

RISK-EFFICIENT SEQUENTIAL ESTIMATION OF THE NUMBER OF MULTINOMIAL CELLS

Hokwon Cho
Department of Mathematical Sciences
University of Nevada, Las Vegas, Las Vegas, NV 89154

Z. Govindarajulu
Department of Statistics
University of Kentucky, Lexington, KY, 40506

August 9, 2004

Abstract

We consider risk-efficient sequential estimation under squared error loss of the number of classes which are equally probable to occur in a given multinomial population. It is assumed that the sampling cost per observation is constant. Large-sample properties of the sequential estimator are studied. Finally, Monte Carlo simulation is carried out in order to investigate its finite sample behavior. The proposed sequential procedure performs better (in the sense of reducing average stopping time and risk) than the one based on the estimator K_n , the number of distinct cells observed in a sample of size n .

Key Words: Sequential estimation; number of classes; Multinomial population; Risk efficiency; Squared error loss; Constant sampling cost.

1 INTRODUCTION

Various methods for point estimation of the number of multinomial classes in a given population have been proposed and developed. Goodman (1949) seems to be the first to consider this problem. We wish to estimate the unknown number of classes (or cells) based on a sample of size n drawn from a multinomial population. This model is applicable to practical situations in biological sciences, numismatics, etc. (For instance, see Arnold and Beaver (1988), p. 413.)

Bunge and Fitzpatrick (1993) provide a comprehensive review of the existing literature relating to this problem. Based on the number of distinct cells in a

⁰*Mathematics Subject Classification:* 62L12, 62F10

random sample, K_n (in fact it is a biased estimator), Cho and Govindarajulu (2002) have considered the sequential estimation of the unknown number of cells with squared error loss plus constant cost of sampling. Here we consider an almost unbiased estimator of the unknown number of classes, and reduce not only the average stopping time but also the risk.

1.1 Almost Unbiased Estimator of k

Let X_1, X_2, \dots be a sequence of independent observations from a multinomial population with unknown number, k (≥ 2) equally probable distinct classes. Then

$$P(X_i \in C_j) = 1/k, \quad i = 1, 2, \dots; \quad j = 1, 2, \dots, k, \quad (1.1)$$

where C_j is j -th class. With a sample (X_1, X_2, \dots, X_n) of size n , we wish to estimate k ($< \infty$), the true number of classes based on an estimator \hat{k}_n of k under the squared error loss function plus constant cost of sampling of the form

$$L_n = (\hat{k}_n - k)^2 + cn. \quad (1.2)$$

where c ($c > 0$) is proportional to the cost per observation.

Let K_n denote the number of distinct cells observed in a sample of size n , then $K_n \leq k$, which is a biased estimator of k . Sufficiency of K_n was noted by Goodman (1949) and the completeness of K_n was established by Harris (1968). Goodman (1953) proposed a heuristic sequential procedure of sampling until L observations repeated classes previously observed. Harris (1968) points out that Goodman's (1953) sequential estimate of k is somewhat crude and gave a slightly improved estimate. Harris (1968) has shown that the best unbiased estimate of k is the ratio of two Stirling numbers of the second kind. For large n , Harris (1968) approximates the best unbiased estimate of k by \hat{k}_n where

$$\hat{k}_n = (n + 1) / R_1 \quad (1.3)$$

and R_1 is the solution of the equation

$$R(1 - e^{-R})^{-1} = (n + 1) / k. \quad (1.4)$$

In the following we try to obtain a closed-form expression for \hat{k}_n in terms of K_n and use it to develop a sequential procedure.

Letting $x = (n + 1) / k$ and $R = x(1 + \varepsilon)$, we have

$$x(1 + \varepsilon) = x \left[1 - e^{-x(1+\varepsilon)} \right], \quad (1.5)$$

or

$$\varepsilon = -e^{-x(1+\varepsilon)} \approx -e^{-x}. \quad (1.6)$$

Hence

$$\begin{aligned}
\hat{k}_n &= \frac{n+1}{[(n+1)/K_n][1 - e^{-(n+1)/K_n}]} \\
&= K_n [1 - e^{-(n+1)/K_n}]^{-1} \\
&\doteq K_n [1 + e^{-(n+1)/K_n}].
\end{aligned} \tag{1.7}$$

