MULTINOMIAL SUBSET SELECTION USING INVERSE SAMPLING AND ITS EFFICIENCY WITH RESPECT TO FIXED SAMPLING

Mausumi Bose and Subir Kumar Bhandari

Indian Statistical Institute, 203 B.T. Road,
Calcutta 700 035, India

ABSTRACT

The multinomial selection problem is considered in its general form where the objective is to select a subset of s cells which contain the t 'best' cells, $s \geq t$. The inverse-sampling procedure is studied for this problem and the LFC is derived under the difference zone. An expression for the relative efficiency of this procedure with respect to the widely used fixed-sample-size selection procedure is obtained and theoretical bounds are derived for this efficiency. It is found that the inverse-sampling procedure performs uniformly better than the usual fixed-sampling procedure in the case $s = t$ and is often more efficient for $s > t$. When the selection goal is to select any c of the t best cells, using a subset of s cells, expressions for efficiency may be similarly obtained.

Key Words: Asymptotic efficiency; Difference zone; Fixed-sample-size procedure; Inverse-sampling procedure; Least favourable configuration; Probability of correct selection

1. INTRODUCTION

The problem of selecting the $t(t \geq 1)$ most probable events from a given mutinomial distribution has drawn the attention of a number of authors and many articles have been published in this area. In this context, two selection procedures, namely the inverse sampling procedure (procedure 1) and the fixed-sample-size procedure (procedure 2) have been studied in the literature.

Let the cell probabilities of a multinomial distribution be written as the vector $p = (p_1, p_2, \ldots, p_k)$, $p_1 \geq p_2 \geq, \ldots, \geq p_k$, where $1 \leq t < k$, $k \geq 3$, $\sum_{i=1}^{k} p_i = 1$. Samples are either drawn sequentially (procedure 1) or one at a time (procedure 2). The problem is to select s cells which will contain the t cells corresponding to p_1, p_2, $\ldots$ p_t. In various areas like marketing research and social survey, these sampling rules can be used to determine the most popular brands of a given product, the most favoured opinions on a political issue, etc. A procedure for such a selection is useful if it reduces the time (and cost) of the experiment. So, a comparison between the two procedures becomes important.

In this paper, we consider this selection problem in a general form where the objective is to select a subset of s cells which contain the t 'best' cells, that is, the cells corresponding to the t highest cell probabilities, $1 \leq t < k$, $k \geq 3$, $s \geq t \geq 1$. This problem was first studied by Bechhofer, Elmaghraby and Morse (1959) for the particular case where $s = t = 1$ and they proposed procedure 2 as the selection procedure. Since then, the study for procedure 2 has been continued by several researchers for different values of s and t and the general form of the problem where $s \geq t \geq 1$, was studied by Chen (1986) and by Bose and Bhandari (1999).

Procedure 1 was proposed by Cacoullos and Sobel (1966) as an alternative procedure for this selection problem. They studied the particular case of the problem where $s = t = 1$. Procedure 1 was studied by Chen and Sobel (1984a) for the case $s = t$. The problem of selecting a subset of size s which contains at least c of the t best cells, where c, s, t are such that $max(1, s + t + 1 - k) \leq c \leq min(s, t)$, was considered by Chen and Sobel (1984b) using procedure 1.

In this paper our objective is to compare procedures 1 and 2. We first consider the case $s \geq t \geq 1$ in detail. For this case, to assess the relative performances of the two procedures we calculate the relative efficiency of procedure 1 with respect to procedure 2 and bounds are derived for this efficiency. We show that procedure 1 is uniformly better than procedure 2 for $s = t$, and is often better for $s > t$. Thus, procedure 1 is a more useful selection procedure for such experiments. We also tabulate some values of this efficiency for different configurations to illustrate this. Then, for general

s and t where the selection goal is as in Chen and Sobel (1984b), we also derive the efficiency expressions for comparing the two procedures.

In the following we describe the two procedures.

Procedure 1: Inverse sampling procedure: Continue sampling one at a time until exactly s many cells reach a frequency of at least r each ($r \geq 1$), where r is a predetermined integer. As soon as this occurs, stop sampling and select these s cells. Under this procedure, let N be the expected sample size or the average sample number (ASN) for the configuration p. Clearly, this N would depend on r.

Procedure 2: Fixed-sample-size procedure: Draw a sample of fixed size n. Select the s cells with the highest frequencies in the sample, with ties broken by randomization. For this procedure, $ASN = n$ for all configurations p.

A usual requirement for selection is that the probability of correct selection (PCS) must not be lower than a prespecified level p^* if the true configuration lies in some preference zone. Two preference zones have been used extensively in the literature. One is the difference zone which is defined by:

$$D(t,k,b) = \left\{ p\colon p_t \geq p_{t+1} + b,\ b \,\text{is a constant in theinterval} \left(0, \frac{1}{t}\right) \right\}, \quad (1.1)$$

and the other is the ratio zone defined by:

$$R(t,\ k,\ b) = \{p\colon p_t \geq \theta p_{t+1},\ \theta \text{ is a constant, } \theta > 1\} \quad (1.2)$$

For the case $s = t = 1$, Cacoullos and Sobel (1966) studied the efficiency of procedure 1 with respect to procedure 2 at the slippage configuration, which is the LFC of procedure 1 in $R(t,\ k,\ \theta)$. Their efficiency measure is based on the ratio of the ASN's of the two procedures at the slippage configuration as $\theta \to 1$. In Chen and Sobel (1984b), the LFC of procedure 1 was derived under the preference zone $R(t,\ k,\ \theta)$. In this paper, when considering the LFC of procedure 1, we restrict our study to the zone $D(t,\ k,\ b)$. This is because our objective is to compare procedure 1 with procedure 2 at all p and at their LFC in particular. Alam and Thompson (1972) have shown that for the case $s \geq t \geq 2$, $R(t,\ k,\ \theta)$ is not suitable as a preference zone for computation of LFC for procedure 2 and $D(t,\ k,\ b)$ should be used in such cases. So, for a meaningful comparison at the LFC, we need the LFC for both procedures in $D(t,\ k,\ b)$.

