it just sums the relevant coordinates of **p** to arrive at $T(\mathbf{p})$. For example, if $r = 2$, then for $\mathbf{p} \in \mathscr{P}_k$,

$$\mathbf{p} \xrightarrow{T} (a_0, a_1, a_2, a_3)$$

where

$$a_i = \sum_j p_j .$$

Here the sum runs over $i2^{k-2} \leqslant j < (i+1)2^{k-2}$.

LEMMA 2. T: $\mathscr{E}_k \to \mathscr{E}_r$ takes $\mathscr{E}_k$ onto the convex region bounded by the $(k+1)$ points $T(\mathbf{e}_l)$, $0 \leqslant l \leqslant k$. $T(\mathbf{e}_l)$ has the interpretation of r balls chosen without replacement from an urn with l balls marked 1 and $k - l$ marked 0.

Proof. This is direct from the definition of T and the fact that linear maps take convex sets into convex sets. The probabilistic description of the extreme points $\mathbf{e}_l$ given in the proof of Theorem 1 leads to the probabilistic interpretation given in the lemma. □

For example, when $r = 2$, the $k + 1$ images are

$$T(\mathbf{e}_l) = \left(\frac{\binom{k-2}{l}}{\binom{k}{l}}, \frac{\binom{k-2}{l-1}}{\binom{k}{l}}, \frac{\binom{k-2}{l-1}}{\binom{k}{l}}, \frac{\binom{k-2}{l-2}}{\binom{k}{l}} \right), \qquad 0 \leqslant l \leqslant k.$$

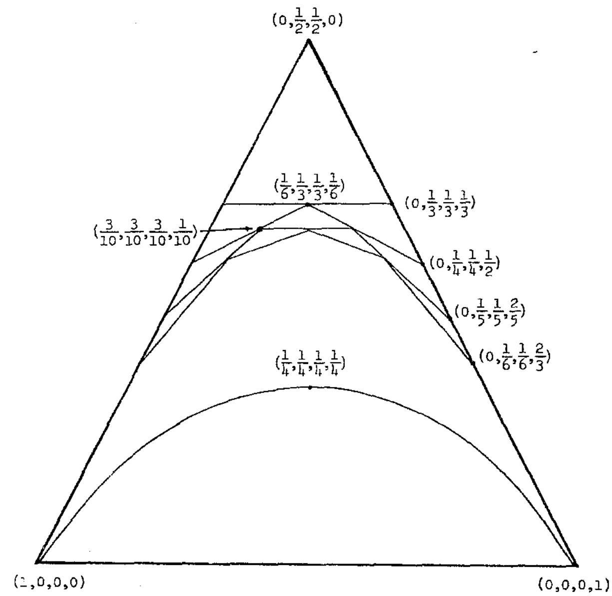

Fig. 5.

Figure 5 shows the various bounding curves for some small values of k when $r = 2$.

The interpretation is that a measure $\mathbf{p} \in \mathscr{E}_2$ which lies above a curve given by the image of the boundary of $\mathscr{E}_k$ cannot be the margin of a measure in $\mathscr{E}_j$ for $j \geqslant k$.

THEOREM 3. Let $\{X_i\}_{i=1}^r$ be an exchangeable sequence which can be extended to an exchangeable sequence of length $k > r$. Then there is a measure μ_k on $[0,1]$ such that if $e_1, \ldots, e_r$ is any sequence of zeros and ones and $h = \Sigma_{i=1}^r e_i$, then

$$\left| P(X_1 = e_1, \ldots, X_r = e_r) - \int_0^1 p^h (1-p)^{r-h} \mathrm{d}\mu_k(p) \right| < \frac{c}{k}$$

where c is a constant which doesn't depend on the sequence e_i.

Proof. For $0 \leqslant l \leqslant k$ let $\mathbf{f}_l$ be $T(\mathbf{e}_l)$. Thus $\mathbf{f}_l = (f_l^0, \ldots, f_l^{2^r-1})$ where, writing $w(j)$ for the number of ones in the binary expansion of the number j,

$$\mathbf{f}_l^j = \frac{\binom{k-r}{l-w(j)}}{\binom{k}{l}}.$$

$\mathbf{f}_l^j$ represents the probability of observing $w(j)$ red balls in the order specified by j in r drawings from an urn containing l red balls and $k - l$ black balls. The binomial approximation to the hypergeometric density shows that

$$\left| f_l^j - \left(\frac{l}{k}\right)^{w(j)} \left(\frac{k-l}{k}\right)^{r-w(j)} \right| \leqslant \frac{c}{k}$$

where c depends on r but not on k, l or j. Let $\mathbf{p} \in \mathscr{E}_k$ be a point representing an extension of $\{X_i\}_{i=1}^r$ to $\mathscr{E}_k$. Let $\{w_i\}_{i=1}^k$ be the barycentric coordinate of $\mathbf{p}$. Thus $\mathbf{p} = \Sigma_{l=0}^k w_l \mathbf{e}_l$ where $w_l \geqslant 0$, $\Sigma_{l=0}^k w_l = 1$. Suppose the sequence $(e_1, e_2, \ldots, e_r)$ given in the statement of the theorem represents the integer j in binary. From Lemma 2 we have $P(X_1 = e_1, \ldots, X_r = e_r) = \Sigma_{l=0}^k w_l f_l^j$. Thus, letting μ_k put mass w_l at the $k + 1$ points l/r in $[0,1]$,

$$\left| p(X_1 = e_1, \ldots, X_r = e_r) - \int_0^1 p^{w(j)} (1-p)^{r-w(j)} \mathrm{d}\mu_k(p) \right|$$

$$\leqslant \Sigma w_l \left| f_l^j - \left(\frac{l}{k}\right)^{w(j)} \left(\frac{k-l}{k}\right)^{r-w(j)} \right| \leqslant \frac{c}{k}$$

as required. □

A straightforward corollary of Theorem 3 is the usual version of de Finetti's theorem.

