

# SEQUENTIAL ESTIMATION OF THE NUMBER OF MULTINOMIAL CELLS

Hokwon Cho and Z. Govindarajulu

Department of Mathematical Sciences  
University of Nevada, Las Vegas, Las Vegas, NV 89154  
and

Department of Statistics  
University of Kentucky, Lexington, KY 40506

## SYNOPTIC ABSTRACT

A sequential risk-efficient estimator with squared error loss is proposed for estimating the number of equally probable cells in a given multinomial distribution. It is assumed that the cost per observation is constant. Large-sample properties of the sequential estimator are investigated and a simulation study is carried out in order to examine its finite sample behavior.

Key Words and Phrases: Multinomial distribution; number of categories; classical occupancy problem; sequential estimator; risk efficiency; squared error loss.

# 1 INTRODUCTION

Suppose we have a problem of estimating the unknown number of categories (or cells) based on a sample of size  $n$  drawn from a multinomial population. For example, numismatists may be interested in estimating the total number of dies used to produce coins in ancient times. Biological scientists or ecologists may be concerned with estimating the number of species in a population of animals or plants.

For this type of problems, various methods have been proposed and developed steadily since Goodman (1949). Bunge and Fitzpatrick (1993) provide a rather comprehensive review of the existing literature with over 550 references, dealing with approaches and statistical models employed.

This class of problems can be thought via the classical occupancy problem because the basic formulation of the problem in modeling is quite similar to that of drawing a sample of size  $n$  with replacement from a population of  $k$  distinct objects. Therefore, the problem turns out to be equivalent to the problem of distributing  $n$  balls into  $k$  distinct boxes, and it is also closely related to the coupon collector's problem.

Boender and Rinnooy Kan (1983) take the Bayesian approach and develop optimal Bayesian stopping rules via the backward recursion, for a sequential sample from a multinomial distribution with an unknown number of cells assuming a uniform prior on the cell probabilities. Further, they assume a cost which is proportional to the relative error of the posterior expectation and fixed sampling cost for each trial.

Boender and Zielinski (1985) consider a cost function that is linear in the sample size and assume a uniform prior on the cell probabilities. They propose to obtain the Bayes stopping rule by the backward induction method.

Finkelstein, Tucker and Veeh (1997) established the almost sure convergence of  $K_n$  to  $k$ .

## SEQUENTIAL ESTIMATION OF MULTINOMIAL CELLS

Since the existing methods have been dealing with the fixed sample size procedures and Bayesian sequential approaches, here, we wish to tackle this problem via a non-Bayesian sequential approach.

### 1.1 FORMULATION

Let  $X_1, X_2, \dots$  be a sequence of independent observations from a multinomial population with unknown  $k$  equally probable distinct cells. With a sample  $(X_1, X_2, \dots, X_n)$  of size  $n$ , we want to estimate the true number of cells,  $k$  ( $< \infty$ ) by  $K_n$  with a loss function of the form

$$L_n = (K_n - k)^2 + cn \quad (1)$$

where  $K_n$  is the number of distinct cells in  $n$  observations,  $c$  ( $c > 0$ ) is proportional to the cost per observation. The risk

$$\begin{aligned} R_n(c) &= E(L_n) = E(K_n - k)^2 + cn \\ &= \text{var}(K_n) + [E(K_n - k)]^2 + cn \end{aligned} \quad (2)$$

where

$$E(K_n) = k \left[ 1 - \left( \frac{k-1}{k} \right)^n \right],$$

the bias, which is defined

$$b = E(K_n) - k,$$

and

$$\text{var}(K_n) = k \left( \frac{k-1}{k} \right)^n - k^2 \left( \frac{k-1}{k} \right)^{2n} + k(k-1) \left( \frac{k-2}{k} \right)^n.$$

The distribution of  $K_n$  is given by

$$P(K_n = s | k, n) = \binom{k}{s} k^{-n} \sum_{v=0}^s (-1)^v \binom{s}{v} (s-v)^n, \quad s = 1, 2, \dots, \min(n, k), \quad (3)$$

which is shown in Govindarajulu (1999, p. 44). Also the distribution of  $k - K_n$  is given in Feller (1968, p. 102). Jordan (1950, p. 178) seems to be the first to obtain the exact distribution of  $k - K_n$ .