For the case $s \geq t$, an expansion for the probability of incorrect selection of procedure 2 was obtained in Bose and Bhandari (1999). For comparing the two procedures, we first need to derive the expansion for probability of incorrect selection for procedure 1.

In Section 2, procedure 1 is studied for the case $s \geq t$. The expansion for the probability of incorrect selection of procedure 1 is obtained. Next, this expansion is used to derive the limiting form of the LFC over $D(t, k, b)$. The proofs of the results in this section use properties of concave functions and the rich-to-poor transfer technique.

In Section 3, we obtain an expression for the efficiency of procedure 1 with respect to procedure 2 and derive bounds for this efficiency. This expression is valid for any general configuration p. In particular, the efficiency is also studied at the common LFC in $D(t, k, b)$ for both the procedures. We show that for $s = t$, procedure 1 is always better. For general $s > t$, the relative performance of procedure 1 varies with p and procedure 1 is often more efficient.

In Section 4, for the selection goal of Chen and Sobel (1984b), we derive the efficiency of procedure 1 with respect to procedure 2 as a consequence of the results derived in Section 3.

2. A STUDY OF PROCEDURE 1 FOR THE CASE $s \geq t$

Let the multinomial probability vector p be written as $p = (p_1, p_2, \ldots, p_k)$, where $p_1 \geq p_2 \geq \cdots \geq p_k$. Then,

$$N = \frac{r}{p_s} + o(r) \text{ and hence } \left(\frac{N}{r}\right) \to \frac{1}{p_s} \text{ as } r \to \infty, \tag{2.1}$$

where N is the ASN for procedure 1.

Let $PCS_1(p)$ be the probability of correct selection at p, when we want to select $s(\geq t)$ cells which contain the t best cells, using procedure 1. In the following two theorems in this section, we obtain an expansion for the PCS and the expression for the limiting form of the LFC for this procedure over $D(t, k, b)$.

Theorem 2.1. Under procedure 1, as $r \to \infty$,

$$\log(1 - PCS_1(p)) = Np_s(s - t + 2)\log\frac{G}{A} + o(N) \tag{2.2}$$

where G and A respectively denote the geometric mean and arithmetic mean of the $s - t + 2$ terms $p_t, p_{t+1}, p_{t+2}, \ldots, p_{s+1}$.

Proof: $1 - PCS_1(p)$ may be expressed in terms of its dominating term as follows:

$$1 - PCS_1(\ p) \simeq$$
$$c \sum_{W_x} \left[\frac{\left\{ \sum_{\substack{i=1 \\ i \neq s}}^{k} x_i + (r-1) \right\}!}{(r-1)! \prod_{\substack{i=1 \\ i \neq s}}^{k} x_i!} \left(\prod_{i=1}^{t-1} p_i^{x_i} \right) p_t^{x_{s+1}} \left(\prod_{i=t}^{s-1} p_{i+1}^{x_i} \right) p_{s+1}^{r} \left(\prod_{i=s+2}^{k} p_i^{x_i} \right) \right] \tag{2.3}$$

where $W_x = \{x = (x_1, x_2, \ldots, x_n) \colon x_1 \geq x_2 \geq \cdots \geq x_{s-1} \geq r \geq \quad x_{s+1} \geq \quad x_i$' $\cdots \geq \ x_k,\ x_1'$s are non-negative integers} and c is a constant.

Using Stirling's approximation, (2.3) simplifies to:

$$1 - PCS_1(\ p) \simeq d \sum_{W_x} \left[\frac{\left\{ \sum_{\substack{i=1 \\ i \neq s}}^{k} x_i + (r-1) \right\} \left(\sum_{\substack{i=1 \\ i \neq s}}^{k} x_i + r - \frac{1}{2} \right)}{(r-1)^{r-\frac{1}{2}} \prod_{\substack{i=1 \\ i \neq s}}^{k} x_i^{x_i+\frac{1}{2}}} \right.$$
$$\left. \times \left(\prod_{i=1}^{t-1} p_i^{x_i} \right) p_t^{x_{s+1}} \left(\prod_{i=t}^{s-1} p_{i+1}^{x_i} \right) p_{s+1}^{r} \left(\prod_{i=s+2}^{k} p_i^{x_i} \right) \right]$$

where d is a constant. Then, after some simplification, we have

$$1 - PCS_1(\ p) \simeq t(r). \sum_{\Omega_q} \left[\frac{\{f(q)\}^r}{g(q)} \right], \tag{2.4}$$

where $\Omega_q = \{q = (q_1, \ldots, q_k) : q = (x/r), x \in W_x\}$,

$$f(q) = \frac{1}{\prod_{\substack{i=1 \\ i \neq s}}^{k} q_i^{q_i}} \left[\left(\sum_{\substack{i=1 \\ i \neq s}}^{k} q_i + 1 \right) \left(\sum_{\substack{i=1 \\ i \neq s}}^{k} q_i + 1 \right) \right.$$
$$\left. \times \left(\prod_{i=1}^{t-1} p_i^{q_i} \right) p_t^{q_{s+1}} \left(\prod_{i=t}^{s-1} p_{i+1}^{q_i} \right) p_{s+1} \left(\prod_{i=s+2}^{k} p_i^{q_i} \right) \right] \tag{2.5}$$

