The Annals of Statistics
1975, Vol. 3, No. 6, 1305–1317

EFFICIENT ESTIMATION OF TRANSITION PROBABILITIES IN A MARKOV CHAIN[1]

BY DO SUN BAI

Korea Advanced Institute of Science, Seoul and State University of New York, Oneonta

Let $\{X_1, \cdots, X_N\}$ be N observations on an m-state Markov chain with stationary transition probability matrix $\mathbf{P} = (p_{ij})$, $p_{ij} > 0$, $i, j = 1, \cdots, m$, where N is a random variable. For any parametric function of $\mathbf{P}$, the information inequality gives a lower bound on the variance of an unbiased estimator; attaining the lower bound depends on whether the sampling plan or stopping rule S, the estimator $f = f(X_1, \cdots, X_N)$, and the function $E(f) = g(\mathbf{P})$ are "efficient". All "efficient triples" (S, f, g) are characterized for the Markov chain in which p_{ij} and $p_{i'j'}$ $(i' \neq i)$ are not related functionally. It is also shown that efficient triples do not exist if $m > 2$ and g is a function of two or more rows of $\mathbf{P}$. For the case $m = 2$, efficient triples in which g's are functions of both rows are characterized.

1. Introduction and summary. Let $\{X_1, \cdots, X_N\}$ be N observations on a Markov chain with stationary transition probability matrix $\mathbf{P} = (p_{ij})$, $p_{ij} > 0$, $i, j = 1, \cdots, m$. It is easy to show that when N is fixed there do not exist unbiased estimators for any function of $\mathbf{P}$ (assuming that p_{ij} and $p_{i'j'}$ $(i' \neq i)$ are not functionally related). Thus, any scheme which yields unbiased estimators of functions of $\mathbf{P}$ must be a sequential estimation scheme. Determining such a scheme involves the problem of defining and then finding optimal stopping rules or sampling plans. The most common criterion of optimality when working in unbiased estimation is defined in terms of the variances of the estimators. A lower bound on the variance of an unbiased estimator is given by the (fixed sample or sequential) Cramêr–Rao information inequality. The equality is attained if and only if the sampling plan S, the estimator f, and the expected value $E(f) = g$ are "efficient" in a sense to be specified.

The problem of characterizing the "efficient" triples (S, f, g) has been studied by Girshick, et al. (1946) and DeGroot (1959) for the case of binomial samples, and by Bhat and Kulkarni (1966) for the case of multinomial samples. In this paper we extend their methods to solve the problem of characterizing the efficient triples for an m-state Markov chain. The purpose of this paper is to indicate the limitations of the searches for unbiased estimators of functions of $\mathbf{P}$. We characterize the functions which admit unbiased estimators with "minimum

Received July 1972; revised May 1975.

[1] This paper is an extended version of a portion of the author's doctoral dissertation at The Ohio State University; it was partially supported by the Air Force Office of the Scientific Research, United States Air Force, under AFOSR Grant No. AF-AFOSR-1305-67.

AMS 1970 *subject classifications*. Primary 62M05; Secondary 62L12.

Key words and phrases. Markov chain, transition probability, information inequality, closed sampling plan, similar sampling plan, efficient triple.

variance" (i.e., variance equal to the lower bound) and the corresponding optimal plans and optimal estimators. We took on this task in the spirit of eliminating the searches for "minimum variance" unbiased estimators of some of the parametric functions. We do not propose that all the functions which can be estimated efficiently are necessarily natural ones to estimate in a given application.

In Section 2, the information inequality for Markov chains is derived, and a necessary and sufficient condition for an estimator to be efficient is given. In Section 3, all possible efficient triples (S, f, g) are characterized, where g is a function of a single row of $\mathbf{P}$. The efficient sampling plans are "similar" to those inverse-type sampling plans in which observations are made one by one until $N_{i,\sigma(1)} + \cdots + N_{i,\sigma(k)} = c$ is attained, where N_{ij} is the number of one-step transitions from state i to state j, $1 \leqq k \leqq m$, $(\sigma(1), \cdots, \sigma(m))$ is a permutation of $(1, \cdots, m)$, and c is a preassigned positive integer. Under these sampling plans the estimator $f = \mu_1 N_{i1} + \cdots + \mu_m N_{im}$ is an efficient estimator for the function $g(\mathbf{P}) = c(\mu_1 p_{i1} + \cdots + \mu_m p_{im})/(p_{i,\sigma(1)} + \cdots + p_{i,\sigma(k)})$. Furthermore, it is shown that if g is a function of two or more rows of $\mathbf{P}$ and $m > 2$, then there does not exist any efficient triple. In the case when $m = 2$, the sampling plan in which observations are made one by one until $N_{12} + N_{21} = 2c$ is attained is efficient, and $f = a(N_{11} + N_{12}) + b(N_{21} + N_{22}) + d$ is an efficient estimator for the function $g = c(a/p_{12} + b/p_{21}) + d$.

2. Preliminaries.

In this section we set up necessary notation. A closed sampling plan is defined and a useful lemma is stated without proof. The lemma is used to derive the information inequality for the Markov chain. Finally the notion of efficiency is introduced and the efficient estimators are characterized.

2.1. Information inequality.

Let $\{X_1, X_2, \cdots\}$ be an m-state Markov chain with stationary transition probability matrix $\mathbf{P} = (p_{ij})$, $0 < p_{ij} < 1$, $i, j, = 1, \cdots, m$, $\sum_{j=1}^m p_{ij} = 1$, and initial probability distribution $\Pr[X_1 = k] = q_k$, $0 < q_k < 1$, $k = 1, \cdots, m$, $\sum_{k=1}^m q_k = 1$. Let $\mathbf{P}_i = (p_{i1}, \cdots, p_{im})$, $i = 1, \cdots, m$, be the row vector denoting the ith row of $\mathbf{P}$, and $\mathbf{q} = (q_1, \cdots, q_m)$ be the vector of initial probabilities.

