FINITE EXCHANGEABLE PROBABILITY

BY

PERSI DIACONIS and DAVID FREEDMAN

TECHNICAL REPORT NO. 111

MAY 1978

PREPARED UNDER THE AUSPICES

OF

NATIONAL SCIENCE FOUNDATION GRANT MCS77-16974

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
FINITE EXCHANGEABLE PROBABILITY

Persi Diaconis\(^1\) and David Freedman\(^2\)

Stanford University and University of California, Berkeley

ABSTRACT

Let \(X_1, X_2, \ldots, X_k, X_{k+1}, \ldots, X_n\) be an exchangeable sequence of random variables taking values in the set \(S\). The variation distance between the distribution of \(X_1, X_2, \ldots, X_k\) and the closest mixture of independent, identically distributed variables is shown to be at most \(ck/n\), where \(c\) is the cardinality of \(S\). If \(c\) is infinite, the bound \(\frac{1}{2} k(k-1)/n\) is obtained. These results imply the most general known forms of de Finetti's theorem. Examples are given to show that the rates \(k/n\) and \(k(k-1)/n\) cannot be improved.

The main tool is a bound on the variation distance between sampling with and without replacement. For instance, suppose an urn contains \(n\) balls, each marked with some element of the set \(S\) (whose cardinality \(c\) is finite). Now \(k\) draws are made at random from this urn, either with or without replacement. This generates two probability distributions on the set of \(k\)-tuples, and the variation distance between them is at most \(ck/n\).

AMS 1970 Subject Classification: Primary 60C05, Secondary 62A15

Key Words and Phrases: de Finetti's Theorem, Exchangeability, Sampling With and Without Replacement

\(^1\)This author's research was partially supported by National Science Foundation Grant MPS 74-21416 and by the Energy Research and Development Administration under Contract EY-76-C-03-0515.

\(^2\)This author's research was partially supported by National Science Foundation Grant MCS 575-09459.
1. Introduction

De Finetti's theorem involves an infinite sequence

\[ X_i \quad 1 \leq i < \infty \] of exchangeable zero- or one-valued random variables.

De Finetti showed that there is a unique probability measure \( \mu \) on the Borel sets of \([0,1]\) such that

\[
(1) \quad P\{X_i = e_i \text{ for } i = 1, \ldots, k\} = \int_{[0,1]} p^j (1-p)^{k-j} \mu(p), \quad j = \sum e_i.
\]

The main results of this paper concern finite exchangeable sequences \( X_i \quad 1 \leq i \leq n \). As is well known, the representation (1) need not hold exactly for finite exchangeable sequences. For example, let

\[
(2) \quad P(X_1 = 0, X_2 = 1) = P(X_1 = 1, X_2 = 0) = \frac{1}{2}, \quad P(X_1 = 0, X_2 = 0) = P(X_1 = 1, X_2 = 1) = 0.
\]

\( X_1 \) and \( X_2 \) are exchangeable; but if a representation like (1) held,

\[
0 = \int_0^1 p^2 \, d\mu(p) = \int_0^1 (1-p)^2 \, d\mu(p).
\]

This implies that \( \mu \) puts mass 1 both at zero and at one, which is impossible.

However, suppose \( k \) is much smaller than \( n \) and \( X_1, \ldots, X_k \) is the beginning of a long exchangeable sequence \( X_1, \ldots, X_k, X_{k+1}, \ldots, X_n \), then (1) should be approximately true. Our main theorem (3) makes this precise, with the universal error bound \( 2k/n \). And, as we will indicate in Section 4, there is an example where the error is essentially \( \frac{1}{\sqrt{2\pi k}} \) so \( k/n \) is the right order of magnitude.
Our results were motivated by questions in the foundations of probability and are discussed from this point of view in Diaconis and Freedman [1978a].

So far, we have discussed only 0-1 valued random variables. However, similar results apply to variables with values in any finite set. There is also a result for infinite sets, although the form of the bound changes to \( \frac{1}{2} k(k-1) \), as shown in Section 3 below.

To state Theorem (3), let \( S \) be a finite set of cardinality \( c \). Let \( S^k \) be the set of \( k \)-tuples of elements of \( S \). A probability \( P \) on \( S^k \) is said to be exchangeable provided it is invariant under permutations. More precisely, if \( \pi \) is a permutation of \( 1, \ldots, k \), then

\[
P\{(s_1, \ldots, s_k)\} = P\{(s_{\pi(1)}, \ldots, s_{\pi(k)})\}.
\]

To state the analog of (1), let \( S^* \) be the set of probabilities on \( S \); geometrically, \( S^* \) is the unit simplex in \( \mathbb{R}^c \). For \( p \in S^* \), let \( p^k \) be the distribution of \( k \) independent picks from \( p \). Formally, \( p^k \) is the power probability on \( S^k \):

\[
p^k\{(s_1, \ldots, s_k)\} = \prod_{j=1}^{k} p\{s_j\}.
\]

If \( \mu \) is a probability on the Borel subsets of \( S^* \), we define the probability \( P_{\mu^k} \) on \( S^k \) as follows: choose \( p \) at random from \( \mu \), then make \( k \) independent picks from \( p \). Formally,

\[
P_{\mu^k}(A) = \int_{S^*} p^k(A) \mu(dp).
\]
In Section 3 we will consider probabilities $Q$ on $S^n$, and denote by $Q_k$ the projection of $Q$ onto $S^k$. More formally, $Q_k$ is the distribution of $(s_1, \ldots, s_k)$ when the $n$-tuple $(s_1, \ldots, s_k, s_{k+1}, \ldots, s_n)$ is picked from $Q$.