In fact, when  $k$  and  $n$  become large such that  $n/k \rightarrow \rho$  ( $0 < \rho < \infty$ ),

$$E(K_n) = k(1 - e^{-n/k}) + O(1),$$

the bias is given by

$$b = -ke^{-n/k} + O(1),$$

and

$$\text{var}(K_n) = ke^{-n/k}(1 - e^{-n/k}) - ne^{-2n/k} + O(1).$$

(See, for instance, Weiss (1958) or Harris (1968).)

Let  $y = e^{-n/k}$ , then the risk in (2) can be written as (ignoring  $O(1)$  term)

$$R_n(c) = k(k-1)y^2 + ky - ck \log y + ky^2 \log y. \quad (4)$$

By setting  $\frac{\partial}{\partial y} [R_n(c)|k] = 0$  (since this requires  $\frac{\partial^2}{\partial y^2} [R_n(c)|k] > 0$ ), the minimum of the risk can be attained by the solution of the equation:

$$g(y) = 2(k-1)y + 1 - \frac{c}{y} + 2y \log y + y = 0. \quad (5)$$

To get the initial value of the solution, we set

$$1 - \frac{c}{y} = 0.$$

So let  $y_0 = c$  denote the initial solution. Then we obtain the next approximation by the Newton-Raphson method as

$$y_1 = c - \frac{g(c)}{g'(c)}$$

## SEQUENTIAL ESTIMATION OF MULTINOMIAL CELLS

where

$$g(c) = (2k - 1)c + 2c \log c,$$

and

$$g'(c) = 2(k - 1) + \frac{1}{c} + 2 \log c + 3.$$

Hence

$$\begin{aligned} y_1 &= c - \frac{(2k - 1)c + 2c \log c}{2k + 1 + 1/c + 2 \log c} \\ &= \frac{2c^2 + c}{(2k + 1)c + 2c \log c + 1}. \end{aligned}$$

Thus

$$e^{-n/k} = \frac{2c^2 + c}{(2k + 1)c + 2c \log c + 1}$$

or

$$n = k [-\log c - \log(2c + 1) + \log \{(2k + 1)c + 2c \log c + 1\}]. \quad (6)$$

Hence we take the optimal fixed-sample size  $n^*$  as the integer such that  $n \leq n^* \leq n + 1$ , for estimating  $k$  when everything is known.

Since  $k$  is fixed,  $c \rightarrow 0$ , expanding  $\log\{(2k + 1)c + 2c \log c + 1\}$  in (6), we obtain (assuming  $(2k + 1)c + 2c \log c < 1$ )

$$\begin{aligned} n^* &= k[-\log c - 2c + (2k + 1)c + 2c \log c + O(c^2)] \\ &\approx k\{(2k - 1)c + (2c - 1) \log c\}. \end{aligned} \quad (7)$$

Then the minimum risk associated with the optimal fixed-sample size  $n^*$  becomes

$$\begin{aligned} R_{n^*}(c) &= ky(1 - y) + (k^2 - n^*)y^2 + cn^* \\ &= kc [1 - 2(k - 1)c + k(k - 1)c^2] \\ &\quad - kc [\log c - \log \{1 + 2(k - 1)c\} + O(c^3)]. \end{aligned} \quad (8)$$

Since  $k$  is unknown, there is no fixed-sample size procedure that will attain the risk (4). So, we resort to the following adaptive sequential procedure: Stop at  $N \equiv N_c$  where

$$N = \inf \{n \geq n_0 : n \geq K_n [-\log c - \log(2c+1) + \log \{(2K_n + 1)c + 2c \log c + 1\}]\} \quad (9a)$$

$$\simeq \inf \{n \geq n_0 : n \geq K_n \{(2K_n - 1)c + (2c - 1) \log c\}\} \quad (9b)$$

where  $n_0$  ( $n_0 \geq 2$ ) denotes the initial sample size,  $K_n$  is the number of distinct cells in  $n$  observations, and we note that  $K_n \leq k$ .