Suppose $\{X_1, \cdots, X_N\}$ is observed, where N is a random variable whose distribution is completely specified by the stopping rule or *sampling plan* under consideration. Define the transition count random variable N_{ij}, $i, j = 1, \cdots, m$, by

$$N_{ij} = n_{ij}(X_1, \cdots, X_N) = \textstyle\sum_{t=2}^N N_{ij}(t) ,$$

where

$$\begin{aligned} N_{ij}(t) &= 1 , \qquad \text{if} \quad X_{t-1} = i \quad \text{and} \quad X_t = j , \\ &= 0 , \qquad \text{otherwise.} \end{aligned}$$

Write $N_{i\cdot} = \sum_{j=1}^m N_{ij}$, $\mathbf{N}_i = (N_{i1}, \cdots, N_{im})$, and $\mathbf{N} = (N_{ij})$. Furthermore define the random variable V_k, $k = 1, \cdots, m$, by

$$\begin{aligned} V_k &= 1 , \qquad \text{if} \quad X_1 = k , \\ &= 0 , \qquad \text{otherwise,} \end{aligned}$$

so that $\Pr[\mathbf{V} = \mathbf{e}_k] = \Pr[X_1 = k] = q_k$, where $\mathbf{V} = (V_1, \cdots, V_m)$ and $\mathbf{e}_k$ is an m-component row vector with unity as the kth component and zeros elsewhere.

Whenever a capital letter is used to denote a random variable (or vector or matrix), the corresponding small letter will denote the value assumed by the random variable (or vector or matrix).

Given a sampling plan S, define

a) $A(S) = \{n : \Pr[N = n \mid S] > 0\}$,
b) $B(S) = \bigcup_{n \in A(S)} \{\mathbf{n} = (n_{ij}) : n_{ij} = n_{ij}(x_1, \cdots, x_n), i, j = 1, \cdots, m\}$,
c) $B^*(S) = \{(\mathbf{v}, \mathbf{n}) : \mathbf{v} \in \{\mathbf{e}_1, \cdots, \mathbf{e}_m\}, \mathbf{n} \in B(S)\}$.

The joint distribution of $(\mathbf{V}, \mathbf{N})$ under S is then given by

$$\begin{aligned} p(\mathbf{v}, \mathbf{n} \mid S) &= \Pr[\mathbf{V} = \mathbf{v}, \mathbf{N} = \mathbf{n} \mid S] \\ &= k(\mathbf{n} \mid \mathbf{v}; S) \cdot (\textstyle\prod_{k=1}^m q_k{}^{v_k}) \cdot (\prod_{i,j=1}^m p_{ij}^{n_{ij}}), \quad (\mathbf{v}, \mathbf{n}) \in B^*(S), \\ &= 0, \qquad \text{otherwise,} \end{aligned} \tag{2.1}$$

where $k(\mathbf{n} \mid \mathbf{v}; S)$ is the number of all possible sequences $(x_1, \cdots, x_n)$, $n \in A(S)$, which yield the same transition counts $\mathbf{N} = \mathbf{n}$, given $\mathbf{V} = \mathbf{v}$.

The following definition of a closed sampling plan is analogous to those of Girshick et al. (1946) and DeGroot (1959).

DEFINITION 1. A sampling plan S will be said to be

i) *closed* if $\sum_{B^*(S)} p(\mathbf{v}, \mathbf{n} \mid S) = 1$,
ii) and *nontrivial* if $\Pr[N \leqq 1 \mid S] = 0$.

In this paper only nontrivial closed sampling plans will be considered.

Under a sampling plan S, an estimate $f(\mathbf{v}, \mathbf{n})$ is a real-valued function defined on $B^*(S)$. It is *unbiased* for its expected value $g(\mathbf{q}, \mathbf{P}) = Ef(\mathbf{V}, \mathbf{N}) = \sum_{B^*(S)} f(\mathbf{v}, \mathbf{n}) \cdot p(\mathbf{v}, \mathbf{n} \mid S)$. This series will be assumed to be absolutely convergent and differentiable termwise with an absolutely convergent derived series for all values of $(\mathbf{q}, \mathbf{P})$.

The objective here is to characterize all sampling plans which admit unbiased estimators of some parametric functions whose variances attain the lower bounds of the information inequality. For the Markov chain described above the information inequality for an estimator $f = f(\mathbf{V}, \mathbf{N})$ of $g(\mathbf{q}, \mathbf{P}) = E(f)$ will be shown to take the form

$$\begin{aligned} \operatorname{Var}(f) \geqq {} & [\textstyle\sum_{k=1}^{m-1} q_k g_k'^2 - (\sum_{k=1}^{m-1} q_k g_k')^2] \\ & + \textstyle\sum_{i=1}^m (EN_{i\cdot})^{-1} \cdot [\sum_{j=1}^{m-1} p_{ij} g_{ij}'^2 - (\sum_{j=1}^{m-1} p_{ij} g_{ij}')^2], \end{aligned} \tag{2.2}$$

where $g_k' = \partial g / \partial q_k$ and $g_{ij}' = \partial g / \partial p_{ij}$ are the partial derivatives of $g(\mathbf{q}, \mathbf{P})$ with respect to q_k and p_{ij}, respectively, for $k, j = 1, \cdots, m - 1$ and $i = 1, \cdots, m$, regarding $q_m = 1 - \sum_{k=1}^{m-1} q_k$ and $p_{im} = 1 - \sum_{j=1}^{m-1} p_{ij}$. For example, for any unbiased estimator f of a transition probability p_{ij}, we have $\operatorname{Var}(f) \geqq p_{ij}(1 - p_{ij}) / (E(N_{i\cdot}))$.

The following lemma will be used to derive (2.2). The proof is straightforward and therefore omitted. (In this section it will be understood that, unless otherwise specified, $i, i' = 1, \cdots, m$ and $k, k', j, j' = 1, \cdots, m-1$.)