Finally, the variation distance $\|P-Q\|$ is defined in usual

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

(3) **Theorem.** Let $S$ be a finite set of cardinality $c$. Let $P$ be an exchangeable probability on $S^n$, where projection onto $S^k$ is denoted $P_k$. Then there exists a probability $\mu$ on the Borel subsets of $S^*$ such that

$$\|P_k - P_{\mu k}\| \leq ck/n$$

where

$$P_{\mu k} = \int_{S^k} P \mu(dp) .$$

This result is almost immediate from the following estimate.

(4) **Theorem.** Suppose an urn $U$ contains $n$ balls, each marked by one or another element of the set $S$, whose cardinality $c$ is finite. Let $H_{Uk}$ be the distribution of $k$ draws made at random without replacement from $U$, and $M_{Uk}$ be the distribution of $k$ draws made at random with replacement ($H$ stands for hypergeometric; $M$ stands for multinomial). Thus, $H_{Uk}$ and $M_{Uk}$ are two probabilities on $S^k$. Then

$$\|H_{Uk} - M_{Uk}\| \leq ck/n .$$
This result will be proved in Section 2.

Proof of Theorem (3). Each extreme exchangeable probability on $S^n$ is of the form $H_{Un}$ for some urn $U$, and any exchangeable probability $P$ on $S^n$ is a unique mixture:

$$P = \sum U w_U H_{Un},$$

where the sum runs over the finite set of all possible urns of the type considered in Theorem (4), the $w_U$'s being nonnegative weights which sum to 1. This may be proved by conditioning on the order statistics, or see de Finetti (1969), Crisma (1971), Diaconis (1977). Clearly,

$$P_k = \sum w_U H_{U_k}$$

and

$$||P_k - \sum w_U M_{U_k}|| \leq \sum w_U ||H_{U_k} - M_{U_k}|| \leq ck/n.$$

This proves Theorem (3), because $M_{U_k} = p^k_U$, where $p_U \in S^k$ is the distribution of one pick from $U$. \[\]

2. Variation bounds on the difference between sampling with and without replacement

Our first object in this section is prove Theorem (4). Since $U$ and $k$ may be taken as fixed, we abbreviate $h$ for $H_{U_k}$ and $m$ for $M_{U_k}$.

To fix notation, we label $S$ as $\{1, \ldots, c\}$. For $1 \leq i \leq c$, let $n_i$ be the number of balls in $S$ which are labelled $i$, so $\sum_{i=1}^c n_i = n$. Without loss of generality, we may suppose $n_i \geq 1$ for all $i$. 

4
Fix $s \in S^k$; for $i \in S$, let $v_i$ be the number of indices $j \leq k$ with $s_j = i$. Then

$$m(s) = \prod_{i=1}^{c} \left( \frac{n_i}{n} \right)^{v_i} = \frac{1}{n^k} \prod_{i=1}^{c} n_i^{v_i}$$

$$h(s) = \frac{(n-k)!}{n!} \prod_{i=1}^{c} \frac{n_i^{v_i}}{(n_i - v_i)!}.$$ 

The main step in proving (4) is

(6) **Lemma.** $h(s) \prod_{i=1}^{c} \left( 1 - \frac{v_i}{n_i} \right) \leq m(s)$. 

**Proof.** The left side vanishes unless $v_i < n_i$ for all $i$, so assume this condition. Now, $m(s) > 0$, and

$$h(s) \prod_{i=1}^{c} \left( 1 - \frac{v_i}{n_i} \right) = \frac{c}{n^k} \prod_{i=1}^{c} \frac{v_i^{v_i}}{n_i^{v_i}} \cdot \frac{n_i^{v_i}}{(n_i - v_i)!} \cdot \frac{1}{n! / n^{(n-k)!}}.$$ 

For $0 \leq x < 1$, let

$$f(x) = -(1-x)\log(1-x) - x = -\sum_{i=1}^{\infty} \frac{x^{i+1}}{i(i+1)}.$$ 

Then we claim

(8) $$\frac{n!}{n^k(n-k)!} \geq e^{nf(k/n)}.$$ 

Indeed, the logarithm of the left side of (8) is

$$\sum_{j=1}^{k-1} \log(1 - \frac{j}{n}) = - \sum_{i=1}^{\infty} \sum_{i=1}^{k-1} \frac{1}{i(n)^i} \geq nf(k/n),$$
using the elementary estimate

\[
\sum_{j=1}^{k-1} j^i \leq \frac{1}{1+i+1} k^{i+1}.
\]

Similarly, with \(n_i\) for \(n\) and \(v_i\) for \(k\), it can be shown that

\[
\left(1 - \frac{v_i}{n_i}\right) \frac{n_i!}{v_i\left(n_i - v_i\right)!} \leq e \frac{n_i f(v_i/n_i)}{n_i}
\]

replacing (9) by

\[
\sum_{j=1}^{v_i} j^i \geq \frac{1}{1+i+1} v_i^{i+1}.
\]

The factor \(\Pi(1 - v_i/n_i)\) was put in precisely to extend the range of summation from \(v_i - 1\) to \(v_i\).