Then, the risk function

$$\begin{aligned}
R_n(c) &= E(\hat{k}_n - k)^2 + cn \\
&= \text{var}(\hat{k}_n) + [E(\hat{k}_n - k)]^2 + cn.
\end{aligned} \tag{1.8}$$

We can write, after letting $f(x) = x(1 + e^{-n/x})$

$$\begin{aligned}
k - \hat{k}_n &= k - f(K_n) = k - f(k) + [f(k) - f(K_n)] \\
&\doteq -ke^{-n/k} + U_n f'(k)
\end{aligned} \tag{1.9}$$

where $U_n = k - K_n$.

The exact mean and variance of K_n obtained by Weiss (1958) are given by (See also Govindarajulu (1999), p. 44)

$$E(K_n) = k [1 - (1 - 1/k)^n], \tag{1.10}$$

and

$$\text{var}(K_n) = k(1 - 1/k)^n - k^2(1 - 1/k)^{2n} + k(k-1)(1 - 2/k)^n. \tag{1.11}$$

Let

$$\alpha_1 = k \log(1 - 1/k) = -1 - (1/2k) - (1/3k^2) - \dots \tag{1.12}$$

and

$$\alpha_2 = k \log(1 - 2/k) = -2 - (1/k) - (2/3k^2) - \dots \tag{1.13}$$

Then, we can rewrite the first two moments of K_n in terms of α_1 and α_2 as

$$E(K_n) = k \left(1 - e^{\alpha_1 n/k}\right), \tag{1.14}$$

and

$$\text{var}(K_n) = ke^{\alpha_1 n/k} - k^2 e^{2\alpha_1 n/k} + k(k-1)e^{\alpha_2 n/k}. \tag{1.15}$$

By letting $\hat{k}_n \doteq K_n [1 + e^{-n/K_n}] \equiv f(K_n)$, we have

$$E(\hat{k}_n) \doteq E[f(k) + (K_n - k)f'(k)] \tag{1.16}$$

where $f'(k) = 1 + (1 + n/k) \exp(-n/k)$.

Hence, the bias term in \hat{k}_n is

$$\begin{aligned} E(\hat{k}_n) - k &= f(k) - k - kf'(k) + f'(k) E(K_n) \\ &= ke^{-n/k} - k \left[1 + (1 + n/k) e^{-n/k} \right] + f'(k) \left[k - ke^{\alpha_1 n/k} \right] \\ &= k \left[e^{-n/k} - f'(k) e^{\alpha_1 n/k} \right] \end{aligned} \quad (1.17)$$

and

$$\begin{aligned} \text{var}(\hat{k}_n) &\doteq [f'(k)]^2 \text{var}(K_n) \\ &= [f'(k)]^2 k \left[e^{\alpha_1 n/k} - ke^{2\alpha_1 n/k} + (k-1) e^{\alpha_2 n/k} \right]. \end{aligned} \quad (1.18)$$

Thus, we have

$$\begin{aligned} E(\hat{k}_n - k)^2 &= k [f'(k)]^2 \left[e^{\alpha_1 n/k} + (k-1) e^{\alpha_2 n/k} \right] \\ &\quad + k^2 \left[e^{-2n/k} - 2f'(k) e^{-n/k + \alpha_1 n/k} \right]. \end{aligned} \quad (1.19)$$

Then, using (1.19) in the risk $R_n(c)$ given by (1.8) and letting $\theta = n/k$, $f'(k) = 1 + (1 + \theta) e^{-\theta}$, we have

$$\begin{aligned} R_n(c)/k \equiv H(\theta) &= [1 + (1 + \theta) e^{-\theta}]^2 \left[e^{\alpha_1 \theta} + (k-1) e^{\alpha_2 \theta} \right] \\ &\quad + k \left[e^{-2\theta} - 2 \{ 1 + (1 + \theta) e^{-\theta} \} e^{-\theta + \alpha_1 \theta} \right] + c\theta. \end{aligned} \quad (1.20)$$