$$g(q) = \left(\sum_{\substack{i=1 \\ i \neq s}}^{k} q_i^{(1/2)} \right) \left(\sum_{\substack{i=1 \\ i \neq s}}^{k} q_i + 1 \right)^{\frac{1}{2}} \tag{2.6}$$

and

$$t(r) = \frac{d}{r^{((k-1)/2)} (1 - (1/r))^{r-(1/2)}} \tag{2.7}$$

Now consider

$$\Gamma_q = \{q = (q_1, q_2, \ldots, q_k) : q_1 \geq q_2 \geq \cdots \geq q_{s-1} \geq q_s \\ = 1 \geq q_{s+1} \geq \cdots \geq q_k\}. \tag{2.8}$$

We use the following two lemmas, the proofs of which are given in the Appendix.

Lemma 2.1. $\max_{\Gamma_q} \log f(q) = (s - t + 2)\log(G/A)$, where A and G are respectively the arithmetic mean and geometric mean of $p_t, p_{t+1}, \ldots, p_s, p_{s+1}$.

Lemma 2.2. $(1 - PCS_1(p))^{(1/r)} \to \max_{\Gamma_q} f(q)$, where $f(q)$ and Γ_q are as in 2.5 and 2.8.

From Lemmas 2.1 and 2.2 it follows that $(1/r)\log(1 - PCS_1(p)) \to (s - t + 2)\log(G/A)$. Hence, from (2.1), Theorem 2.1 follows.

In the following theorem, the expansion of Theorem 2.1 is used to obtain the limiting form of the LFC using the rich-to-poor transfer technique. The proof also uses a result of Bose and Bhandari (1999) which we state below, for the sake of completeness, as Lemma 2.3.

Lemma 2.3. In a group of m elements, $e_1, e_2, \ldots, e_m$, not all equal, if e_i is increased to $e_i + h$, for some h, for all $i = 1, 2, \ldots, m$, then the geometric mean of the m elements increases by an amount $\geq h$, for small h.

Theorem 2.2. For the case $s \geq t$, the limiting form of the LFC for procedure 1 over $D(t, k, b)$ is given by: $\{p = (p_1, p_2, \ldots : p_k) : p_1 = p_2 = \cdots = p_{t-1} = p_t = p_{t+1} + b > p_{t+1} = p_{t+2} = \cdots = p_s = p_{s+1} > p_{s+2} = \cdots = p_k = 0\}$.

Proof: The theorem may be proved using the following transfer steps used to narrow down the search for the limiting form of the LFC in $D(t, k, b)$. By Theorem 2.1, this limting form will be given by the point which maximizes $p_s log(G/A)$.

(1) Transfer from $p_{s+2}, \ldots, p_k$ to p_1 until $p_{s+2} = \cdots = p_k = 0$. This transfer leaves p_s, G and A unaffected. So now the LFC over $D(t, k, b)$ lies in the subclass

$$\wp_1 = \{p : p_1 \geq \cdots \geq p_t \geq p_{t+1} + b > p_{t+1} \geq \cdots \geq p_{s+1} \geq p_{s+2} \\ = \cdots = p_k = 0\} \subset D(t, k, b)$$

(2) Continue rich-to-poor transfer among $p_{t+1} \ldots, p_{s+1}$ until $p_{t+1} = p_{t+2} = \cdots = p_{s+1}$. This transfer leaves A unchanged and increases G. Moreover, by considering the directional derivatives of $p_s log(G/A)$ with

$p_{t+1} = p_{t+2} = \cdots = p_s > p_{s+1}$, it can be seen that this transfer increases $p_s log(G/A)$. Now transfer from p_t to $p_{t+1}, \ldots, p_{s+1}$ until $p_t = p_{t+1} + b, p_{t+1} = \cdots = p_{s+1}$. This transfer again leaves A unchanged, increases G and increases p_s. Hence, by (2.2) the LFC over $D(t,\ k,\ b)$ lies in the subclass

$$\begin{aligned}\wp_2 = \{p : p_1 \geq \cdots \geq p_{t-1} \geq p_t \geq p_{t+1} + b > p_{t+1} \\ = \cdots = p_{s+1} > p_{s+2} = \cdots = p_k = 0\} \subset \wp_1\end{aligned}$$

(3) Transfer from $p_2, p_3, \ldots, p_{t-1}$ until $p_1 > p_2 = \cdots = p_{t-1} = p_t$. This transfer leaves p_s, A and G unchanged. So, the LFC lies in

$$\begin{aligned}\wp_3 = \{p : p_1 \geq p_2 = \cdots = p_t = p_{t+1} + b > p_{t+1} = \cdots = p_{s+1} > p_{s+2} \\ = \cdots = p_k = 0\} \subset \wp_2\end{aligned}$$

(4) Finally, transfer from p_1 to each of $p_2, p_3, \ldots, p_{t-1}, p_t$, $p_{t+1}, \ldots, p_{s+1}$ by equal amounts until $p_1 = p_2$. This increases p_s and by Lemma 2.3, increases G/A. No more transfers are possible towards increasing p_s and G/A. Hence the LFC over $D(t, k, b)$ lies in

$$\begin{aligned}\wp = \{p : p_1 = \cdots = p_t = p_t = p_{t+1} + b > p_{t+1} \\ = \cdots = p_{s+1} > p_{s+2} = \cdots = p_k = 0\} \subset \wp_3 \subset D(t,\ k,\ b).\end{aligned}$$

Remark 2.1. As is known from the literature, there does not seem to be a single LFC which works for all n. So, the limiting form of the LFC becomes interesting in this case.