To complete the proof, notice that \(f\) is concave. By Jensen's inequality

\[
\sum_{i=1}^{c} n_i f(v_i/n_i) \leq n f(k/n).
\]

Proof of Theorem (4). Lemma (6) implies

\[
m(s) - h(s) \geq - h(s) \sum_{i=1}^{c} \frac{v_i}{n_i}.
\]

Writing \(x^- = \max\{-x, 0\}\), we get

\[
(m(s) - h(s))^- \leq h(s) \sum_{i=1}^{c} \frac{v_i}{n_i}.
\]
Now \[ \|m - h\|_s = \sum_s (m(s) - h(s))^2. \]

Recall that \( v_i \) is the number of \( j \leq k \) with \( s_j = i \), so summing the right side of (12) over \( s \) is tantamount to computing a certain expectation, relative to \( h \); since 
\[ E[v_i] = k \frac{n_i}{n}, \]
the sum of the right side is \( ck/n \). \[\square\]

3. Results for general range spaces

In this section we give finite forms of de Finetti’s theorem for variables taking values in an arbitrary space. The most general known infinite version of de Finetti’s theorem are derived by taking limits.

Let \((\Omega, B)\) be a measurable space and write \(\Omega^*\) for the set of probabilities on \((\Omega, B)\). Endow \(\Omega^*\) with the smallest \(\sigma\)-algebra \(B^*\) making \(P \to P(A)\) measurable for all \(A \in B\). By \(P^k\) we mean the product probability on \((\Omega^k, B^k)\); \(P \to P^k(A)\) in \(B^*\) measurable for any \(A \in B^k\).

If \(\mu\) is a probability on \((\Omega^k, B^k)\), define the probability \(P_{\mu^k}\) on \((\Omega^k, B^k)\) as follows:
\[ P_{\mu^k}(A) = \int_{\Omega^k} P^k(A) \mu(dp). \]

The main result of this section can be stated as follows:
(13) **Theorem.** Let $P$ be an exchangeable probability on $(\Omega^n, B^n)$, whose projection into $(\Omega^k, B^k)$ is denoted $P_k$. Then there exists a probability $\mu$ on $(\mathcal{S}^*, B^*)$ such that

$$\|P_k - P_{\mu k}\| \leq \frac{1}{2} k(k-1)/n .$$

**Proof.** For $\omega \in \Omega^n$, let $U(\omega)$ be the urn containing $n$ balls, marked $\omega_1, \ldots, \omega_n$, respectively. Let $H_{U(\omega)k}$ be the distribution of $k$ draws made at random without replacement from $U(\omega)$, while $M_{U(\omega)k}$ is the distribution with replacement. Thus, $H_{U(\omega)k}$ and $M_{U(\omega)k}$ are probabilities on $(\Omega^k, B^k)$; the maps $\omega \mapsto H_{U(\omega)k}(A)$ and $\omega \mapsto M_{U(\omega)k}(A)$ are measurable on $(\Omega^n, B^n)$ for each $A \in B^k$. As is easily verified, the exchangeability of $P$ entails

$$P(A) = \int_{\Omega^n} H_{U(\omega)k}(A) \, P(d\omega) \quad \text{for} \quad A \in B^n .$$

So,

$$P_k(A) = \int_{\Omega^n} H_{U(\omega)k}(A) \, P(d\omega) \quad \text{for} \quad A \in B^k .$$

Clearly,

$$M_{U(\omega)k} = (M_{U(\omega)1})^k .$$

Now $\omega \mapsto M_{U(\omega)1}$ is a measurable map from $(\Omega^n, B^n)$ to $(\Omega^*, B^*)$. Let $\mu$ be the image of $P$ under this map. So

$$P_k(A) - P_{\mu k}(A) = \int_{\Omega^n} [H_{U(\omega)k}(A) - M_{U(\omega)k}(A)] \, P(d\omega) .$$
But Freedman (1977) shows

\[ |H_{U(w)k}^{(A)} - M_{U(w)k}^{(A)}| \leq \frac{1}{2} k(k-1)/n. \]

(14) Remark. This estimate implies the most general known representation theorems for an infinite exchangeable sequence of random variables. For instance, suppose \( \Omega \) is compact Hausdorff, \( B \) is the Baire \( \sigma \)-field in \( \Omega \), and \( P \) is an exchangeable probability on \( (\Omega^\infty, B^\infty) \).

For each \( k \) and \( n \) with \( 1 \leq k \leq n \), Theorem (13) yields a probability \( \mu_{kn} \) on \( (\Omega^*, B^*) \) such that

\[ \|P_k - P_{\mu_{kn}}\| \leq \frac{1}{2} k(k-1)/n. \]

\( \Omega^* \) is compact for weak* convergence. Letting \( n \to \infty \), perhaps through a net, we find a probability \( \mu_k \) on \( (\Omega^*, B^*) \) with \( P_k = P_{\mu_k} \). Letting \( k \to \infty \), we find a \( \mu \) with \( P = P_\mu \); and this is Theorem (7.2) of Hewitt-Savage (1955).