LEMMA 1. i) *If* $f = f(\mathbf{V}, \mathbf{N})$ *is an estimator of* $E(f) = g(\mathbf{q}, \mathbf{P})$, *then*

$$E[(q_m V_k - q_k V_m) \cdot f] = q_k q_m g_k' , \tag{2.3}$$
$$E[(p_{im} N_{ij} - p_{ij} N_{im}) \cdot f] = p_{ij} p_{im} g_{ij}' ,$$

where

$$q_m = 1 - \textstyle\sum_{k=1}^{m-1} q_k \qquad \textit{and} \qquad p_{im} = 1 - \textstyle\sum_{j=1}^{m-1} p_{ij} .$$

In particular, with $f = 1$, *we have*

$$E(q_m V_k - q_k V_m) = E(p_{im} N_{ij} - p_{ij} N_{im}) = 0 , \tag{2.4}$$

and

$$E(N_{ij}) = p_{ij} E(N_{i\bullet}) . \tag{2.5}$$

ii)

$$\begin{aligned} E(q_m V_k - q_k V_m)(q_m V_{k'} - q_{k'} V_m) &= q_k q_m (q_k + q_m) , \quad & k' = k , \\ &= q_k q_{k'} q_m , & k' \neq k , \end{aligned} \tag{2.6}$$

and

$$\begin{aligned} E(p_{im} N_{ij} - p_{ij} N_{im})&(p_{i'm} N_{i'j'} - p_{i'j'} N_{im}) \\ &= p_{ij} p_{im} (p_{ij} + p_{im}) \cdot E(N_{i\bullet}) , \quad & i' = i , \ j' = j , \\ &= p_{ij} p_{ij'} p_{im} \cdot E(N_{i\bullet}) , & i' = i , \ j' \neq j , \\ &= 0 , & i' \neq i . \end{aligned} \tag{2.7}$$

iii)

$$E(q_m V_k - q_k V_m)(p_{im} N_{ij} - p_{ij} N_{im}) = 0 . \tag{2.8}$$

To derive (2.2) we first define random variables U_k and W_{ij} as the partial derivatives of the log-likelihood function

$$U_k = \partial \log p(\mathbf{V}, \mathbf{N} \mid S) / \partial q_k = (q_m V_k - q_k V_m) / q_k q_m , \tag{2.9}$$
$$W_{ij} = \partial \log p(\mathbf{V}, \mathbf{N} \mid S) / \partial p_{ij} = (p_{im} N_{ij} - p_{ij} N_{im}) / p_{ij} p_{im} .$$

Using Lemma 1 we obtain their moments;

$$E(U_k) = E(W_{ij}) = 0 , \tag{2.10}$$

$$\begin{aligned} E(U_k U_{k'}) &= 1/q_k + 1/q_m , \quad & k' = k , \\ &= 1/q_m , & k' \neq k , \end{aligned} \tag{2.11}$$

$$\begin{aligned} E(W_{ij} W_{i'j'}) &= (1/p_{ij} + 1/p_{im}) \cdot E(N_{i\bullet}) , \quad & i' = i , \ j' = j , \\ &= (1/p_{im}) \cdot E(N_{i\bullet}) , & i' = i , \ j' \neq j , \\ &= 0 , & i' \neq i , \end{aligned} \tag{2.12}$$

$$E(U_k W_{ij}) = 0 . \tag{2.13}$$

Let

$$\begin{aligned}\mathbf{Z} &= (Z_1, \cdots, Z_{m-1}, Z_m, \cdots, Z_{2m-2}, \cdots, Z_{m(m-1)+1}, \cdots, Z_{(m+1)(m-1)}) \\ &= (U_1, \cdots, U_{m-1}, W_{11}, \cdots, W_{1,m-1}, \cdots, W_{m1}, \cdots, W_{m,m-1}), \\ \mathbf{\Lambda} &= (\mathbf{\Lambda}^{(0)}, \cdots, \mathbf{\Lambda}^{(m)}),\end{aligned}$$

where

$$\begin{aligned}\mathbf{\Lambda}^{(0)} &= (\lambda_1, \cdots, \lambda_{m-1}), \\ \mathbf{\Lambda}^{(i)} &= (\lambda_{i(m-1)+1}, \cdots, \lambda_{(i+1)(m-1)}), \\ \lambda_j &= E(f \cdot Z_j), \qquad j = 1, \cdots, (m+1)(m-1).\end{aligned}$$

Using (2.10) through (2.13), we then obtain

$$\mathbf{\Sigma} = (EZ_jZ_{j'}) = \begin{bmatrix} \mathbf{\Sigma}_0 & 0 \cdots 0 \\ 0 & \mathbf{\Sigma}_1 \cdots 0 \\ \vdots & \vdots \quad \vdots \\ 0 & 0 \quad \mathbf{\Sigma}_m \end{bmatrix},$$

where $\mathbf{\Sigma}_0 = (EU_kU_{k'})$ and $\mathbf{\Sigma}_i = (EW_{ij}W_{ij'})$.

It is then easy to show (for example see Graybill (1969), Chapter 8) that

$$|\mathbf{\Sigma}_0| = (q_1 \cdots q_m)^{-1} > 0, \quad \mathbf{\Sigma}_0^{-1} = (\sigma_0^{kk'}), \tag{2.14}$$

where

$$\begin{aligned}\sigma_0^{kk'} &= q_k(1 - q_k), \quad k' = k, \\ &= -q_kq_{k'}, \quad k' \neq k,\end{aligned}$$

and

$$|\mathbf{\Sigma}_i| = (p_{i1} \cdots p_{im})^{-1} \cdot (EN_{i\bullet})^{m-1}, \quad \mathbf{\Sigma}_i^{-1} = (\sigma_i^{jj'}), \tag{2.15}$$

where

$$\begin{aligned}\sigma_i^{jj'} &= p_{ij}(1 - p_{ij})/(EN_{i\bullet}), \quad j' = j, \\ &= -p_{ij}p_{ij'}/(EN_{i\bullet}), \quad j' \neq j.\end{aligned}$$

Since $p_{ij} > 0$ for all $i, j = 1, \cdots, m$ and the sampling plan S is assumed to be nontrivial, we have $E(N_{i\bullet}) > 0$ for all $i = 1, \cdots, m$. Therefore $|\mathbf{\Sigma}_i| > 0$, and $|\mathbf{\Sigma}| = \prod_{i=1}^m |\mathbf{\Sigma}_i| > 0$. The inverse $\mathbf{\Sigma}^{-1}$ of the matrix $\mathbf{\Sigma}$ is then given by

$$\mathbf{\Sigma}^{-1} = \begin{bmatrix} \mathbf{\Sigma}_0^{-1} & 0 & \cdots & 0 \\ 0 & \mathbf{\Sigma}_1^{-1} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \mathbf{\Sigma}_m^{-1} \end{bmatrix}.$$