THEOREM 4. Let $\{X_i\}_{i=1}^r$ be an exchangeable sequence of random variables taking only values zero and one. If, for each $n > r$, $\{X_i\}$ can be embedded in a sequence of length n, then there exists a measure μ on $[0,1]$ such that for each sequence of zeros and ones e_i with $\Sigma_{i=1}^r e_i = h$,

$$p(X_1 = e_1, \ldots, X_r = e_r) = \int_0^1 p^h(1-p)^{r-h} \mathrm{d}\mu(p).$$

Proof. Theorem 3 gives a sequence of measures μ_n on $[0,1]$. By standard weak compactness arguments there is a measure μ and a sequence μ_{k_i} such that for any bounded continuous function f on $[0,1]$, $\int f \,\mathrm{d}\mu_{k_i} \to \int f \,\mathrm{d}\mu$. Noting that the bound given in Theorem 3 does not depend on the choice of the extension of $X_1, \ldots, X_r$ to an exchangeable sequence of length n, we see that

$$\left| p(X_1 = e_1, \ldots, X_r = e_r) - \int p^h(1-p)^{r-h} \mathrm{d}\mu(p) \right| \leqslant \frac{c}{k_i} + \epsilon.$$

Since k_i may be chosen arbitrarily large and ϵ arbitrarily small, this implies the statement of the theorem. □

The extension of the results of this paper to random variables taking on a finite number of values greater than two is straightforward. It is not clear how to carry through a result like Theorem 3 to variables taking values in a more general space such as a complete separable metric space. The fault may well lie with the notion of distance between probabilities used in Theorem 3. Some more tractable (and possibly scientifically more useful) notion of distance such as weak convergence of Bayes risk will be required.

3. CONCLUDING REMARKS

Assumptions of infinite extendability of an exchangeable allocation make certain types of knowledge impossible to incorporate. For example, if we are reasonably sure an urn contains 2 red and 2 black balls, and observe 2 drawings showing red and black, then our probability assignment for the outcome of the next two drawings will be in some neighborhood of the point $(0, \frac{1}{2}, \frac{1}{2}, 0)$ shown in Figure 5. The theory shows that there does not exist a prior on $[0,1]$ which can incorporate the outcome (red, black) into a posterior probability close to the point $(0, \frac{1}{2}, \frac{1}{2}, 0)$.

An example of complete agreement between the finite and infinite views is found in considering flat priors. One way to quantify little prior knowledge has been to assign a uniform prior to [0,1]. If s successes and $f = n - s$ failures are observed in a sample of size n, Bayes' theorem yields Laplace's law of succession: The probability of success on the next trial is $(s + 1)/(n + 2)$. Suppose now that we only assume that the sample of size n could have been extended to an exchangeable sample of size $n + k$ for $k \geqslant 1$ fixed. The theory developed above says that we may behave as if we had a prior distribution over the $n + k + 1$ possible urns. If we quantify our prior knowledge by a uniform prior over the urns, what is the probability of success on the next trial? Curiously, the answer is the same, $(s + 1)/(n + 2)$, no matter what k is. An easy way to see this is to imagine the following scheme for generating a uniform mixture of the $n + k + 1$ possible urns: Choose a probability, p, uniformly in [0,1], and draw a sample of size $n + k$ from a binomial random variable with this parameter p. Then, as Thomas Bayes showed in his original memoir, the number of successes is uniformly distributed between 0 and $n + k$.

de Finetti (1972) cites two papers in Italian that I have been unable to consult. These are de Finetti (1969) and Crisma (1971). Both papers contain conditions under which n exchangeable events can be embedded in a finite sequence of $r > n$ events.

ACKNOWLEDGEMENTS

It is a pleasure to thank Henry Landau and Colin Mallows for useful discussions during a summer stay at Bell Labs. Tom Cover and Ron Giere made many helpful comments on an earlier draft of this paper. This research was partially supported by National Science Foundation Grant MP574–21416.

Stanford University

REFERENCES

Crisma, L.: 1971, 'Sulla proseguibilità di processi scambiabili,' *Rend. Matem. Trieste* **3**, 96–124.

Ericson, W. A.: 1973, 'A Bayesian Approach to 2-Stage Sampling,' Technical Report No. 26, Department of Statistics, University of Michigan, Ann Arbor.

Fienberg, S. E.: 1968, 'The Geometry of an $r \times c$ Contingency Table,' *Ann. Math. Stat.* **39**, 1186–90.

Fienberg, S. E. and Gilbert, J. P.: 1970, 'The Geometry of a Two by Two Contingency Table,' *Jour. Amer. Stat. Assoc.* **65**, 695–701.

de Finetti, B.: 1964, 'Foresight: Its Logical Laws, Its Subjective Sources,' in Kyburg, H. E. and Smokler, H. E. (eds.), *Studies in Subjective Probability*, Wiley, New York.

de Finetti, B.: 1969, 'Sulla proseguibilità di processi aleatori scambiabili,' *Rend. Matem. Trieste* **1**, 53–67.

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

Hewitt, E. and Savage, L. J.: 1955, 'Symmetric Measures on Cartesian Products,' *Trans. Amer. Math. Soc.* **80**, 470–501.

Kendall, D. G.: 1967, 'On Finite and Infinite Sequences of Exchangeable Events,' *Studia Sci. Math. Hung.* **2**, 319–327.