Now,

$$\begin{aligned} \partial H(\theta) / \partial \theta \equiv g(\theta) &= -2\theta e^{(\alpha_1 - 1)\theta} + \alpha_1 e^{\alpha_1 \theta} + 2(1 + \theta) \alpha_1 e^{(\alpha_1 - 1)\theta} \\ &\quad + (k-1) \alpha_2 e^{\alpha_2 \theta} + 2k(1 - \alpha_1) e^{-\theta(1 - \alpha_1)} \\ &\quad - 2ke^{-2\theta} + c + o(e^{-2\theta}). \end{aligned} \quad (1.21)$$

By setting $\partial H(\theta) / \partial \theta = 0$, we solve for θ for which the risk is minimum. Ignoring terms inferior to $e^{-2\theta}$ in $g(\theta)$, we solve for θ by setting $g(\theta) = 0$.

In order to get the initial solution, we solve the equation

$$\alpha_1 e^{\alpha_1 \theta} = -c \quad (1.22)$$

and obtain

$$\theta_1 = \alpha_1^{-1} \log(-c/\alpha_1). \quad (1.23)$$

Now, expanding $g(\theta)$ given by (1.21) about θ_1 , we obtain

$$g(\theta) \doteq g(\theta_1) + (\theta - \theta_1) g'(\theta_1) = 0, \quad (1.24)$$

hence

$$\theta_2 = \theta_1 - g(\theta_1) / g'(\theta_1), \quad (1.25)$$

where

$$g'(\theta) \doteq -2e^{(\alpha_1-1)\theta} - 2\theta(\alpha_1-1)e^{(\alpha_1-1)\theta} - 2\theta e^{-\theta}\alpha_1 e^{\alpha_1\theta} + \alpha_1^2 e^{-\alpha_1\theta} - 2(1-\alpha_1)^2 e^{-\theta(1-\alpha_1)}, \quad (1.26)$$

after ignoring terms inferior to $e^{-2\theta}$.

Then, we have

$$g(\theta_1)/g'(\theta_1) \doteq \alpha_1^{-2} e^{-\theta_1} \left[-2\theta_1 + 2(1+\theta_1)\alpha_1 + (k-1)\alpha_2 e^{(\alpha_2-\alpha_1+1)\theta_1} + 2k(1-\alpha_1) - 2ke^{-(1+\alpha_1)\theta_1} \right]. \quad (1.27)$$

Hence, the second approximation for θ is

$$\theta_2 = \theta_1 - e^{-\theta_1} \alpha_1^{-2} \left[-2\theta_1 + 2(1+\theta_1)\alpha_1 + (k-1)\alpha_2 e^{(\alpha_2-\alpha_1+1)\theta_1} + 2k(1-\alpha_1) - 2ke^{-(1+\alpha_1)\theta_1} \right]. \quad (1.28)$$

Since $\theta = n/k$, the optimal fixed-sample size when everything is known is given by

$$n^* = k \left[\theta_1 - \alpha_1^{-2} e^{-\theta_1} \left\{ -2\theta_1 + 2(1+\theta_1)\alpha_1 + (k-1)\alpha_2 e^{(\alpha_2-\alpha_1+1)\theta_1} + 2k(1-\alpha_1) - 2ke^{-(1+\alpha_1)\theta_1} \right\} \right], \quad (1.29)$$

where θ_1 is given by (1.23), α_1 and α_2 are defined by (1.12) and (1.13) respectively.

Then, the minimum risk associated with the optimal fixed-sample size n^* is

$$R_{n^*}(c) = E \left(\hat{k}_{n^*} - k \right)^2 + cn^*. \quad (1.30)$$

However, since k is unknown, there is no fixed-sample size procedure that will attain the risk (1.8). So, we resort to the following adaptive sequential procedure: Stop sampling at N where

$$N = \inf \left\{ n \geq n_0 : n \geq \hat{k}_n \left[\hat{\theta}_1 - \hat{\alpha}_1^{-2} e^{-\hat{\theta}_1} \left\{ -2\hat{\theta}_1 + 2(1+\hat{\theta}_1)\hat{\alpha}_1 + (\hat{k}_n-1)\hat{\alpha}_2 e^{(\hat{\alpha}_2-\hat{\alpha}_1+1)\hat{\theta}_1} + 2\hat{k}_n(1-\hat{\alpha}_1) - 2\hat{k}_n e^{-(1+\hat{\alpha}_1)\hat{\theta}_1} \right\} \right] \right\} \quad (1.31)$$

and n_0 (≥ 2) denotes the initial sample size, \hat{k}_n is given by (1.7), and $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are obtained by replacing k by \hat{k}_n , and $\hat{\theta}_1$ is obtained by replacing α_1 by $\hat{\alpha}_1$ and α_2 by $\hat{\alpha}_2$.