We also have, from (2.3), $\mathbf{\Lambda}^{(0)} = (g_1', \cdots, g_{m-1}')$ and $\mathbf{\Lambda}^{(i)} = (g_{i1}', \cdots, g_{i,m-1}')$. We then have

$$\mathbf{\Lambda}\mathbf{\Sigma}^{-1}\mathbf{\Lambda}^t = \textstyle\sum_{i=0}^m \mathbf{\Lambda}^{(i)}\mathbf{\Sigma}_i^{-1}\mathbf{\Lambda}^{(i)t}, \tag{2.16}$$

where $\mathbf{\Lambda}^t$ denotes the transpose of $\mathbf{\Lambda}$ and

$$(2.17)\qquad \mathbf{\Lambda}^{(0)}\mathbf{\Sigma}_0^{-1}\mathbf{\Lambda}^{(0)t} = \sum_{k=1}^{m-1}\sum_{k'=1}^{m-1}\sigma_0^{kk'}g_k'g_{k'}' = \sum_{k=1}^{m-1} q_k g_k'^2 - \left(\sum_{k=1}^{m-1} q_k g_k'\right)^2 ,$$

and similarly

$$(2.18)\qquad \mathbf{\Lambda}^{(i)}\mathbf{\Sigma}_i^{-1}\mathbf{\Lambda}^{(i)t} = [E(N_{i\bullet})]^{-1}\cdot\left[\sum_{j=1}^{m-1} p_{ij}g_{ij}'^2 - \left(\sum_{j=1}^{m-1} p_{ij}g_{ij}'\right)^2\right] .$$

Let $f^* = \mathbf{\Lambda\Sigma}^{-1}\mathbf{Z}^t$. Then we have

$$\mathrm{Var}\,(f^*) = E(f^{*2}) = \mathbf{\Lambda\Sigma}^{-1}\mathbf{\Lambda}^t ,$$
$$\mathrm{Cov}\,(f, f^*) = E(f\cdot f^*) = \mathbf{\Lambda\Sigma}^{-1}\mathbf{\Lambda}^t ,$$

and

$$\mathrm{Corr}^2\,(f, f^*) = \mathbf{\Lambda\Sigma}^{-1}\mathbf{\Lambda}^t/\mathrm{Var}\,(f) .$$

Therefore,

$$(2.19)\qquad \mathrm{Var}\,(f) \geqq \mathbf{\Lambda\Sigma}^{-1}\mathbf{\Lambda}^t .$$

Thus we have the following theorem on the information inequality for a Markov chain.

THEOREM 1. *Under any nontrivial closed sampling plan S, the variance of any unbiased estimator $f = f(\mathbf{V}, \mathbf{N})$ of $E(f) = g(\mathbf{q}, \mathbf{P})$ is bounded below by* (2.2).

Equality holds in (2.2) *if and only if $f(\mathbf{v}, \mathbf{n})$ is a linear function of u_k's and w_{ij}'s for all $(\mathbf{v}, \mathbf{n}) \in B^*(S)$.*

2.2. Efficient estimators. The notion of "efficient" estimators used in this paper is the one introduced by R. A. Fisher in connection with unbiased estimation. The following definition is analogous to the one used by DeGroot (1959) for the binomial case.

DEFINITION 2. i) For a given sampling plan S, a nonconstant estimator $f = f(\mathbf{V}, \mathbf{N})$ is said to be *efficient* for $E(f) = g(\mathbf{q}, \mathbf{P})$ at $(\mathbf{q}^*, \mathbf{P}^*)$ if equality holds in (2.2) when $\mathbf{q} = \mathbf{q}^*$ and $\mathbf{P} = \mathbf{P}^*$. g is then said to be *efficiently estimable* at $(\mathbf{q}^*, \mathbf{P}^*)$.

ii) If f is efficient at all values of $(\mathbf{q}, \mathbf{P})$, then f is said to be efficient for $g(\mathbf{q}, \mathbf{P})$, and g is said to be efficiently estimable.

iii) A sampling plan S is said to be *efficient* if it admits at least one nonconstant efficient estimator.

The following corollary, which is an immediate consequence of Theorem 1, characterizes efficient estimators.

COROLLARY 1. *Under a nontrivial closed sampling plan S, a nonconstant estimator $f = f(\mathbf{V}, \mathbf{N})$ is efficient for $E(f)$ at $(\mathbf{q}, \mathbf{P})$ if and only if there exist constants a_k, b_{ij}, $k, j = 1, \cdots, m-1$, $i = 1, \cdots, m$, not all zero, and d such that*

$$(2.20)\qquad f(\mathbf{v}, \mathbf{n}) = \sum_{k=1}^{m-1} a_k(q_m v_k - q_k v_m) + \sum_{i=1}^{m}\sum_{j=1}^{m-1} b_{ij}(p_{im}n_{ij} - p_{ij}n_{im}) + d$$

for all $(\mathbf{v}, \mathbf{n}) \in B^*(S)$.

3. Efficient estimation of functions of P. In this section the results of Section

2 are applied to various functions of **P**. First we consider estimating functions of a single row of **P**. Two useful lemmas are proved and then additional notation is introduced. *Similar* sampling plans are defined and examples are displayed. Theorems 2 and 3 give the efficient triples for estimating functions of a single row of **P**. Next, the problem of estimating functions of two or more rows of **P** is considered, and it is shown that unless the chain has only two states ($m = 2$) no efficient triples exist. Furthermore for the case $m = 2$, Theorem 4 yields efficient triples.

3.1. Efficient estimation of functions of a row of P. In this subsection we characterize all efficient triples (S, f, g) when g's are functions of a single row of **P** (without loss of generality we take the first row $\mathbf{P}_1 = (p_{11}, \cdots, p_{1m})$ of **P**). Henceforth two distinct values $\mathbf{P}_1^{(0)} = (p_{11}^{(0)}, \cdots, p_{1m}^{(0)})$ and $\mathbf{P}_1^{(1)} = (p_{11}^{(1)}, \cdots, p_{1m}^{(1)})$ of $\mathbf{P}_1$ will be said to be *equivalent with respect to* $g(\mathbf{P}_1)$ if $g(\mathbf{P}_1^{(0)}) = g(\mathbf{P}_1^{(1)})$, and the initial probabilities $\mathbf{q} = (q_1, \cdots, q_m)$ will be regarded as nuisance parameters.