Then, the risk function associated with the stopping time  $N$  is given by

$$R_N(c) = E \{(K_N - k)^2 + cN\}. \quad (10)$$

## **2 ASYMPTOTIC BEHAVIOR OF THE PROCEDURE**

In this section, we investigate the asymptotic behavior of the sequential procedure  $(N_c, K_N)$  and study some properties of the stopping time  $N$ .

### **2.1 FINITE SURE TERMINATION**

Here, we would like to show the fundamental property that the proposed stopping rule terminates finitely almost surely.

**Theorem 1.** *Let  $N$  denote the stopping time in the sequential procedure. Then  $P(N = \infty) = 0$ .*

## SEQUENTIAL ESTIMATION OF MULTINOMIAL CELLS

**Proof.** From (9b) we have (since  $K_n \leq k$ )

$$\begin{aligned}
 P(N = \infty) &= \lim_{n \rightarrow \infty} P\{N > n\} \\
 &\leq \lim_{n \rightarrow \infty} P\{n < K_n\{(2K_n - 1)c + (2c - 1)\log c\}\} \\
 &= \lim_{n \rightarrow \infty} P\left\{K_n \geq \frac{n}{(2k - 1)c + (2c - 1)\log c}\right\} \\
 &= 0.
 \end{aligned} \tag{11}$$

This establishes the finite sure termination of the sequential procedure. ■

## 2.2 FIRST ORDER ASYMPTOTIC RESULTS

For sufficiently small  $c$ , the stopping time  $N$  of the sequential procedure  $(N_c, K_N)$  in (9a) can be rewritten as

$$N = \inf \left\{ n \geq n_0 : \frac{n}{k(-\log c)} \geq \frac{K_n}{k} \left[ 1 + \frac{\log(2c + 1)}{\log c} - \frac{\log\{(2K_n + 1)c + 2c \log c + 1\}}{\log c} \right] \right\}. \tag{12}$$

Now, set  $Y_n = (K_n/k)[1 + \{\log(2c + 1)/\log c\} - \{\log\{(2K_n + 1)c + 2c \log c + 1\}/\log c\}]$ , let  $f(n) = n$ , and  $t = -k \log c$ . Then  $Y_n$  is a sequence of random variables such that  $Y_n > 0$  a.s.,  $\lim_{n \rightarrow \infty} Y_n = 1$  a.s.,  $\lim_{n \rightarrow \infty} f(n) = \infty$ , and  $\lim_{n \rightarrow \infty} f(n)/f(n - 1) = 1$ .

Since the stopping rule  $N$  is well-defined and non-decreasing as a function of  $t$ , by applying the results of Chow and Robbins (1965) we obtain the following first order asymptotic result.

### **Result 1.**

- (i)  $\lim_{c \rightarrow 0} N = \infty$  a.s.,  $\lim_{c \rightarrow 0} E(N) = \infty$ ,
- (ii)  $\lim_{c \rightarrow 0} N/n^* = 1$  a.s.,
- (iii)  $\lim_{c \rightarrow 0} E(N)/n^* = 1$ .

**Proof.** (i) is easily verified by using (9b). (ii) follows from the facts that  $n^*$  given in (7) and  $\lim_{c \rightarrow 0} N/(-k \log c) = 1$  a.s.,  $\lim_{c \rightarrow 0} E(N)/(-k \log c) = 1$ . (iii) follows from  $E(\sup_n Y_n) \leq E[1 + \{\log(2c + 1)/\log c\} - \lfloor \log\{(2K_n + 1)c + 2c \log c + 1\} \rfloor / \log c] < \infty$ . ■

### **3 PERFORMANCE OF THE PROCEDURE**

The performance of the stopping rule in the sequential procedure  $(N_c, K_N)$  is usually evaluated by comparing two risks; the one is  $R_N(c)$ , the risk involved in sequential estimation of  $k$  using the proposed procedure  $(N_c, K_N)$ , and the other is  $R_{n^*}(c)$ , the risk associated with the optimal fixed-sample size  $n^*$ . The comparison would be carried out by considering the ratio and the difference as follows;

(i) the risk efficiency:  $R_N(c)/R_{n^*}(c)$ ,

(ii) the regret:  $R_N(c) - R_{n^*}(c)$ .

In most cases, there does not exist sequential procedures which are uniformly risk efficient or which have uniformly minimum regret. Therefore, we consider the risk efficiency in the asymptotic sense and the regret.