To extend the result to more general spaces we need some notation: For any probability \( \mu \) on \( (\Omega^*, B^*) \), let \( \sim \mu \) be its regular Borel extension. Similarly, if \( Q \) is a probability on \( (\Omega^\infty, B^\infty) \), let \( \sim Q \) be the regular extension to the Borel \( \sigma \)-field \( B \) of \( \Omega^\infty \). By definition,

\[ P_{\mu}^{(A)} = \int_{\Omega^*} P^{\infty} (A) \mu(dp). \]

We now show \( P \to \sim Q \) is Borel for \( A \in B \), and

(15) \[ P_{\mu}^{(A)} = \int_{\Omega^*} P^{\infty} (A) \mu(dp). \]
Proof. Suppose $f_\alpha$ is continuous on $\Omega_\alpha^\infty$ and $f_\alpha \uparrow f$. Then $P \rightarrow \int f_\alpha \, dP_\alpha^\infty$ is continuous, and increases to $P \rightarrow \int f \, dP^\infty_\omega$. Thus,

$$\int f \, dP^\infty_\omega = \int \int f \, dP^\infty_\omega \mu(dp).$$

(See, for instance, Theorem (11.5(IV)), p. 190 in Choquet (1969).)

(16) Remark. The representation (15) may be extended even further, to the $\sigma$-field of $A$'s which are measurable relative to all regular Borel probabilities. As is early verified from (15) itself, $P \rightarrow P^\infty_\omega(A)$ is measurable relative to any regular Borel $\mu$. This leads to a variation on de Finetti's theorem which includes all others we know: For random variables taking values in a universally measurable subset of compact Hausdorff space, equipped with the $\sigma$-field of universally measurable subsets (by "universal measurability" we mean relative to regular Borel probabilities). Olshen (1973) contains related results and references.

It may be noted that Dubins and Freedman (1978) recently gave an example to show that de Finetti's theorem fails for separable metric spaces.

4. The rates are sharp

We begin by showing that in a sense (13) is best possible.

(17) Proposition. Let $\Omega_n$ be the integers $\{1, \ldots, n\}$. For each $n$, there is an exchangeable probability $P$ on $\Omega_n^k$ such that for every probability $\mu$ on $\Omega_n^k$ and $k < n$,
\begin{equation}
\| P_k - P_{\mu_k} \| \geq 1 - \frac{n!/(n-k)!}{n^k} \geq 1 - e^{-\frac{1}{2} \frac{k(k-1)}{n}}.
\end{equation}

(19) Remark. If \( n \) and \( k \) tend to infinity in such a way that \( \frac{k}{\sqrt{n}} \geq \epsilon > 0 \), then (18) shows that no mixture of independent identically distributed variables approximates \( P_k \) since \( \| P_k - P_{\mu_k} \| \to 0 \). If \( k = o(\sqrt{n}) \), then, as shown in the proof of (18), there is a \( \mu \) which minimizes the left side of (18). For this \( \mu \), \( \| P_k - P_{\mu} \| = \frac{1}{2} \frac{k(k-1)}{n} \) so (13) is sharp.

Proof. The \( P \) corresponding to \( n \) assigns equal weight \( 1/n! \) to all \( n! \) permutations of \( \Omega_n^k = \{1, \ldots, n\} \). Thus, \( P \) is a probability on \( \Omega_n^k \), and \( P_k \) is the distribution of \( k \) draws made at random without replacement from \( \{1, \ldots, n\} \). Let \( u \in \Omega_n^k \) give equal weight \( 1/n \) to all elements of \( \Omega_n^k \). We will argue that the \( \mu \) minimizing the left side of (18) is \( \delta_u \), namely, a point mass at \( u \). Granting this, Freedman (1977) shows

\[ \| P_k - u \| \geq 1 - e^{-\frac{1}{2} \frac{k(k-1)}{n}} = \frac{1}{2} \frac{k(k-1)}{n} \]

provided \( k = o(\sqrt{n}) \).

One way to see that \( \delta_u \) minimizes the left side of (18) is to introduce the set

\[ B = \{ \omega \in \Omega_n^k \text{ and } \omega_i = \omega_j \text{ for some } i, j \text{ with } 1 \leq i < j \leq k \} \]

We claim
(20) for any probability \( P \) on \( \{1, \ldots, n\} \), \( P^k(B) \geq u^k(B) \).

This can be deduced from the easily verified Schur convexity of \( 1_B \), using Rinott (1973, p. 68). A direct proof is in Munford (1977).

Relation (20) shows that \( P_{\mu k}(B) \geq u^k(B) \); clearly, \( P_k(B) = 0 \).

So, \( \|P_k - P_{\mu k}\| \geq u^k(B) \). Then, Freedman (1977) shows that

\[
u^k(B) = 1 - \frac{n!/(n-k)!}{\binom{n}{k}} \geq 1 - \frac{1}{2} \frac{k(k-1)}{n}.
\]

We will now argue that the rate \( k/n \) in Theorems (3) and (4) cannot be improved. This is done in two stages. In Proposition (23) we show that the variation distance between sampling with and without replacement from an urn with two colors is asymptotically of order \( k/n \).