When k is sufficiently large, say $k \geq 25$ (independently of c getting small), $\alpha_1 \sim -1$, $\alpha_2 \sim -2$, $\theta_1 \sim -\log c$, and hence the stopping time given by (1.31) takes

$$N = \inf \left\{ n \geq n_0 : n \geq \hat{k}_n (1+4c)(-\log c) \right\}. \quad (1.32)$$

Then, the risk function corresponding to the random stopping time N is given by

$$R_N(c) = E \left[(\hat{k}_N - k)^2 + cN \right]. \quad (1.33)$$

2 ASYMPTOTIC BEHAVIOR OF THE PROCEDURE

In this section, we investigate the asymptotic behavior of the sequential procedure (N, \hat{k}_N) and study some properties of the stopping time N .

2.1 Finite Sure Termination

Here, we want to establish the fundamental property that the proposed stopping rule terminates finitely almost surely.

Theorem 2.1 *Let N denote the stopping time in the sequential procedure. Then, assuming that $c > 0$, $P(N = \infty) = 0$.*

Proof. From the stopping time for the sequential procedure we have,

$$\begin{aligned} P(N = \infty) &= \lim_{n \rightarrow \infty} P(N > n) \\ &\leq \lim_{n \rightarrow \infty} P \left\{ n < \hat{k}_n \left[\hat{\theta}_1 - \hat{\alpha}_1^{-2} e^{-\hat{\theta}_1} \left\{ -2\hat{\theta}_1 + 2 \left(1 + \hat{\theta}_1 \right) \hat{\alpha}_1 \right. \right. \right. \\ &\quad \left. \left. \left. + \left(\hat{k}_n - 1 \right) \hat{\alpha}_2 e^{(\hat{\alpha}_2 - \hat{\alpha}_1 + 1)\theta_1} + 2\hat{k}_n (1 - \hat{\alpha}_1) - 2\hat{k}_n e^{-(1 + \hat{\alpha}_1)\hat{\theta}_1} \right\} \right] \right\} \\ &= 0 \end{aligned} \tag{2.1}$$

due to the fact that $\hat{k}_n \rightarrow k$, $\hat{\alpha}_1 \rightarrow k \log(1 - 1/k)$, and $\hat{\alpha}_2 \rightarrow k \log(1 - 2/k)$ as $n \rightarrow \infty$ since \hat{k}_n is a function of K_n and $K_n \rightarrow k$ almost surely (a.s.) as $n \rightarrow \infty$. (See Finkelstein, Tucker and Veeh (1998).) Thus, n is finite with probability one. ■

2.2 First Order Asymptotic Results

For sufficiently small c , the stopping time N of the sequential procedure (N, \hat{k}_N) given by (1.31) can be rewritten as

$$\begin{aligned} N &= \inf \left\{ n \geq n_0 : \frac{n}{k\theta_1} \geq \frac{\hat{\theta}_1 \hat{k}_n}{\theta_1 k} \left[1 - \frac{\hat{\alpha}_1^{-2} e^{-\hat{\theta}_1}}{\hat{\theta}_1} \left\{ -2\hat{\theta}_1 + 2 \left(1 + \hat{\theta}_1 \right) \hat{\alpha}_1 \right. \right. \right. \\ &\quad \left. \left. \left. + \left(\hat{k}_n - 1 \right) \hat{\alpha}_2 e^{(\hat{\alpha}_2 - \hat{\alpha}_1 + 1)\theta_1} + 2\hat{k}_n (1 - \hat{\alpha}_1) - 2\hat{k}_n e^{-(1 + \hat{\alpha}_1)\hat{\theta}_1} \right\} \right] \right\}. \end{aligned} \tag{2.2}$$