#### **3.1 RISK EFFICIENCY**

Now, we would like to show that the sequential procedure  $(N_c, K_n)$  is asymptotically risk-efficient, i.e., the ratio  $R_N(c)/R_{n^*}(c) \rightarrow 1$  or  $R_N(c) \sim R_{n^*}(c)$  as  $c \rightarrow 0$ . This can be shown by establishing the following theorem.

**Theorem 2.** *Let  $X_1, X_2, \dots$  be a sequence of independent observations from a multinomial population with unknown  $k$  equally probable distinct cells.*

## SEQUENTIAL ESTIMATION OF MULTINOMIAL CELLS

For  $c > 0$  and  $k < \infty$ , define the stopping time  $N \equiv N_c$  by

$$N = \inf \{n \geq n_0 : n \geq K_n [-\log c - \log(2c + 1) + \log \{(2K_n + 1)c + 2c \log c + 1\}]\},$$

where  $n_0$  ( $n_0 \geq 2$ ) is the initial sample size. Then the sequential procedure  $(N_c, K_N)$  is asymptotically risk-efficient, i.e.,

$$\lim_{c \rightarrow 0} \frac{R_N(c)}{R_{n^*}(c)} = 1. \quad (13)$$

**Remark 1.** To prove the asymptotic risk efficiency, we notice that some uniform integrability conditions are required because of the second moment of  $K_N$  in the risk. Clearly, since  $K_N \leq k < \infty$ , the sequence of random variables  $\{K_N, N \geq 1\}$  is uniformly integrable. The proof is trivial because self-evidently  $\sup_{N \geq 1} E(K_N) = k < \infty$ . Hence all conditions of uniform integrability and immediate consequences hold for our case.

**Proof.** Obviously  $N < \infty$  a.s., and as  $c \rightarrow 0$ ,  $N \rightarrow \infty$  a.s. Towards the risk efficiency, we have to show that

$$\frac{E(K_N - k)^2 + cE(N)}{E(K_{n^*} - k)^2 + cn^*} \rightarrow 1 \text{ as } c \rightarrow 0.$$

Taking limit on LHS,

$$\begin{aligned} \text{LHS} &= \lim_{c \rightarrow 0} \frac{E(K_N^2) - 2kE(K_N) + k^2 + cE(N)}{E(K_{n^*}^2) - 2kE(K_{n^*}) + k^2 + cn^*} \\ &= \lim_{c \rightarrow 0} \frac{E(K_N^2/k^2) - 2E(K_N/k) + 1 + cE(N/k^2)}{E(K_{n^*}^2/k^2) - 2E(K_{n^*}/k) + 1 + c(n^*/k^2)}. \end{aligned}$$

Since we have  $K_N/k \rightarrow 1$  a.s., and  $K_{n^*} \rightarrow k$  a.s. as  $c \rightarrow 0$ , so  $E(K_N/k) \rightarrow 1$ .

Also, since  $K_N$  is bounded by  $k$ ,  $K_N^2/k^2 \rightarrow 1$  a.s. and so  $E(K_N^2/k^2) \rightarrow 1$ .

From (7) we obtain

$$cn^*/k^2 = 2c - k^{-1}c + k^{-1}(2c - 1)c \log c,$$

which tends to zero as  $c \rightarrow 0$ . Further, we can write

$$\frac{cE(N)}{k^2} = \frac{cn^*}{k^2}E(N/n^*).$$

Now the proof is completed upon noting that  $E(N/n^*) \rightarrow 1$  as  $c \rightarrow 0$ . ■

In addition, we note that Chow and Yu (1981) generalized the asymptotic risk-efficiency of a sequential procedure with similar loss structure but without assuming any distributional form. Also, Lai (1996) established the asymptotic risk efficiency for more general stochastic sequences by using the uniform integrability and moment convergence of certain variables.

### 3.2 REGRET

One of the desirable properties that we studied above, namely, asymptotic risk efficiency can be strengthened by the property that regret goes to zero, which is considerably stronger than the asymptotic risk efficiency. Chow and Martinsek (1982) point out that uniform integrability results alone are not enough to prove the bounded regret property because some sort of cancellation is needed in the difference between  $R_N(c)$  and  $R_{n^*}(c)$ .