Remark 2.2. The LFC as obtained in Theorem 2.2 coincides with the LFC obtained for the fixed-sample-size procedure in Bose and Bhandari (1999).

3. A COMPARISON OF PROCEDURE 1 VERSUS PROCEDURE 2 FOR $s \geq t \geq 1$

Let $PCS_2(p)$ be the probability of correct selection at p, when we want to select $s(\geq t)$ cells which contain the $t \geq 1$ best cells, using procedure 2 with a sample of size n. Then by Theorem 2.1 of Bose and Bhandari (1999), as $n \to \infty$,

$$log(1 - PCS_2(p)) = n \log[1 - (s - j_0 + 2)(A_0 - G_0)] + o(n), \tag{3.1}$$

where j_0 is the largest integer among $t, t+1, \ldots, s$ such that $p_{j_0} > G_0$, and

$$G_0 = (p_t p_{j_0+1}, \ldots, p_s p_{s+1})^{(1/s-j_0+2)}, \qquad A_0 = \frac{1}{s-j_0+2}\left(p_t + \sum_{i=j_0+1}^{s+1} p_i\right).$$

The efficiency of one procedure with respect to another is defined as the ratio of the ASN required by each of the procedures to achieve the same probability of correct selection (α, say) for a given configuartion p in $D(t, k, b)$. Thus the asymptotic relative efficiency of procedure 1 with respect to procedure 2 is defined as:

$$e = \lim_{\alpha \to 1} \frac{ASN \text{ of procedure } 2}{ASN \text{ of procedure } 1}.$$

From 2.2 and 3.1, we then have

$$e = \frac{p_s(s-t+2)\log(G/A)}{\log[1-(s-j_0+2)(A_0-G_0)]} \tag{3.2}$$

If e takes a value greater than unity, we say that procedure 1 is more efficient than procedure 2 in the sense that for a given configuration p in $D(t, k, b)$, procedure 1 requires, on the average, a fewer number of observations than procedure 2, to achieve the same probability of correct selection.

Now, we study the possible value of e for general p and $s \geq t$.

Case 1. $s = t$

If $s = t$, then $s = j_0 = t$ and e as in (3.2) simplifies to

$$e = \frac{2p_t \log(2\sqrt{p_t p_{t+1}}/(p_t + p_{t+1}))}{\log[1 + (\sqrt{p_t p_{t+1}} - (p_t + p_{t+1}/2))]}.$$

Now, it can be shown using routine algebra that $e > 1$ and so procedure 1 is always more efficient than procedure 2. To illustrate the actual values of this efficiency, the value of e has been computed in Tables 1 and 2 for some values of s and t, $s = t$.

Case 2. $s > t$. For this case, the value of e depends on the relative values of $p_1, p_2, \ldots, p_k$ and s, t. The following theorem gives bounds for e.

Theorem 3.1. The efficiency e of procedure 1 with respect to procedure 2 satisfies the following inequality:

$$\frac{(s-t+2)(A-G)}{(s-j_0+2)(A_0-G_0)}\,\frac{p_s}{A} < e < \frac{(s-t+2)(A-G)}{(s-j_0+2)(A_0-G_0)}\,\frac{p_s}{G},$$

where e, A, G, A_0, G_0 are as in (3.2), (2.2) and (3.1) respectively.

Remark 3.1. When $j_0 = t$, the bounds of Theorem 3.1 simplify to

$$\frac{p_s}{A} < e < \frac{p_s}{G}.$$

Remark 3.2. At LFC, $j_0 = t$ and so the bounds of Remark 3.1 apply.

Case (i). At LFC, for $s = t$, $(p_s/A) = (p_t/A) > 1$. Hence, $e > 1$ and so, procedure 1 is uniformly more efficient than procedure 2.

Case (ii). At LFC, for $s > t$, as $p_s < G$, by Remark 3.1 it follows that $e < 1$. However, for b small, (p_s/A) is close to unity and so, e is also close to unity. Hence procedure 1 is almost as efficient as procedure 2 and this efficiency decreases if b increases.

Table 1 gives some illustrative e values for these cases.

Remark 3.3. Outside LFC, if $j_0 = t$, then e can be more than unity. To see this, starting from any p, we keep p_s fixed and transfer to $p_1, p_2, \ldots, p_{t-1}$ from $p_t, p_{t+1}, \ldots, p_{s-1}, p_{s+1}$.

If j_0 does not change by this transfer, A will decrease and (p_s/A) will increase. This will increase e and e can become more than unity. If j_0 increases by the above transfer and becomes greater than t, then, $((s-t+2)(A-G)/(s-j_0+2)(A_0-G_0)) > 1$ and so, again e increases.

Table 2 gives the values of e for some configurations outside the LFC.

Remark 3.4. The above discussion shows that for the multinomial selection problem, procedure 1 is always more efficient than procedure 2 when $s = t$ and even when $s > t$, procedure 1 often requires substantially fewer observations than procedure 2. So, procedure 1 is extremely useful as it is simple and at the same time it reduces the cost of the experiment.