Now, set

$$\begin{aligned} W_n &= \frac{\hat{\theta}_1 \hat{k}_n}{\theta_1 k} \left[1 - \frac{\hat{\alpha}_1^{-2} e^{-\hat{\theta}_1}}{\hat{\theta}_1} \left\{ -2\hat{\theta}_1 + 2 \left(1 + \hat{\theta}_1 \right) \hat{\alpha}_1 + 2\hat{k}_n (1 - \hat{\alpha}_1) \right. \right. \\ &\quad \left. \left. + \left(\hat{k}_n - 1 \right) \hat{\alpha}_2 e^{(\hat{\alpha}_2 - \hat{\alpha}_1 + 1)\theta_1} - 2\hat{k}_n e^{-(1 + \hat{\alpha}_1)\hat{\theta}_1} \right\} \right]. \end{aligned} \tag{2.3}$$

Then, W_n is a sequence of random variables such that $W_n > 0$ a.s., $\lim_{n \rightarrow \infty} W_n = 1$ a.s., and as before $K_n \rightarrow k$ a.s. as $n \rightarrow \infty$.

Using Lemmas 1 and 2 of Chow and Robbins (1965), we obtain the following first order asymptotic results.

Theorem 2.2 For the stopping time N ,

- (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. Using (1.31), (i) is verified without difficulty. Towards the proof of (ii), Setting $f(n) = n$, $t = k\theta_1$ in Chow and Robbins (1965) we have for $N > n_0 \geq 2$

$$W_N < f(N)/t \text{ and also } W_{N-1} > f(N-1)/t. \quad (2.4)$$

Hence,

$$W_N < \frac{f(N)}{t} = \frac{f(N)}{f(N-1)} \cdot \frac{f(N-1)}{t} < \frac{f(N)}{f(N-1)} W_{N-1}. \quad (2.5)$$

Now taking limits on both sides of the above inequality, we obtain $\lim_{N \rightarrow \infty} f(N)/t = 1$ a.s. Towards the proof of (iii), we note that since $-\alpha_1 > 1$, $\log[c/(-\alpha_1)] < \log c$. Also

$$\begin{aligned} -\alpha_1 &= 1 + (1/2k) + (1/3k^2) + \dots < 1 + (1/2k)(1 + 1/k + 1/k^2 + \dots) \\ &= 1 + [2(k-1)]^{-1} < 1 + 1/2. \end{aligned} \quad (2.6)$$

Hence,

$$\frac{\hat{\theta}_1}{\theta_1} = \frac{\alpha_1 \log(c - \hat{\alpha}_1)}{\hat{\alpha}_1 \log(c - \alpha_1)} \leq \frac{-1}{-(3/2)} \frac{\log(2c/3)}{\log c} = \frac{2 \log(2c/3)}{3 \log c} = A \text{ (say)}. \quad (2.7)$$

Consequently,

$$W_n \leq \left[\left(\frac{\hat{\theta}_1}{\theta_1} \right) \left(\frac{\hat{k}_n}{k} \right) \right] \leq A \left(\frac{\hat{k}_n}{k} \right).$$

Thus,

$$\sup_n W_n < A \sup_n \hat{k}_n/k. \quad (2.8)$$

In order to apply Lemma 2 of Chow and Robbins (1965) for obtaining Result (iii), it suffices to show that $E(\sup_n \hat{k}_n/k) < \infty$. Toward this, we have

$$\begin{aligned} \hat{k}_n/k &= (K_n/k) \left[1 + e^{-(n+1)/K_n} \right] \leq \left[1 + e^{-\{(n+1)/k\}/(K_n/k)} \right] \\ &\leq \left[1 + e^{-(n+1)/k} \right] \leq \left[1 + e^{-(n_0+1)/k} \right] \end{aligned} \quad (2.9)$$

where n_0 is the initial sample size.

Therefore

$$\sup_n \left(\frac{\hat{k}_n}{k} \right) \leq 1 + e^{-(n_0+1)/k} \leq 1 + e^{-3/k} \quad (2.10)$$

since $n_0 \geq 2$. Hence $E \left[\sup_n \left(\frac{\hat{k}_n}{k} \right) \right] < \infty$. This completes the proof of (iii).