In the course of the proof we show that the binomial frequency function is first above, then below, then above the hypergeometric frequency function. The crossover occurs at \( np \pm \sqrt{npq + o(\sqrt{n})} \) (see Figure 1).

In Lemma (36) we show that any mixture of coin tossing has smaller density than coin tossing in this central interval. This leads to Theorem (41): sampling with replacement is asymptotically the closest variation approximation to the hypergeometric among all mixtures.

To state the results, let

\[
\gamma = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |1-u^2| e^{-u^2/2} \, du = \frac{1}{\sqrt{2\pi e}}
\]

(21)

\[
\phi(\alpha) = \frac{1}{2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sqrt{1-\alpha} e^{-u^2/2} \, du \quad \text{for } 0 < \alpha < 1
\]

(22)
(23) **Proposition.** Suppose the urn $U_n$ contains $n$ red balls and $n$ black balls. Let $H_{kn}$ be the distribution of the number of red balls in $2k$ draws made at random without replacement from $U_n$, and let $M_{kn}$ be the distribution with replacement. So, $M_{kn}$ is binomial with parameters $2k$ and $1/2$. Then

\begin{equation}
\|H_{kn} - M_{kn}\| = \gamma \frac{k}{n} + o\left(\frac{k}{n}\right) \text{ if } k = o(n),
\end{equation}

\begin{equation}
\|H_{kn} - M_{kn}\| \to \phi(\alpha) \text{ if } \frac{k}{n} \to \alpha \text{ and } 0 < \alpha < \frac{1}{2}.
\end{equation}

Here, $\gamma$ and $\phi(\alpha)$ are defined by (21) and (22).

The computation is a bit intricate, but the following heuristic may help. Let $\alpha = k/n$. The probabilities $H_{kn}$ and $M_{kn}$ are both essentially normal, with the same mean $k$. The variance of $M_{kn}$ is $2k \cdot 1/4$, and the variance of $H_{kn}$ differs by the finite population correction factor, which is essentially $1 - \alpha$. Thus, $\|H_{kn} - M_{kn}\|$ is essentially the variation difference between two normal distributions with the same mean and variances differing by the factor $1 - \alpha$, namely, $\phi(\alpha)$ motivating (22). But $\phi(\alpha) \approx \gamma \alpha$ for small $\alpha$, motivating (21).

Since the draws from the urn are exchangeable, whether they are made with or without replacement, the variation distance between the two distributions for the draws is exactly equal to the variation distance between the two distributions for the number of reds among the draws. Thus, the variation distance considered in
Theorem (4) coincides with the one in Proposition (23). In particular, the \( k/n \) in Theorem (4) is sharp, although the constant \( c \) is not.

**Proof of Proposition (23).** Suppose \( 2k < n \). For \( -k \leq j \leq k \), let

\[
h(j) = H_{kn}^{k+j} = \binom{n}{k+j} \binom{n}{k-j} \left( \frac{2n}{2k} \right)^{2k},
\]

\[
m(j) = M_{kn}^{k+j} = \binom{2k}{k+j} \left( \frac{1}{2} \right)^{2k}.
\]

As usual, \( h(j)/m(j) = N/D \), where

(26) \[
N = \left( \frac{2n-2k}{n-k+j} \right)^{2n-2k} 
\]

(27) \[
D = \left( \frac{2n}{n} \right)^{2n} \left( \frac{1}{2} \right)^{2n} 
\]

To prove (23), the binomial probabilities \( N \) and \( D \) may be estimated using the local Berry-Esséen theorem, as in Petrov (1975, p. 197).

\[
N = \frac{1}{\sqrt{\pi(n-k)}} e^{-\frac{x^2}{2}} + o\left( \frac{1}{\sqrt{n}} \right), \quad x = \frac{\sqrt{2}j}{\sqrt{n-k}}
\]

\[
D = \frac{1}{\sqrt{\pi n}} + o\left( \frac{1}{\sqrt{n}} \right),
\]

the error term in \( N \) being uniform in \( j \). Thus,

(28) \[
\frac{h(j)}{m(j)} = \frac{N}{D} = \frac{1}{\sqrt{1 - \frac{k}{n}}} e^{-\frac{x^2}{2}} + o(1),
\]

14
the error term being uniform in \( j \). Now \( ||H_{kn} - M_{kn}|| \) is

\[
(29) \quad \frac{1}{2} \sum \left| \frac{h(j)}{m(j)} - 1 \right| m(j) \rightarrow \frac{1}{2} E \left| \frac{1}{\sqrt{1-\alpha}} e^{-\frac{1}{2} x^2 \frac{\alpha}{1-\alpha}} - 1 \right|
\]

where \( X \) is \( N(0,1) \) by the classical Central Limit Theorem applied to \( m(j) \). The right side of (29) is \( \phi(\alpha) \), as defined by (22), by the change of variables \( \frac{x}{\sqrt{1-\alpha}} = U \). This proves (25).