LEMMA 2. *Let S be a given nontrivial closed sampling plan for which there exists a nonconstant estimator f which is efficient for some function $g(\mathbf{P}_1)$ at two values of $\mathbf{P}_1$ which are not equivalent with respect to $g(\mathbf{P}_1)$.*

Then there exist constants $\mu_1, \cdots, \mu_m$, not all zero, and $\xi \neq 0$ such that

$$(3.1) \qquad \mu_1 n_{11} + \cdots + \mu_m n_{1m} = \xi \qquad \textit{for all} \quad \mathbf{n}_1 = (n_{11}, \cdots, n_{1m}) \in B_1(S)\,,$$

where

$$B_i(S) = \{\mathbf{n}_i = (n_{i1}, \cdots, n_{im}) : \mathbf{n} \in B(S)\}\,, \quad i = 1, \cdots, m\,.$$

PROOF. Since g is a function of $\mathbf{P}_1$ alone, Theorem 1 shows that for any unbiased estimator f of g,

$$(3.2) \qquad \mathrm{Var}\,(f) \geqq (EN_{1\bullet})^{-1} \cdot \{\textstyle\sum_{j=1}^{m-1} p_{ij} g_{ij}'^2 - (\sum_{j=1}^{m-1} p_{ij} g_{ij}')^2\}\,,$$

and by Corollary 1 f is efficient at $\mathbf{P}_1$ if and only if there exist constants $a_1, \cdots, a_{m-1}$, not all zero, and b such that

$$(3.3) \qquad f = \textstyle\sum_{j=1}^{m-1} a_j \cdot (p_{1m} n_{1j} - p_{1j} n_{1m}) + b \qquad \text{for all} \quad \mathbf{n}_1 \in B_1(S)\,.$$

Suppose f is efficient at $\mathbf{P}_1^{(0)} = (p_{11}^{(0)}, \cdots, p_{1m}^{(0)})$ and $\mathbf{P}_1^{(1)} = (p_{11}^{(1)}, \cdots, p_{1m}^{(1)})$. Then there exist constants $a_j^{(0)}, a_j^{(1)}, j = 1, \cdots, m-1$, $b^{(0)}$ and $b^{(1)}$ such that

$$\begin{aligned} f &= \textstyle\sum_{j=1}^{m-1} a_j^{(0)} \cdot (p_{1m}^{(0)} n_{1j} - p_{1j}^{(0)} n_{1m}) + b^{(0)} \\ &= \textstyle\sum_{j=1}^{m-1} a_j^{(1)} \cdot (p_{1m}^{(1)} n_{1j} - p_{1j}^{(1)} n_{1m}) + b^{(1)}\,. \end{aligned}$$

From this (3.1) easily follows. Since $b^{(i)} = g(\mathbf{P}_1^{(i)}) = E_i(f)$, $i = 0, 1$, where $E_i(f)$ is the expectation of f when $\mathbf{P}_1 = \mathbf{P}_1^{(i)}$, and $\mathbf{P}_1^{(0)}$ and $\mathbf{P}_1^{(1)}$ are not equivalent with respect to $g(\mathbf{P}_1)$, we have $\xi \neq 0$, and not all μ_j's are zeros.

Lemma 2 shows that if condition (3.1) is not satisfied, then f cannot be efficient at two or more nonequivalent values of $\mathbf{P}_1$.

LEMMA 3. *Let S be a nontrivial closed sampling plan for which* $\Pr[N_{1\bullet} = 0 \mid S] = 0$.

Then $\mathbf{N}_1 = (N_{11}, \cdots, N_{1m})$ *can be written as*

$$\mathbf{N}_1 = \textstyle\sum_{k=1}^{N_{1\cdot}} \mathbf{Z}_k \, , \tag{3.4}$$

where $\{\mathbf{Z}_k\}$ *are independently and identically distributed random vectors with*

$$\Pr[\mathbf{Z}_1 = \mathbf{e}_j] = p_{1j} \, , \qquad j = 1, \cdots, m \, . \tag{3.5}$$

PROOF. Define random variables $\{M_k\}$ by

$$\begin{aligned} M_1 &= \min\{t : X_t = 1\} \, , \\ M_k &= \min\{t : X_t = 1, t > M_{k-1}\} \, , \qquad k = 2, 3, \cdots \, . \end{aligned} \tag{3.6}$$

Then since $\Pr[N_{1\cdot} = 0 \mid S] = 0$, $q_k > 0$, and $p_{1j} > 0$, $j, k = 1, \cdots, m$, we have $\Pr[M_k < \infty \mid S] = 1$, $k = 1, 2, \cdots$.

Let $\mathbf{Z}_k$, $k = 1, 2, \cdots$ be an m-component random vector defined by

$$\mathbf{Z}_k = \mathbf{e}_j \quad \text{if} \quad X_{M_k+1} = j \, , \qquad j = 1, \cdots, m \ \text{ and } \ k = 1, 2, \cdots \, . \tag{3.7}$$

That is, $\mathbf{Z}_k = \mathbf{e}_j$ if the kth visit to state 1 by the chain is followed by a visit to state j. Then we obtain

$$\begin{aligned} \Pr[\mathbf{Z}_k = \mathbf{e}_j] &= \Pr[X_{M_k+1} = j] \\ &= \textstyle\sum_{m_k} \Pr[X_{m_k+1} = j \mid M_k = m_k] \cdot \Pr[M_k = m_k] \\ &= p_{1j} \cdot \textstyle\sum_{m_k} \Pr[M_k = m_k] \\ &= p_{1j} \, , \qquad j = 1, \cdots, m \, , \end{aligned}$$

and it can easily be seen that $\{\mathbf{Z}_k\}$ are independent. Thus, $\{\mathbf{Z}_k\}$ are independently and identically distributed random vectors with distribution (3.5). Furthermore, it can be seen, from the definition of $\mathbf{Z}_k$, that $\mathbf{N}_1$ can be represented as the sum (3.4) of $N_{1\cdot}$ independent observations on a random vector $\mathbf{Z}$ with the same distribution as given in (3.5).

The following notation and definition will be used in characterizing efficient sampling plans.