Now we prove Theorem 3.1. We first state and prove a lemma which is required to prove the Theorem.

Lemma 3.1. $((s-t+2)(A-G)/(s-j_0+2)(A_0-G_0)) \geq 1$.

Proof: Let the arithmetic mean (A.M.) and the geometric mean (G.M.) of $p_{t+1}, p_{t+2} \ldots, p_{j_0}$ be denoted by A_1 and G_1 respectively. Let $w_0 = s - j_0 + 2$ and $w_1 = j_0 - t$.

Noting that G as in (2.2) is a weighted G.M. of G_0 and G_1, with weights w_0 and w_1 respectively, it follows that

$$\begin{aligned} A - G &= (A - \text{weighted G.M. of } G_0, G_1) \\ &\geq (A - \text{weighted G.M. of } G_0,\ A_1) \\ &\geq (A - \text{weighted A.M. of } G_0,\ A_1). \end{aligned}$$

Again, noting that A is the similarly weighted A.M. of A_0 and A_1, it now follows that

$$\begin{aligned} (s-t+2)(A-G) \geq (s-t+2)(&\text{weighted A.M. of } A_0,\ A_1 \\ &- \text{weighted A.M of } G_0,\ A_1) = (s-j_0+2)(A_0-G_0). \end{aligned}$$

Hence Lemma.

Table 1. Efficiency at LFC

t	s	b	e	t	s	b	e
1	1	0.01	1.0100	2	3	0.005	0.9927
		0.05	1.0500			0.01	0.9866
		0.09	1.0900			0.05	0.9316
		0.30	1.3000			0.09	0.8740
		0.60	1.6000			0.10	0.8596
		0.90	1.9000			0.20	0.7029
2	2	0.01	1.0151			0.30	0.5246
		0.05	1.0774	2	4	0.005	0.9934
		0.09	1.1435			0.01	0.9874
		0.10	1.1606			0.05	0.9357
		0.20	1.3493			0.09	0.8815
		0.30	1.5863			0.10	0.8676
		0.40	1.9372			0.20	0.7178
3	3	0.01	1.0202			0.30	0.5448
		0.05	1.1069	2	5	0.005	0.9935
		0.2	1.5669			0.01	0.9879
1	2	0.005	0.9950			0.05	0.9382
		0.01	0.9900			0.09	0.8858
		0.05	0.9500			0.10	0.8724
		0.09	0.9100			0.20	0.7272
		0.10	0.9000			0.30	0.5578
		0.30	0.7000				

Proof of Theorem 3.1: From (3.2),

$$\begin{aligned} e &= \frac{p_s(s-t+2)log[1-(A-G/A)]}{log[1-(s-j_0+2)(A_0-G_0)]} \\ &= p_s(s-t+2)\frac{\sum_{k=1}^{\infty}((A-G)^k/A^k k)}{\sum_{k=1}^{\infty}(s-j_0+2)^k((A_0-G_0)^k/k)} \\ &= p_s(s-t+2)\frac{M}{N}, \; say \qquad \cdots(3.3) \end{aligned}$$

The ratio of the kth term in M and the kth term in N is:

$$\frac{[(s-t+2)(A-G)]^k}{[(s-j_0+2)(A_0-G_0)]^k}\;\frac{1}{[(s-t+2)A]^k} \qquad \cdots(3.4)$$

Since $(s-t+2)A$ is less than unity, $(1/[(s-t+2)A]^k)$ is increasing in k. This together with Lemma 3.1 implies that the expression in (3.4) is also increasing in k. Hence, (M/N) is greater than the ratio of the first terms of M and N. So, from (3.4) it follows that

$$e > \frac{(s-t+2)(A-G)}{(s-j_0+2)(A_0-G_0)}\;\frac{p_s}{A}. \qquad \cdots(3.5)$$

Again, since the ratio of kth term in M and the kth term in N is increasing in k, if we multiply the kth terms of both M and N by k, for $k = 1, 2, \ldots, \infty$, then the ratio (M/N) increases. So,

$$e < p_s(s-t+2)\frac{\sum_{k=1}^{\infty}((A-G)^k/A^k)}{\sum_{k=1}^{\infty}(s-j_0+2)^k(A_0-G_0)^k}.$$

On simplification, it follows that

$$\begin{aligned} e &< p_s(s-t+2)\frac{((A-G)/A)/(1-(1-(G/A)))}{(s-j_0+2)(A_0-G_0)/(1-(s-j_0+2)(A_0-G_0))} \\ &= \frac{(s-t+2)(A-G)}{(s-j_0+2)(A_0-G_0)}\;\frac{p_s}{G}[1-(s-j_0+2)(A_0-G_0)]. \end{aligned}$$

Hence,

$$e < \frac{(s-t+2)(A-G)}{(s-j_0+2)(A_0-G_0)}\;\frac{p_s}{G} \qquad \cdots(3.6)$$

From (3.5) and (3.6), the theorem follows.