The argument for (24) is similar but more complicated. We assume \( k = o(n) \). The probabilities \( N \) and \( D \) in (25) and (27) must be estimated to within \( o(k/n) \), and this can be done using the Edgeworth expansion. It is allowable to restrict \( j \) to the range

\[ |j| \leq \min\{k, 2\sqrt{k \log(n/k)}\}, \]

for \( M_{kn} \) and \( H_{kn} \) assign mass \( o(k/n) \) outside this range, as in Petrov (1975), p. 205:

\[
(30) \quad D = \frac{1}{\sqrt{n}} \left( 1 + o\left(\frac{1}{n}\right) \right)
\]

\[
(31) \quad N = \frac{1}{\sqrt{n(n-k)}} e^{-\frac{x^2}{2}} \left\{ 1 + \frac{\beta}{n-k} H_4(x) \right\} + o\left(\frac{1}{(n-k)^{3/2}}\right)
\]

\[
= \frac{1}{\sqrt{n(n-k)}} e^{-\frac{x^2}{2}} + 0\left(\frac{1}{(n-k)^{3/2}}\right)
\]

where \( \beta \) is a constant, \( x = \sqrt{2/(n-k)} \) \( j \), the \( o \) is uniform in \( j \), and \( H_4 \) is the fourth Hermite polynomial

\[ H_4(x) = x^4 - 6x^2 + 3. \]

Since \( \sqrt{n/(n-k)} = 1 + \frac{1}{2} \frac{k}{n} + o\left(\frac{k}{n}\right) \), we have
\[ \frac{N_D}{d} = \left(1 + \frac{1}{2} \frac{k}{n}\right) e^{-\frac{x^2}{2}} + o\left(\frac{k}{n}\right). \]

Since \( x^4 = o(k/n) \) uniformly in the range \(|j| \leq 2\sqrt{k \log \frac{n}{k}}\), we have

\[ \frac{h(j)}{m(j)} = \frac{N_D}{d} = 1 + \frac{1}{2} \left(\frac{k}{n} - \frac{2j^2}{n-k}\right) + o\left(\frac{k}{n}\right) = 1 + \frac{1}{2} \frac{k-2j^2}{n} + o\left(\frac{k}{n}\right), \]

the \( o \) being uniform in \(|j| \leq 2\sqrt{k \log \frac{n}{k}}\). Again,

\[ \|H_{jk} - M_{jk}\| = \frac{1}{2} \sum_{|j| \leq 2\sqrt{k \log \frac{n}{k}}} \left| \frac{h(j)}{m(j)} - 1 \right| m(j) \]

\[ + \frac{1}{2} \sum_{|j| > 2\sqrt{k \log \frac{n}{k}}} \left| h(j) - m(j) \right|. \]

The second sum is \( o(k/n) \) by the inequalities of Bernstein (see Petrov (1975, p. 52)), and Hoeffding (1963). Thus,

(32) implies

\[ \frac{n}{k} \|H_{kn} - M_{kn}\| = \frac{1}{4} \sum_{|j| \leq 2\sqrt{k \log \frac{n}{k}}} \left| 1 - \frac{2j^2}{k} \right| m(j) + o(1). \]

Let \( S_{2k} \) be the number of heads in \( 2k \) tosses of a fair coin. The sum on the right of side (33) is essentially \( E|1 - (2(S_{2k} - k)^2)/k| \). This tends to the \( \gamma \) of (21) by De Moivre's Central Limit Theorem; and by observing that \( (S_{2k} - k)^2/k \) is uniformly integrable, because \( E((S_{2k} - k)^4/k^2) \) is uniformly bounded. \( \square \)

(34) Remark. A similar argument, which will not be given in detail, gives results for the non-symmetric case. Let the urn \( U_n \) contain \( r_n \) red balls and \( b_n \) black balls where \( r_n + b_n = n \). If for \( n \) sufficiently
large there is an \( \varepsilon > 0 \) for which \( \varepsilon \leq \frac{r_n}{n} \leq 1 - \varepsilon \), and if \( k_n \to \infty \) with \( \frac{k_n}{n} \to 0 \), then writing \( H_n \) for the hypergeometric probability of \( k_n \) draws from \( U_n \) without replacement and \( M_n \) for the binomial probability of \( k_n \) draws from \( U_n \) with replacement, it can be shown that

\[
||H_n - M_n|| = \gamma \frac{k_n}{n} + o\left(\frac{k_n}{n}\right) \quad \text{with } \gamma \text{ defined in (21)}.
\]

The error term in (35) is uniform if \( \frac{r_n}{n} \) is bounded away from 0 and 1. This assumption is essential since, for an urn with 1 red ball and \( n-1 \) black balls, the variation distance on the left side of (35) can be shown to be of order \( (\frac{k_n}{n})^2 \); and the closest binomial measure \( M_n^k \) is at variation distance \( (\frac{k_n}{n})^3 \).

We next show that no mixture of coin tossing can be closer to the hypergeometric probability \( H_{kn} \) of Proposition (23) than fair coin tossing. \( H_{kn} \) is the probability density of \( S_{2k} \)--the number of red balls in \( 2k \) draws without replacement from an urn containing \( n \) red and \( n \) black balls. If \( \mu \) is a measure on the Borel sets of \( [0,1] \), we write \( P_{\mu_k} \) for the \( \mu \) mixture of coin tossing with parameters \( 2k \) and \( p \in [0,1] \). Let \( S_{\mu_k} \) be the number of red balls in a sample from \( P_{\mu_k} \).