Let c and c^* be preassigned positive integers and $W = w(X_1, \cdots, X_n)$ be a random variable such that $w = w(x_1, \cdots, x_n)$, $n \in A(S)$, is a positive integer-valued function of $(x_1, \cdots, x_n)$, nondecreasing in n. Let $W^* = w^*(X_1, \cdots, X_n)$ be another such random variable. $S(c; W)$ will denote the sampling plan in which observations are continued until exactly $W = c$ is attained.

DEFINITION 3. A sampling plan $S(c^*; W^*)$ will be said to be *similar* to $S(c; W)$ if $w = w(x_1, \cdots, x_n) = c$ for all $(x_1, \cdots, x_n)$ whenever $w^* = w^*(x_1, \cdots, x_n) = c^*$ and $n \in A[S(c^*; W^*)]$; that is, if $W = c$ at the termination of the sampling under $S(c^*; W^*)$.

EXAMPLE 1. Let

$$\begin{aligned} X^{(i)} &= x^{(i)}(X_1, \cdots, X_N) \\ &= \text{number of times that} \quad X_t = i \, , \quad t = 1, \cdots, N \, . \end{aligned}$$

Then we have

$$X^{(i)} = \textstyle\sum_{t=1}^{N} I_{\{i\}}(X_t) = N_{1\cdot} + I_{\{i\}}(X_N)\,, \tag{3.8}$$

where $I_{\{\cdot\}}$ is an indicator function.

Now, the plan $S(c+1; X^{(i)})$ can be seen to have the property that

1. $x_n = i$ for all $n \in A[S(c+1; X^{(i)})]$,
2. $n_{i\cdot} = c$ for all $\mathbf{n}_i \in B_i[S(c+1; X^{(i)})]$.

Thus, $S(c+1; X^{(i)})$ is *similar* to $S(c; N_{i\cdot})$.

EXAMPLE 2. Let

$$\begin{aligned} R^{(i)} &= r^{(i)}(X_1, \cdots, X_N) \\ &= \text{number of runs of } i\text{'s in } N \text{ observations.} \end{aligned}$$

Then we have

$$R^{(i)} = \textstyle\sum_{j=1, j\neq i}^{m} N_{ij} + I_{\{i\}}(X_N)\,. \tag{3.9}$$

The plan $S(c+1; R^{(i)})$ can be seen to have the property that

1. $x_n = i$ for all $n \in A[S(c+1; R^{(i)})]$,
2. $\sum_{j=1, j\neq i}^{m} n_{ij} = c$ for all $\mathbf{n}_i \in B_i[S(c+1; R^{(i)})]$.

Thus, $S(c+1; R^{(i)})$ is *similar* to $S(c; \sum_{j=1, j\neq i}^{m} N_{ij})$.

Bhat and Kulkarni (1966) have extended the result of DeGroot (1959) for the binomial population to the multinomial case, and have characterized efficient sampling plans, which turn out to be single (fixed sample size) sampling plans or inverse multinomial sampling plans (Tweedie (1952)). The following theorem is a consequence of Lemmas 2 and 3 and the Bhat–Kulkarni result (their Theorem 2)[2] and is presented without proof.

THEOREM 2. *Let S be a given nontrivial closed sampling plan such that*

1. $\Pr[N_{1\cdot} = 0 \mid S] = 0$,
2. *there exist constants $\mu_1, \cdots, \mu_m$, not all zero, and $\xi \neq 0$ such that*

$$\mu_1 n_{11} + \cdots + \mu_m n_{1m} = \xi \qquad \textit{for all } \mathbf{n}_1 \in B_1(S)\,.$$

Then S is either $S(c; \sum_{k=1}^{k_0} N_{1,\sigma(k)})$ for some positive integer c or a sampling plan similar to it, where $(\sigma(1), \cdots, \sigma(m))$ is a permutation of $(1, \cdots, m)$ and k_0 is an integer, $1 \leqq k_0 \leqq m$.

From Lemmas 2 and 3 and Theorem 2, it is seen that the only sampling plans which could be efficient for some function $g(\mathbf{P}_1)$ are $S(c; \sum_{k=1}^{k_0} N_{1,\sigma(k)})$ or plans similar to them. The following theorem shows that these plans are indeed efficient. This, combined with Lemma 2, will then imply that a nonconstant estimator f is efficient for $g(\mathbf{P}_1)$ for all values of $\mathbf{P}_1$ if and only if it is efficient at two distinct nonequivalent values of $\mathbf{P}_1$.

[2] To be precise, their Lemma 1 is in error. However, their Theorem 2, which partially depends on Lemma 1, is essentially true.

THEOREM 3. *The sampling plan* $S(c;\ \sum_{k=1}^{k_0} N_{1,\sigma(k)})$, $1 \leqq k_0 \leqq m$, *or plans similar to it are efficient and any nonconstant function* f *of the form*

$$f = \mu_1 N_{11} + \cdots + \mu_m N_{1m} \tag{3.10}$$

is efficient for

$$c(\mu_1 p_{11} + \cdots + \mu_m p_{1m})/(p_{1,\sigma(1)} + \cdots + p_{1,\sigma(k_0)})\,. \tag{3.11}$$

PROOF. It will be proved that $S(c;\ \sum_{j=1}^{k_0} N_{1j})$, $1 \leqq k_0 \leqq m$, is efficient and (3.10) is the efficient estimator for

$$g(\mathbf{P}_1) = c(\mu_1 p_{11} + \cdots + \mu_m p_{1m})/(p_{11} + \cdots + p_{1k_0})$$

by showing that there exist constants a_j, $j = 1, \cdots, m-1$, and b such that equation (3.3) holds for all $\mathbf{n}_1 \in B_1[S(c;\ \sum_{j=1}^{k_0} N_{1j})]$. Write

$$p_{1j} n_{1m} = -(p_{1m} n_{1j} - p_{1j} n_{1m}) + p_{1m} n_{1j}\,, \qquad j = 1, \cdots, k_0\,. \tag{3.12}$$

Since

$$\sum_{j=1}^{k_0} n_{1j} = c \qquad \text{for all} \quad \mathbf{n}_1 \in B_1[S(c;\ \textstyle\sum_{j=1}^{k_0} N_{1j})]\,,$$

summing both sides of (3.12), we obtain

$$(p_{11} + \cdots + p_{1k_0}) \cdot n_{1m} = -\sum_{j=1}^{k_0} (p_{1m} n_{1j} - p_{1j} n_{1m}) + c p_{1m}\,.$$