Table 2. Efficiency at Some p Vectors Not Equal to the LFC

k	p	t	s	b	e
4	(0.500 0.490 0.005 0.005)	1	2	0.01	3.2031
	(0.490 0.480 0.020 0.010)	1	2	0.01	2.6710
	(0.420 0.410 0.090 0.080)	1	2	0.01	1.9717
	(0.420 0.410 0.090 0.080)	1	1	0.01	1.0120
	(0.500 0.480 0.010 0.010)	1	2	0.02	2.8253
	(0.480 0.460 0.030 0.030)	1	2	0.02	2.3822
	(0.450 0.430 0.060 0.060)	1	2	0.02	2.0938
	(0.450 0.430 0.060 0.0600)	1	1	0.02	1.0227
	(0.800 0.100 0.090 0.010)	2	3	0.01	2.2396
	(0.900 0.050 0.040 0.010)	2	3	0.01	1.6008
	(0.750 0.100 0.090 0.060)	2	3	0.01	1.2231
	(0.840 0.085 0.065 0.010)	2	3	0.02	1.7564
	(0.850 0.080 0.060 0.010)	2	3	0.02	1.6889
	(0.870 0.070 0.050 0.010)	2	3	0.02	1.5363
	(0.700 0.150 0.120 0.030)	2	3	0.03	1.5759
	(0.800 0.100 0.070 0.030)	2	3	0.03	1.1660
	(0.900 0.060 0.030 0.010)	2	3	0.03	1.0194
	(0.750 0.140 0.090 0.020)	2	3	0.05	1.3435
	(0.800 0.100 0.095 0.0050	2	3	0.005	2.7972
5	(0.800 0.050 0.050 0.050 0.050)	1	1	0.750	2.0173
	(0.500 0.470 0.010 0.010 0.010)	1	1	0.03	1.0309
	(0.500 0.470 0.010 0.010 0.010)	1	2	0.03	2.7568
	(0.500 0.495 0.002 0.002 0.001)	1	1	0.005	1.0050
	(0.500 0.495 0.002 0.002 0.001)	1	2	0.005	3.6903
	(0.400 0.395 0.002 0.002 0.001)	1	2	0.005	1.6259
	(0.400 0.395 0.200 0.004 0.001)	2	3	0.195	1.6764
	(0.400 0.395 0.200 0.004 0.001)	2	2	0.195	1.3432
6	(0.400 0.395 0.200 0.003 0.001 0.001)	2	2	0.195	1.3432
	(0.400 0.395 0.200 0.003 0.001 0.001)	3	3	0.197	3.4000
	(0.400 0.395 0.200 0.003 0.001 0.001)	1	3	0.005	1.8277
	(0.400 0.300 0.295 0.003 0.001 0.001)	2	3	0.005	3.5981
	(0.300 0.300 0.295 0.050 0.050 0.005)	2	2	0.005	1.0084
	(0.300 0.300 0.295 0.050 0.050 0.005)	5	5	0.045	2.3400
7	(0.200 0.200 0.200 0.195 0.195 0.005 0.005)	3	3	0.005	1.0137
	(0.400 0.395 0.200 0.002 0.001 0.001 0.001)	3	3	0.198	3.6651
	(0.200 0.200 0.200 0.195 0.195 0.005 0.005)	3	5	0.005	3.2596

4. A COMPARISON OF PROCEDURE 1 AND PROCEDURE 2 FOR THE SELECTION GOAL OF CHEN AND SOBEL (1984B)

A selection problem with general s and t was considered in Chen and Sobel (1984b). Their goal was to select a subset of size s which contains at least c of the t best cells, where c, s, t are such that $max(1, s+t+1-k) \leq c \leq min(s, t)$. For this problem, Chen and Sobel (1984b) used procedure 1 and the s cells so selected were to contain at least c of the t best cells.

With this formulation of correct selection, we could again compare procedure 1 with procedure 2 in a way similar to what was done in Section 3. For this comparison, the expansions for the probabilities of incorrect selection with the two procedures are required under this formulation. These expansions are given below:

Let $PCS_3(p)$ and $PCS_4(p)$ be the probabilities of correct selection at p under the formulation of Chen and Sobel (1984b) for procedure 1 and procedure 2 respectively.

Corollary 4.1. Under procedure 1, as $r \to \infty$,

$$\log(1 - PCS_3(p)) = Np_s(s-c+2)\log\frac{G}{A} + o(N) \tag{4.1}$$

where G and A respectively denote the geometric mean and arithmetic mean of the $s-c+2$ terms $p_c, p_{c+1}, p_{c+2}, \cdots p_{s+1}$.

Proof: Note that as in (2.3), for this formulation, $1 - PCS_3(p)$ may be expressed in terms of its dominating term by

$$1 - PCS_3(p) \simeq$$
$$d\sum_{W_x}\left[\frac{\left\{\sum_{\substack{i=1\\i\neq s}}^{k} x_i + (r-1)\right\}!}{(r-1)!\prod_{\substack{i=1\\i\neq s}}^{k} x_i!}\left(\prod_{i=1}^{c-1} p_i^{x_i}\right)p_c^{x_{s+1}}\left(\prod_{i=c}^{s-1} p_{i+1}^{x_i}\right)p_{s+1}^{r}\left(\prod_{i=s+2}^{k} p_i^{x_i}\right)\right]$$

where $W_x = \{x = (x_1, x_2, \ldots, x_n) : x_1 \geq x_2 \geq \cdots \geq x_{s-1} \geq r \geq x_{s+1} \geq \cdots \geq x_k$, x_i's are non-negative integers$\}$ and d is a constant.

Then, Corollary 4.1 follows along the line of the proof of Theorem 2.1.

Similarly, following the proof of Theorem 2.1 of Bose and Bhandari (1999), the following may be derived.