**Theorem 3.** *Let  $X_1, X_2, \dots$  be a sequence of independent observations from a multinomial population with unknown  $k$  equally probable distinct cells. For  $c > 0$  and  $k < \infty$ , define the stopping time  $N \equiv N_c$  by*

$$N = \inf \{n \geq n_0 : n \geq K_n [-\log c - \log(2c + 1)] + \log \{(2K_n + 1)c + 2c \log c + 1\}\},$$

where  $n_0$  ( $n_0 \geq 2$ ) is the initial sample size. Then for the sequential procedure  $(N_c, K_N)$ ,

$$\lim_{c \rightarrow 0} \{R_N(c) - R_{n^*}(c)\} = 0. \tag{14}$$

## SEQUENTIAL ESTIMATION OF MULTINOMIAL CELLS

**Proof.** From (10) and since the loss function for the optimal sample size  $n^*$  is  $L_{n^*} = (K_{n^*} - k)^2 + cn^*$ ,

$$\begin{aligned} LHS &= \lim_{c \rightarrow 0} [E(K_N - k)^2 + cE(N) - E(K_{n^*} - k)^2 - cn^*] \\ &= \lim_{c \rightarrow 0} [E(K_N^2) - E(K_{n^*}^2) - 2k \{E(K_N) - E(K_{n^*})\} + \{cE(N) - cn^*\}]. \end{aligned}$$

Since  $\lim_{c \rightarrow 0} \{cE(N) - cn^*\} = \lim_{c \rightarrow 0} [cn^* \{E(N/n^*) - 1\}]$ ,  $E(N/n^*) \rightarrow 1$  as  $c \rightarrow 0$  due to (iii) in Result 1, and furthermore,  $cn^* \approx -kc \log c \rightarrow 0$  as  $c \rightarrow 0$ , so  $\lim_{c \rightarrow 0} \{cE(N) - cn^*\} = 0$ . Then,

$$\begin{aligned} LHS &= \lim_{c \rightarrow 0} [E(K_N^2) - E(K_{n^*}^2) - 2k \{E(K_N) - E(K_{n^*})\}], \\ &= \lim_{c \rightarrow 0} \left[ k^2 \left\{ E\left(\frac{K_N^2}{k^2}\right) - E\left(\frac{K_{n^*}^2}{k^2}\right) \right\} - 2 \left\{ E\left(\frac{K_N}{k}\right) - E\left(\frac{K_{n^*}}{k}\right) \right\} \right]. \end{aligned}$$

Now, using similar arguments as in the proof of Theorem 2, we come to the conclusion that  $LHS = 0$ . This completes the proof. ■

## 4 SIMULATION STUDY

### 4.1 MONTE CARLO EXPERIMENTATION

For a simulation study, the Monte Carlo method is used in order to illustrate the performance of the proposed stopping rule in the sequential procedure. The numerical results indicate the small sample behavior and provide support for the asymptotic behavior of the sequential procedure as  $c \rightarrow 0$ .

The results of the Monte Carlo simulation, based on the sequential rule (9a), are summarized in the following tables, which contain the average of estimates  $\hat{k}$  of  $K_n$ , the average of stopping time,  $E(N)$ , the average risk associated with the stopping time  $N$ ,  $Risk(N)$ , the risk under the optimal fixed-sample size  $n^*$ ,  $Risk(n^*)$ , the risk efficiency which is the ratio  $R(n^*)/R(N)$ , and the regret which is the difference  $R(N) - R(n^*)$ .