Then

$$\begin{aligned} n_{1m} = &-\sum_{j=1}^{k_0} (p_{11} + \cdots + p_{1k_0})^{-1} \cdot (p_{1m} n_{1j} - p_{1j} n_{1m}) \\ &+ c p_{1m}(p_{11} + \cdots + p_{1k_0})^{-1}\,. \end{aligned} \tag{3.13}$$

Write

$$p_{1m} n_{1j} = (p_{1m} n_{1j} - p_{1j} n_{1m}) + p_{1j} n_{1m}\,,$$

or

$$n_{1j} = p_{1m}^{-1} \cdot (p_{1m} n_{1j} - p_{1j} n_{1m}) + p_{1m}^{-1} p_{1j} n_{1m}\,.$$

Then

$$\begin{aligned} f = \sum_{j=1}^{m} \mu_j n_{1j} = &\sum_{j=1}^{m-1} \mu_j p_{1m}^{-1} \cdot (p_{1m} n_{1j} - p_{1j} n_{1m}) \\ &+ p_{1m}^{-1}(\textstyle\sum_{j=1}^{m} \mu_j p_{1j}) \cdot n_{1m}\,. \end{aligned} \tag{3.14}$$

Substituting (3.13) into (3.14), we get, after some simplification,

$$\begin{aligned} f = &\sum_{j=1}^{k_0} p_{1m}^{-1}[\mu_j - (p_{11} + \cdots + p_{1k_0})^{-1} \cdot \textstyle\sum_{k=1}^{m} \mu_k p_{1k}] \cdot (p_{1m} n_{1j} - p_{1j} n_{1m}) \\ &+ \sum_{j=k_0+1}^{m-1} \mu_j p_{1m}^{-1}(p_{1m} n_{1j} - p_{1j} n_{1m}) \\ &+ c(p_{11} + \cdots + p_{1k_0})^{-1} \cdot \textstyle\sum_{j=1}^{m} \mu_j p_{1j}\,, \end{aligned} \tag{3.15}$$

which is of the form (3.3). Hence f is efficient for

$$E(f) = c(p_{11} + \cdots + p_{1k_0})^{-1} \cdot \sum_{j=1}^{m} \mu_j p_{1j} = g(\mathbf{P}_1)\,.$$

We note that, in particular, $S(c;\ N_{1\bullet})$ is efficient and (3.10) is an efficient estimator for $c \cdot \sum_{j=1}^{m} \mu_j p_{1j}$.

3.2. Efficient estimation of functions of two or more rows of P. We now consider the problem of characterizing efficient triples (S, f, g) when g's are functions of two rows of $\mathbf{P}$, say $\mathbf{P}_1$ and $\mathbf{P}_2$.

We note that, when p_{ij} and $p_{i'j'}$ $(i' \neq i)$ are not functionally related, a) terms involving p_{ij}'s in the joint probability displayed in (2.1) can be factored out so that terms involving $\mathbf{P}_i = (p_{i1}, \cdots, p_{im})$ involve only $\mathbf{n}_i = (n_{i1}, \cdots, n_{im})$, and b) the lower bound in the information inequality (2.2) and the form of efficient estimators (2.20) have an additive property in the sense that they are, in part, sums of terms involving $\mathbf{P}_i$ and $\mathbf{n}_i$ only. These facts plus Theorem 2 indicate that in order for a sampling plan S to be efficient for some $g(\mathbf{P}_1, \mathbf{P}_2)$, S must have the property that for some fixed positive integers c_1 and c_2,

$$\sum_{k=1}^{k_0} n_{1,\sigma(k)} = c_1 \quad \text{and} \quad \sum_{\delta=1}^{\delta_0} n_{2,\sigma(\delta)} = c_2 \tag{3.16}$$

for all $(\mathbf{n}_1, \mathbf{n}_2)$ for which $\mathbf{n} \in B(S)$, where k_0 and δ_0 are positive integers, $1 \leqq k_0$, $\delta_0 \leqq m$.

However, it is easy to see that condition (3.16) cannot be satisfied if $m > 2$. The case $m = 2$ will be discussed in the next subsection.

Thus, if $m \geqq 3$, there does not exist a sampling plan that is efficient for g whenever g is a function of two or more rows of $\mathbf{P}$.

3.3. Dependent Bernoulli trials (the case when $m = 2$). Let $\{X_1, \cdots\}$ be Markov dependent Bernoulli trials such that

$$\begin{aligned} &\Pr[X_1 = 1] = 1 - \Pr[X_1 = 0] = p\,, && 0 < p < 1\,, \\ &\Pr[X_t = 1 \mid X_{t-1} = 1] = \alpha\,, && 0 < \alpha < 1\,, \\ &\Pr[X_t = 1 \mid X_{t-1} = 0] = \beta\,, && 0 < \beta < 1\,. \end{aligned}$$

Write $1 - p = q$, $1 - \alpha = \bar{\alpha}$, and $1 - \beta = \bar{\beta}$. We note that we have here simplified the notation so that p, q, α, $\bar{\alpha}$, β, $\bar{\beta}$, x_1, n_{11}, n_{10}, n_{01}, n_{00} are q_1, q_2, p_{11}, p_{12}, p_{21}, p_{22}, v_1, n_{11}, n_{12}, n_{21}, n_{22}, respectively, of the earlier sections.

Then the joint probability (2.1) becomes

$$p(x_1, \mathbf{n} \mid S) = k(\mathbf{n} \mid x_1; S) \cdot p^{x_1} q^{1-x_1} \alpha^{n_{11}} \bar{\alpha}^{n_{10}} \beta^{n_{01}} \bar{\beta}^{n_{00}}\,, \tag{3.17}$$

the information inequality (2.2) reduces to

$$\begin{aligned} \operatorname{Var}(f) \geqq pq g_p'^2(p, \alpha, \beta) &+ \alpha\bar{\alpha} g_\alpha'^2(p, \alpha, \beta)/E(N_{1\bullet}) \\ &+ \beta\bar{\beta} g_\beta'^2(p, \alpha, \beta)/E(N_{0\bullet})\,, \end{aligned} \tag{3.18}$$

and the efficient estimators (2.20) have the form

$$f(x_1, \mathbf{n}) = a(x_1 - p) + b_1(\bar{\alpha} n_{11} - \alpha n_{10}) + b_2(\bar{\beta} n_{01} - \beta n_{00}) + d\,. \tag{3.19}$$