Corollary 4.2. $PCS_4(p)$ admits the expansion

$$log(1 - PCS_4(p)) = n[1 - (s - j_0 + 2)(A_0 - G_0)] + o(n), \quad \text{as } n \to \infty,$$

where j_0 is the largest integer among $c, c+1, \ldots, s$ such that $p_{j_0} > G_0$, and

$$A_0 = \left(p_c + \sum_{i=j_0+1}^{s+1} p_i\right)\frac{1}{s - j_0 + 2}, \qquad G_0 = (p_c \cdot p_{j_0+1}, \ldots, p_s p_{s+1})^{\frac{1}{s-j_0+2}}. \tag{4.2}$$

For the selection goal of Chen and Sobel (1984b), using Corollaries 4.1 and 4.2, it may be shown following arguments as in Section 3 that the expression for efficiency of procedure 1 with respect to procedure 2 will be given by

$$e = \frac{p_s(s - c + 2)\log(G/A)}{\log[1 - (s - j_0 + 2)(A_0 - G_0)]} \tag{4.3}$$

where A, A_0, G, G_0 are as defined in Corollaries 4.1 and 4.2.

All other results in Section 3 can be extended for the formulation of Chen and Sobel (1984b) easily.

APPENDIX

Proof of Lemma 2.1: From (2.5)

$$\begin{aligned}\log f(q) = &\left(\sum_{\substack{i=1\\ i\neq s}}^{k} q_i + 1\right)\log\left(\sum_{\substack{i=1\\ i\neq s}}^{k} q_i + 1\right)\\ &- \sum_{\substack{i=1\\ i\neq s}}^{k} q_i \log q_i + \sum_{i=1}^{t-1} q_i \log p_i + q_{s+1}\log p_t\\ &+ \sum_{i=t}^{s-1} q_i \log p_{i+1} + \log p_{s+1} + \sum_{i=s+2}^{k} q_i \log p_i \qquad \cdots (\text{A.1})\end{aligned}$$

From (A.1), it is clear that $\log f(q)$ is concave in each q_i, $i = 1, \ldots, k$, $i \neq s$, separately and so the solutions q_i^* of

$$\frac{\partial \log f(q)}{\partial q_i} = 0 \qquad \cdots (\text{A.2})$$

for $i = 1, \ldots, k,\ i \neq s$, will give the maximizer point $q^* = (q_1^*, q_2^*, \ldots, q_k^*)$ of $\log f(q)$. From (A.1) and (A.2), we obtain

$$\left.\begin{aligned} q_i^* &= p_{i+1}\left(1+\sum_{\substack{i=1\\ i\neq s}}^{k} q_i^*\right) \quad \text{for } i = t, t+1, \ldots, s-1 \\ &= p_t\left(1+\sum_{\substack{i=1\\ i\neq s}}^{k} q_i^*\right) \quad \text{for } i = s+1 \\ &= p_i\left(1+\sum_{\substack{i=1\\ i\neq s}}^{k} q_i^*\right) \quad \text{for } i = 1, 2, \ldots, t-1 \\ &\quad \text{and } i = s+2, s+3, \ldots, k \end{aligned}\right\} \qquad \cdots \text{(A.3)}$$

Now we have to check if $q^* \in \Gamma_q$, where Γ_q is as in (2.8). Since $p \in D(t, k, b)$ from (1.1) it follows that $q_{s+1}^* > q_i^*$ for $i = t, t+1, \ldots, s-1$ and so by (2.8), q^* as given by (A.3) cannot belong to Γ_q which requires $q_i \geq 1 \geq q_{s+1}$ for $i = t, t+1, \ldots, s-1$. So, remembering that $\log f(q)$ is concave in q_{s+1} and in each of q_i, $i = t, \ldots, s-1$, it now follows that the maximizer point q^{**} of $\log f(q)$ over Γ_q must have $q_t^{**} = q_{t+1}^{**} = \cdots = q_{s-1}^{**} = q_s^{**} = 1 = q_{s+1}^{**}$.

Now the remaining q_i^* values may be obtained by solving (A.2) for $i = 1, 2, \ldots, t-1$ and $i = s+2, \ldots, k$. Hence, after some simplification using $\sum_{i=1}^{k} p_i = 1$, we have

$$\left.\begin{aligned} q_i^{**} &= \frac{1}{A} p_i \quad \text{for } i = 1, \ldots, t-1 \\ &= 1 \quad \text{for } i = t, t+1, \ldots, s-1, s, s+1 \\ &= \frac{1}{A} p_i \quad \text{for } i = s+2, \ldots, k \end{aligned}\right\} \qquad \cdots \text{(A.4)}$$

where A is the arithmetic mean of the terms $p_t, p_{t+1}, \ldots, p_s, p_{s+1}$. Clearly q^{**} is in Γ_q and so from (A.1) and (A.4), it follows on simplification that

$$\max_{\Gamma_q} \log f(q) = \log f(q^{**}) = (s - t + 2) \log \frac{G}{A} \qquad \cdots \text{(A.5)}$$

where G is the geometric mean of $p_t, p_{t+1}, \ldots, p_s, p_{s+1}$. Hence Lemma 2.1.

Proof of Lemma 2.2: Consider an infinite sequence of concentric spheres in $\Re^k$ centered at the origin. Let the sphere with radius n be denoted by C_n.

For fixed η, η small, let $\epsilon > 0$ be the area of the set $Q = \{q : \max_{\Gamma_q} f(q) > f(q) - \eta\}$. Also, for some $\delta > 0$, $g(q^{**}) + \delta \geq \max_Q g(q)$, where q^{**} is as in (A.4). Lemma 2.1 shows that q^{**} is unique point and so for large r, there will be a n_0 such that q^{**} will be contained in C_n for all $n \geq n_0$.