We first show that symmetric measures \( \mu \) give better approximations to \( H_{kn} \) than nonsymmetric measures. Let \( \bar{\mu} \) be the image measure of \( \mu \) under the map \( p \to 1 - p \). The distribution of \( S_{2k} \) is invariant under the transformation \( T \) which switches reds and blacks. \( T \) transforms \( P_{\mu_k} \) to \( P_{\bar{\mu}_k} \). Thus, \( P_{\mu_k} \) and \( P_{\bar{\mu}_k} \) are equidistant from \( H_{kn} \); and since

\[
||H_{kn} - \frac{1}{2}(P_{\mu_k} + P_{\bar{\mu}_k})|| \leq \frac{1}{2} ||H_{kn} - P_{\mu_k}|| + \frac{1}{2} ||H_{kn} - P_{\bar{\mu}_k}|| = ||H_{kn} - P_{\mu_k}||,
\]

17
it follows that the symmetric mixture \( \frac{1}{2} (p_{\mu k} + p_{-\mu k}) = p_{1/2(\mu+\mu)k} \) is closer to \( H_n \) in variation distance than \( p_{\mu k} \).

As an aid to understanding the next lemma and Theorem (41), consider the following picture:

![Diagram of probability distributions](image)

**Figure 1**

First consider the curves representing the probability density of the hypergeometric and fair coin tossing. Elementary considerations, as in Feller (1968, p. 171), show that the coin tossing density in first larger than, then less than, and then larger
than the hypergeometric. Equation (32) shows that the crossover points are at \( k \pm \sqrt{k/2} + o(\sqrt{k}) \). In Lemma (36) we show that the probability density for \( P_{\mu k} \) lies below the density for fair coin tossing in this range. Since the variation distance between \( H_{kn} \) and \( P_{\mu k} \) is the sum of the difference \( H_{kn}(j) - P_{\mu k}(j) \) over \( j \) where the difference is positive, this will imply that fair coin tossing is closer to \( h \) than any symmetric mixture and thus that the rate \( k/n \) in Theorem (3) is best possible.

(36) Lemma. Let \( P_{1/2k} \) denote the binomial distribution with parameters \( 2k \) and \( 1/2 \). Let \( \mu \) be a symmetric measure on the Borel sets of \([0,1]\) and write \( P_{\mu k} \) for the \( \mu \) mixture of coin tossing with parameters \( 2k \) and \( P \). For any \( \varepsilon > 0 \) there is \( K > 0 \) so that \( k > K \) implies

\[
(37) \quad P_{\mu k}(k+j) \leq P_{1/2k}(k+j) \quad \text{if} \quad |j| < \sqrt{\frac{1}{2} - \varepsilon}k.
\]

Proof. By symmetry:

\[
\frac{P_{\mu k}(k+j)}{P_{1/2k}(k+j)} = \int_0^1 f(k,j;p)\mu(dp),
\]

where

\[
f(k,j,p) = [4p(1-p)]^k \left( \frac{p}{1-p} \right)^j + \left( \frac{1-p}{p} \right)^j.
\]

Let \( p = \frac{1+x}{2} \) for \(-1 < x < 1\), so
\[ 4 \ p(1-p) = 1-x^2, \quad \frac{p}{1-p} = \frac{1+x}{1-x}, \quad \frac{1-p}{p} = \frac{1-x}{1+x} \]

and

\[ f(k,j,p) = (1-x^2)^k \left( \frac{1+x}{1-x} \right)^j + \frac{1-x}{1+x} \]

To prove (37) we show the range \([0,1]\) can be broken into four zones by choosing parameters \(\delta\) and \(A\) such that \(f(k,j,p)\) is less than 1 in each zone for \(k\) sufficiently large.

The zones are:

I. \[ |p - \frac{1}{2}| \leq \delta/\sqrt{k} \quad \text{or} \quad |x| < 2\delta/\sqrt{k} \]

II. \[ \delta/\sqrt{k} < |p - \frac{1}{2}| \leq A/\sqrt{k} \quad \frac{2\delta}{\sqrt{k}} < |x| \leq 2A/\sqrt{k} \]

III. \[ A/\sqrt{k} < |p - \frac{1}{2}| \leq \delta \quad \frac{2A}{\sqrt{k}} < |x| \leq 2\delta \]

IV. \[ \delta < |p - \frac{1}{2}| \quad |x| > 2\delta \]

In Zone IV, \( |x| > 2\delta \), it is enough to show that for \(2\delta < x < 1 \)

\[ \left( \frac{1+x}{1-x} \right)^j < \left( \frac{1}{1-x^2} \right)^k \]

that is,

\[ \frac{j}{k} < -\log(1-x^2) \quad \log \frac{1+x}{1-x} \]

(38)

As a function of \(x\), the right hand side of (38) is continuous on \([\delta,1]\) and has a minimum \(\mu > 0\). Thus, choosing \(k\) so large that \(\sqrt{(\frac{1}{2} - \varepsilon)k} < k\mu\) suffices.