TABLE 1. The True Number of Cells,  $k = 5$ 

$c$	$\hat{k}$	$E(N)$	$Risk(N)$	$Risk(n^*)$	$Risk\ Eff$	$Regret$
.10	4.08	10.99	3.269	1.726	.528	1.543
.05	4.45	14.23	1.906	1.010	.530	.896
.02	4.77	19.09	.817	.492	.602	.325
.01	4.90	23.53	.370	.281	.759	.089
.005	4.96	26.79	.188	.157	.837	.031
.002	4.98	31.84	.085	.072	.852	.013
.001	4.99	34.94	.044	.040	.909	.004

TABLE 2. The True Number of Cells,  $k = 10$ 

$c$	$\hat{k}$	$E(N)$	$Risk(N)$	$Risk(n^*)$	$Risk\ Eff$	$Regret$
.10	8.99	27.65	6.566	3.670	.559	2.787
.05	9.40	32.66	3.525	2.102	.596	1.383
.02	9.76	40.83	1.211	1.001	.826	.193
.01	9.88	46.50	.614	.565	.920	.047
.005	9.94	53.63	.334	.316	.945	.024
.002	9.98	62.86	.150	.144	.964	.007
.001	9.99	69.90	.080	.079	.988	.001

TABLE 3. The True Number of Cells,  $k = 15$ 

$c$	$\hat{k}$	$E(N)$	$Risk(N)$	$Risk(n^*)$	$Risk\ Eff$	$Regret$
.10	14.09	47.96	8.407	5.822	.693	2.584
.05	14.41	53.64	4.753	3.247	.751	1.081
.02	14.73	62.88	1.691	1.519	.898	.172
.01	14.84	71.22	.899	.853	.949	.046
.005	14.92	80.60	.491	.475	.967	.016
.002	14.97	93.80	.223	.217	.973	.006
.001	14.98	103.88	.121	.119	.983	.002

TABLE 4. The True Number of Cells,  $k = 20$ 

$c$	$\hat{k}$	$E(N)$	$Risk(N)$	$Risk(n^*)$	$Risk\ Eff$	$Regret$
.10	19.24	70.48	9.821	8.082	.823	1.739
.05	19.46	76.45	5.206	4.447	.854	.759
.02	19.67	86.38	2.268	2.050	.904	.218
.01	19.82	97.07	1.206	1.145	.949	.061
.005	19.90	108.48	.662	.636	.961	.026
.002	19.96	125.70	.297	.290	.976	.007
.001	19.99	138.85	.161	.158	.984	.003

## SEQUENTIAL ESTIMATION OF MULTINOMIAL CELLS

In the above tables, we observe that all of Monte Carlo estimate  $\hat{k}$  is uniformly smaller than  $k$ , i.e., it underestimates  $k$ . However, the estimate  $\hat{k}$  of  $K_n$  converges to the corresponding true number of cells as  $c \rightarrow 0$ . The risk efficiency approaches one, and the regret gets close to zero as  $c \rightarrow 0$  respectively, which provide a substantial amount of numerical evidence and strong belief, for concluding that the proposed sequential estimator is quite good.

We note that both  $E(N)$  and  $Risk(N)$  are based on averages of the observed values given by (9a) for 5,000 independent experiments for each selected values of  $c$ . We see from tables 1 through 4 that  $E(N)$  increases as the sampling cost per observation,  $c$ , gets smaller. However, the average risk under stopping time  $N$  decreases dramatically as  $c \rightarrow 0$ . Recall that the loss function we assumed incorporates both losses from estimation and costs from sampling. In this context, the value of  $c$  plays an ‘*inflating-sample factor*’ in the sequential procedure.

### **4.2 METHODS OF APPROXIMATION AND EXAMPLES**

For the approximation of the procedure, we consider the case of  $c \rightarrow 0$ . From the stopping time (6),

$$\begin{aligned} n &= k[-\log c - \log(2c + 1) + \log\{(2k + 1)c + 2c \log c + 1\}] \\ &= k[-\log c - 2c + (2k + 1)c + 2c \log c + O(c^2)] \\ &\approx k\{(2k - 1)c + (2c - 1) \log c\}, \end{aligned} \tag{15}$$

assuming  $(2k + 1)c + 2c \log c < 1$ .

The adaptive sequential procedure is: Stop at  $N$  such that

$$N = \inf[n \geq n_0 : n \geq K_n\{(2K_n - 1)c + (2c - 1) \log c\}], \tag{16}$$

and using the expansion on  $y = \exp(-n/k)$ , the minimum risk  $R^*(c)$  becomes

$$\begin{aligned} R^*(c) &= kc^2(k-1)\{1-4(k-1)c+O(c^2)\} + kc\{1-2(k-1)c+O(c^2)\} \\ &\quad -kc\log[c\{1-2(k-1)c+O(c^2)\}] + kc^2\{1-4(k-1)c+O(c^2)\} \\ &\quad \cdot \log[c^2\{1-4(k-1)c+O(c^2)\}] \end{aligned} \quad (17)$$

$$\begin{aligned} &= kc\{1+O(c^2)\} - kc\log\{c+O(c^2)\} \\ &\approx kc(1-\log c). \end{aligned} \quad (18)$$

We also note that one can easily get another term in the expansion for the minimum  $R^*$  involving  $O(c^2)$  term.