Hence, for large r and large n,

$$\frac{\epsilon\{\max\limits_{\Gamma_q} f(q) - \eta\}^r}{g(q^{**}) + \delta} \leq \sum_{\Omega_q \cap C_n} \frac{\{f(q)\}^r}{g(q)} \qquad \cdots(\mathrm{A}.6)$$

Note that $g(q) = 0$ if at least one $q_i = 0$, $i = 1, \ldots, k$.

There exists an open set B around the zeros of g, for which the Lebesgue measure of B is less than ϵ^*, for some small $\epsilon^* > 0$. For fixed n, since C_n is compact, g is bounded below in $\Omega_q \cap C_n - B$ and so

$$\sum_{\Omega_q \cap C_n} \frac{\{f(q)\}^r}{g(q)} \leq C\Sigma\{f(q^{**})\}^r, \qquad \cdots(\mathrm{A}.7)$$

for some constant C, where the summation is taken over all the points in $\Omega_q \cap C_n - B$. Since C_n is of radius n, it can contain at most $(2nr)^k$ points of Ω_q and so

$$C\sum\{f(q^{**})\}^r \leq C(2nr)^k\{f(q^{**})\}^r \qquad \cdots(\mathrm{A}.8)$$

Hence, from (A.6), (A.7) and (A.8) for each $n \geq n_0$,

$$\frac{\epsilon\{f(q^{**}) - \eta\}^r}{g(q^{**}) + \delta} \leq \sum_{\Omega_q \cap C_n} \frac{\{f(q)\}^r}{g(q)} \leq C(2nr)^k\{f(q^{**})\}^r. \qquad \cdots(\mathrm{A}.9)$$

Now, the R.H.S. of (2.4) is bounded above by 1 and $E(q) \rightarrow (p/p_s)$ for large r. Hence by Markov inequality on the coordinates of q, it follows that for all r, the R.H.S. of (2.4) is uniformly dominated. Hence for some large n^*, the proportion of the portion of the R.H.S. of (2.4) not contained in C_{n^*} can be made smaller than any given ϵ^*. So,

$$\sum_{\Omega_q} t(r)\frac{\{f(q)\}^r}{g(q)} = \sum_{\Omega_q \cap C_{n^*}} t(r)\frac{\{f(q)\}^r}{g(q)}[1 + \epsilon^*]. \qquad \cdots(\mathrm{A}.10)$$

Hence, from (A.9) and (A.10),

$$t(r)\frac{\epsilon\{f(q^{**}) - \eta\}^r}{g(q^{**}) + \delta} \leq \sum_{\Omega_q \cap C_{n^*}} t(r)\frac{\{f(q)\}^r}{g(q)} = \sum_{\Omega_q} t(r)\frac{\{f(q)\}^r}{g(q)}\frac{1}{1 + \epsilon^*}$$

$$\leq Ct(r)(2n^*r)^k\{f(q^{**})\}^r.$$

Now taking the r-th root, the lemma follows.

Proof of equation (2.1): Let $x = (x_1, \ldots, x_k)$ denote the observation vector corresponding to the k multinomial cells and let $n_{i(r)} = x_1 + \cdots + x_k$ when $x_i = r$ for the first time. Let n be the sample size for procedure 1 at the time of stopping. Let A_s be the event that a cell with cell probability p_s has observation r at the stopping point of procedure 1.

Then, by SLLN, $(n/x_s) \to (1/p_s)$ and $P(A_s) \to 1$. Now, $(x_s/r)1_{A_s} \to 1$. Hence, $(n/r) \to (1/p_s)$.

Note that $0 \leq (n/r) \leq Y_r$, where $Y_r = (1/r)(n_{1(r)} + \cdots + n_{k(r)})$ and $Y_r \to \sum_{i=1}^{k}(1/p_i)$ and $E(Y_r) = \sum_{i=1}^{k}(1/p_i)$. Hence, by Billingsley (1991), it follows that

$$E\left(\frac{n}{r}\right) \to \frac{1}{p_s}.$$

REFERENCES

Alam, K. and Thompson, J.R. (1972). On selecting the least probable multinomial even. *Ann. Math. Statist.* **43**, 1981–1990.

Bechhofer, R.E., Elmaghraby, S.A. and Morse, N. (1959). A single-sample multiple decision procedure for selecting the multinomial event which has the largest probability. *Ann. Math. Statist.* **30**, 102–119.

Billingsley, P. (1991). *Probability and Measure. John Wiley and Sons, Inc, Singapore* pp. 222, 16.6a.

Bose, M. and Bhandari, S.K. (1999). Selecting the *t*-best cells of a multinomial distribution. *SankhyaĀ* **61**, 139–147.

Cacoullos, T. and Sobel, M. (1966). An inverse sampling procedure for selecting the most probable event in a multinomial distribution. *Proceedings of the 1st International Symposium on Multivariate Analysis ed. by P.R. Krishnaish, Academic Press*, New York, 423–455.

Chen, P. (1986). On the least favourable configuration in multinomial selection problems, *Commun. Statist.-Theor. Meth.* **15**(2), 367–385.

Chen, P. and Sobel, M. (1984a). Selecting the *t* best cells of a multinomial using inverse sampling inequalities, *Statistcs and Probability, IMS Lecture Notes-Monograph Series* Vol. 5, 206–210.

Chen, P. and Sobel, M. (1984b). Selecting among multinomial cells using inverse sampling – a generalized goal, *Statistics and Decisions, Supplement Issue* **1**, 285–295.

Received October 1999
Revised February 2001
Recommended by M. Ghosh