20
In Zone III, again by symmetry we need only satisfy (38) for $2A/\sqrt{k} < x < 2\delta$. Then a Taylor series expansion shows
\[
- \log(1-x^2) = \frac{x}{2} + O(x^2)
\]
\[
\log \frac{1+x}{1-x}
\]
uniformly in $0 \leq x \leq \frac{1}{2}$. Thus, for some positive constant $c_1$
\[
- \log(1-x^2) \geq \frac{c_1 x}{2} \geq c_1 A/\sqrt{k}.
\]
The inequality (38) will be satisfied if $j \leq c_1 A/\sqrt{k}$. This can be satisfied by choosing $A$ sufficiently large.

In Zone II, with $\delta$ and $A$ fixed, $y = |x|/\sqrt{k}$ is bounded between $\delta$ and $A$. Hence,
\[
(1-x)^k = e^{-y^2}(1+o(1)), \left(\frac{1+x}{1-x}\right)^j = e^{2jx}(1+o(1))
\]
and
\[
\left(\frac{1-x}{1+x}\right)^j = e^{-2jx}(1+o(1))
\]
The $o$ terms are uniform in $j$ satisfying (37) but depend on $\delta$ and $A$.
The expression $e^{2jx} + e^{-2jx}$ achieves a maximum for $j = \sqrt{\frac{1}{2}}k$ and
there takes value $e^{\sqrt{2}y} + e^{-\sqrt{2}y}$. To show $f(k,j,p) < 1$ we must show
\[
e^{-\frac{y^2}{2}} \frac{e^{\sqrt{2}y} + e^{-\sqrt{2}y}}{2} < 1 \quad \text{or} \quad \frac{e^{\sqrt{2}y} + e^{-\sqrt{2}y}}{2} < e^{y^2}
\]
This can be seen by taking Taylor expansions of both sides and checking term by term.

In Zone I, consider \( 0 < x < 2\delta/\sqrt{k} \).

\[
(1-x^2)^k = 1 - kx^2 + o(kx^2) \quad ,
\]

(39)

\[
\left(\frac{1+x}{1-x}\right)^j = 1 + 2jx + 2j^2x^2 + o(kx^2) \quad ,
\]

(40)

\[
\left(\frac{1-x}{1+x}\right)^j = 1 - 2jx + 2j^2x^2 + o(kx^2) \quad .
\]

The bounds (39) and (40) hold uniformly in \( |j| \leq \sqrt{\frac{1}{2k}} \). Averaging (39) and (40) yields \( 1 + 2j^2x^2 + o(kx^2) \) so that

\[
f(k,j,p) = 1 + (2j^2 - k)x^2 + o(kx^2)
\]

\[
= 1 - 2\varepsilon kx^2 + o(kx^2) < 1 \quad \text{for } \delta \text{ sufficiently small} .
\]

Thus, \( \delta \) should be fixed so that this last inequality holds; then \( A \) should be chosen sufficiently large for the inequality in Zone III to hold. Then \( K \) can be chosen so large that \( k > K \) implies the remaining inequalities are valid. \( \square \)

Returning to the notation of Theorem (3), let \( P \) be the measure induced by \( 2k \) draws without replacement from an urn with \( n \) red and \( n \) black balls with \( n > k \).

(41) Theorem. For any measure \( \mu \) on the Borel sets of \([0,1]\) and any \( \varepsilon > 0 \), if \( P_{\mu k} \) denotes the \( \mu \) mixture of coin tossing with parameters \( 2k \) and \( p \), then \( \|P - P_{\mu k}\| \geq (\gamma - \varepsilon) \frac{k}{n} \) for \( k \) and \( n \) sufficiently large, with \( \gamma \) as in (21).
Proof. As remarked earlier, we may assume that \( \mu \) is symmetric.

If

\[
A = \{ j : |j - k| < \sqrt{\left(\frac{1}{2} - \varepsilon\right) k} \},
\]

then

\[
\| P - P_{\mu k} \| \geq |P(A) - P_{\mu k}(A)| \geq P(A) - P_{\frac{1}{2k}}(A)
\]

Referring to Equation (32), identifying \( P \) with \( h \) and \( P_{\mu k} \) with \( m \), we see that for \( k \) sufficiently large, letting \( A'' = \{ j : |j - k| < \sqrt{\left(\frac{1}{2} + \varepsilon\right) k} \} \), there is a set \( A' \) satisfying \( A \subseteq A' \subseteq A'' \) such that \( P(A') - P_{\frac{1}{2k}}(A') = ||\pi - \pi_{\frac{1}{k}}|| = \gamma \frac{k}{n} + o\left(\frac{k}{n}\right) \). Since the approximations used in (32) are valid for \( A \in A'' - A' \), we see

\[
P(A) - P_{\frac{1}{2k}}(A) = P(A') - P_{\frac{1}{2k}}(A')
\]

\[
+ 0 \left( \frac{k}{n} \sum_{k < |j-k| < \sqrt{\left(\frac{1}{2} + \varepsilon\right) k}} P_{\frac{1}{2k}}(j) \left| 1 - \frac{2j^2}{k} \right| \right)
\]

\[
\geq (\gamma - \varepsilon) \frac{k}{n} \quad \text{for} \quad k \quad \text{and} \quad n \quad \text{sufficiently large.} \]

Acknowledgements. We thank Andre Cabannes and Michael Perlman for their helpful comments.
REFERENCES