**Example 1.** Suppose that  $k = 5$ ,  $c = 0.01$ . Then

$$\begin{aligned} n^* &= 5[(9)0.01 + (0.02 - 1)\log 0.01] \\ &= 23.02, \end{aligned}$$

and the minimum  $R_{n^*}$ ,

$$\begin{aligned} R_{n^*} &= 5(0.01)[1 - \log 0.01] \\ &= 0.280, \end{aligned}$$

whereas the averages of the simulated values based on 5000 replications are  $n = 23.53$  and  $R = 0.281$  respectively. So the values obtained by the expansion are fairly close to the simulated values.

**Example 2.** Suppose that  $k = 10$ ,  $c = 0.005$ . Then

$$\begin{aligned} n^* &= 10[(19)0.005 + (0.01 - 1)\log 0.005] \\ &= 53.40, \end{aligned}$$

and the minimum  $R_{n^*}$ ,

$$\begin{aligned} R_{n^*} &= 10(0.005)[1 - \log(0.005)] \\ &= 0.315, \end{aligned}$$

## SEQUENTIAL ESTIMATION OF MULTINOMIAL CELLS

whereas the averages of the simulated values based on 5000 replications are  $n = 53.63$  and  $R = 0.316$  respectively. So the values obtained by the expansion are in good agreement with the simulated values. This implies that we are able to get better approximations when the inflating-sample factor,  $c$ , gets smaller, which increases the sample size in the sequential procedure.

### ACKNOWLEDGMENT

We thank the referee for a careful reading of the manuscripts.

### REFERENCES

- Boender, C. and Rinnooy Kan, A. (1983). A Bayesian analysis of the number of cells of a multinomial distribution. The Statistician, 32, 240-248.
- Boender, C. and Zielinski, R. (1985). A sequential Bayesian approach to estimating the dimension of a multinomial distribution. Sequential Methods in Statistics, Banach Center Publications, 16, 37-41, PWN-Polish Scientific Publishers, Warsaw.
- Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: a review. Journal of the American Statistical Association, 88, 364-373.
- Chow, Y. S. and Martinsek, A. T. (1982). Bounded regret of a sequential procedure for estimation of the mean. The Annals of Statistics, 10, 909-914.
- Chow, Y. S. and Robbins, H. (1965). On the asymptotic theory of fixed width sequential confidence intervals for the mean. The Annals of Mathematical Statistics, 36, 457-462.
- Chow, Y. S. and Yu, K. F. (1981). The performance of a sequential procedure for the estimation of the mean. The Annals of Statistics, 9, 184-189.
- Feller, W. (1968). An Introduction to Probability Theory and its Applications, Vol. 1, 3rd Ed. John Wiley, New York.
- Finkelstein, M., Tucker, H. and Veeh, J. (1998). Confidence intervals for the number of unseen types. Statistics and Probability Letters, 37, 423-430.

Goodman, L. (1949). On the estimation of the number of classes in a population. The Annals of Mathematical Statistics, 20, 572-579.

Govindarajulu, Z. (1999). The Elements of Sampling Theory and Methods. Prentice-Hall, Inc., New Jersey.

Harris, B. (1968). Statistical inference in the classical occupancy problem unbiased estimation of the number of classes. Journal of the American Statistical Association, 63, 837-847.

Jordan, C. (1950). Calculus of Finite Differences, 2nd Ed., Chelsea Publishing Co., New York.

Lai, T. L. (1996) On uniform integrability and asymptotically risk-efficient sequential estimation. Sequential Analysis. 15, 237-251.

Weiss, I. (1958). Limiting Distributions in some Occupancy Problems. The Annals of Mathematical Statistics, 29, 878-884.