Theorems 2 and 3 show that a given nontrivial closed sampling plan S is efficient (for functions of α alone) if and only if S is either one of $S(c; N_{1\bullet})$, $S(c; N_{11})$, $S(c; N_{10})$ or a plan similar to one of them. The efficient estimators f and efficiently estimable functions $g(\alpha)$, under these plans, are

$$\begin{aligned} &S(c; N_{1\bullet})\,; && f = aN_{11} + b\,, && g(\alpha) = ca\alpha + b\,, \\ &S(c; N_{11})\,; && f = aN_{1\bullet} + b\,, && g(\alpha) = ca/\alpha + b\,, \\ &S(c; N_{10})\,; && f = aN_{1\bullet} + b\,, && g(\alpha) = ca/\bar{\alpha} + b\,. \end{aligned} \tag{3.20}$$

Note that if we let $X = x(X_1, \cdots, X_N) = \sum_{t=1}^{N} X_t$, then $S(c; X)$, a sampling plan in which trials are continued until exactly c successes are attained, is similar to $S(c-1; N_{1\bullet})$.

By replacing $N_{1\bullet}$, N_{11}, N_{10} and α by $N_{0\bullet}$, N_{00}, N_{01} and $\bar{\beta}$, we obtain the characterization of efficient triples (S, f, g) when g's are functions of β alone.

We now consider the problem of characterizing efficient triples when g's are functions of both α and β.

Out of the nine possible combinations between $n_{1\bullet}$, n_{11}, $n_{10} = c_1$ and $n_{0\bullet}$, n_{01}, $n_{00} = c_2$ which satisfy the condition (3.16), the only one that a sampling plan can possibly satisfy is $(n_{10} = c_1; n_{01} = c_2)$. Moreover, for any sampling plan S,

$$n_{10} - n_{01} = x_1 - x_n \qquad \text{for all} \quad n \in A(S) . \tag{3.21}$$

Thus,

$$x_1 - x_n = c_1 - c_2 .$$

This implies that $x_1 = x_n = 0$ or 1 for all $n \in A(S)$. In that case, $c_1 = c_2 = c$. That is, $n_{10} = n_{01} = c$ or equivalently $n_{10} + n_{01} = 2c$ for all (n_{10}, n_{01}) for which $\mathbf{n} \in B(S)$.

THEOREM 4. *The only efficient sampling plan for functions of both parameters α and β is $S(2c; N_{10} + N_{01})$ or plans similar to it, and any nonconstant function f of the form*

$$f = aN_{1\bullet} + bN_{0\bullet} + d \tag{3.22}$$

is an efficient estimator for

$$g(\alpha, \beta) = c(a/\bar{\alpha} + b/\beta) + d . \tag{3.23}$$

In particular, $f = N$ is efficient for $g(\alpha, \beta) = c(1/\bar{\alpha} + 1/\beta) + 1$.

PROOF. The above arguments show that the only sampling plan that could be efficient is $S(2c; N_{10} + N_{01})$ (or plans similar to it). We now show that this plan is indeed efficient. Since $n_{10} = n_{01} = c$ for all (n_{10}, n_{01}) for which $\mathbf{n} \in B[S(2c; N_{10} + N_{01})]$, we have

$$\bar{\alpha} n_{11} - \alpha n_{10} = \bar{\alpha} n_{1\bullet} - c.$$

Hence

$$n_{1\bullet} = (1/\bar{\alpha})(\bar{\alpha} n_{11} - \alpha n_{10}) + c/\bar{\alpha} .$$

Similarly

$$n_{0\bullet} = (-1/\beta)(\bar{\beta} n_{01} - \beta n_{00}) + c/\beta .$$

Thus,

$$\begin{aligned} f &= a n_{1\bullet} + b n_{0\bullet} + d \\ &= (a/\bar{\alpha})(\bar{\alpha} n_{11} - \alpha n_{10}) + (-b/\beta)(\bar{\beta} n_{01} - \beta n_{00}) + c(a/\bar{\alpha} + b/\beta) + d \end{aligned}$$

for all $\mathbf{n} \in B[S(2c; N_{10} + N_{01})]$, which is of the form (3.19). Hence f is efficient for

$$E(f) = c(a/\bar{\alpha} + b/\beta) + d = g(\alpha, \beta) .$$

4. Conclusions. The purpose of this paper has been to display the restrictions on the problem of obtaining minimum variance unbiased (m.v.u.) estimators for functions of the transition probabilities in a Markov chain. We have characterized those parametric functions and corresponding sampling plans which admit m.v.u. estimators. We have not claimed that these functions are necessarily natural ones to estimate in a given application. However it is not difficult to visualize a situation where we would want to estimate one of these efficiently estimable functions. For example, in an experiment using Bernoulli trials, we might want to estimate the probability ratio $\bar{\alpha}/\alpha = 1/\alpha - 1$, in which case $S(c; N_{11})$ would be a good sampling scheme to adopt.

Acknowledgments. The author is grateful to Professor J. S. Rustagi for his guidance and encouragement, and to L. S. Mayer and the referee for their valuable comments and suggestions.

REFERENCES

[1] BHAT, B. R. and KULKARNI, N. V. (1966). On efficient multinomial estimation. *J. Roy. Statist. Soc. Ser. B* **28** 45-52.

[2] DEGROOT, M. H. (1959). Unbiased sequential estimation for binomial population. *Ann. Math. Statist.* **30** 80-101.

[3] GIRSHICK, M. A., MOSTELLER, F., and SAVAGE, L. J. (1946). Unbiased estimation for certain binomial sampling problems with applications. *Ann. Math. Statist.* **17** 13-23.

[4] GRAYBILL, F. A. (1969). *Introduction to Matrices with Applications in Statistics.* Duxbury Press, Mass.

[5] TWEEDIE, M. C. K. (1952). The estimation of parameters from sequentially sampled data on a discrete distribution. *J. Roy. Statist. Soc. Ser. B* 14 238-245.

DEPARTMENT OF INDUSTRIAL ENGINEERING
KOREA ADVANCED INSTITUTE OF SCIENCE
P.O. BOX 150 CHONGYANGNI
SEOUL, KOREA