Similarly, we can establish $E \left[\sup_n \left(\frac{\hat{k}_n^2}{k^2} \right) \right] < \infty$. ■

3 PERFORMANCE OF THE PROCEDURE

The performance of the sequential procedure 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, and the other is $R_{n^*}(c)$, the risk associated with the optimal fixed-sample size n^* . The comparison would be made by studying the ratio and the difference of two risks given by:

- (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 that are uniformly risk-efficient or which have uniformly minimum regret. Therefore, we consider the risk-efficiency and the regret as c tends to zero.

3.1 Risk-Efficiency

Now, we would like to show that the sequential procedure 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 3.1 *Let X_1, X_2, \dots be a sequence of independent observations from a multinomial distribution having unknown k equally probable distinct classes. For $c > 0$ and $k < \infty$, define the stopping time N by (1.31). Then the sequential procedure is asymptotically risk-efficient, i.e.,*

$$\lim_{c \rightarrow 0} R_N(c)/R_{n^*}(c) = 1. \quad (3.1)$$

Proof. As $c \rightarrow 0, N \rightarrow \infty$ a.s. Towards the risk efficiency, we have to show that

$$\frac{E(\hat{k}_N - k)^2 + cE(N)}{E(\hat{k}_{n^*} - k)^2 + cn^*} \rightarrow 1 \text{ as } c \rightarrow 0. \quad (3.2)$$

Taking limit on left-hand side (LHS),

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

Since we have $\hat{k}_N/k \rightarrow 1$ a.s. (see (1.7)), and $\hat{k}_{n^*} \rightarrow k$ a.s. as $c \rightarrow 0$, so $E(\hat{k}_N/k) \rightarrow 1$ and $E(\hat{k}_{n^*}^2/k^2) \rightarrow 1$. (See the proof of Theorem 2.2 (iii)). From (1.29) we obtain

$$cn^*/k \leq c\theta_1 = c\alpha_1^{-1} \log(c/\alpha_1) \leq (-c \log c)/(-\alpha_1) \leq -c \log c,$$

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

$$cE(N)/k^2 = (cn^*/k^2) E(N/n^*). \quad (3.3)$$

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

3.2 Regret

The asymptotic risk-efficiency is established in Section 3.1 assuming a certain uniform integrability result. However, the property of risk-efficiency can be strengthened by showing that the regret goes to zero. Chow and Martinsek (4) point out that uniform integrability results alone are not sufficient to prove that the regret is bounded because cancellation of some terms is required in the difference between $R_N(c)$ and $R_{n^*}(c)$, especially when both become large.

Theorem 3.2 *Let X_1, X_2, \dots be a sequence of independent observations from a multinomial distribution with unknown k equally probable distinct cells. For $c > 0$ and $k < \infty$, define the stopping time N by (1.31). Then for the sequential procedure (N, \hat{k}_N) ,*

$$\lim_{c \rightarrow 0} \{R_N(c) - R_{n^*}(c)\} = 0. \quad (3.4)$$

Proof. Since the loss function for the optimal sample size n^* is $L_{n^*} = (\hat{k}_{n^*} - k)^2 + cn^*$,

$$\begin{aligned} \text{LHS of (3.4)} &= \lim_{c \rightarrow 0} \left[E(\hat{k}_N - k)^2 + cE(N) - E(\hat{k}_{n^*} - k)^2 - cn^* \right] \\ &= \lim_{c \rightarrow 0} \left[E(\hat{k}_N^2) - E(\hat{k}_{n^*}^2) - 2k \left\{ E(\hat{k}_N) - E(\hat{k}_{n^*}) \right\} + \{cE(N) - cn^*\} \right]. \end{aligned}$$

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) of Theorem 2.2, and furthermore, $cn^* \approx kc\theta_1 \rightarrow 0$ as $c \rightarrow 0$, so $\lim_{c \rightarrow 0} \{cE(N) - cn^*\} = 0$. Hence,

$$\begin{aligned} \text{LHS} &= \lim_{c \rightarrow 0} \left[E(\hat{k}_N^2) - E(\hat{k}_{n^*}^2) - 2k \left\{ E(\hat{k}_N) - E(\hat{k}_{n^*}) \right\} \right], \\ &= \lim_{c \rightarrow 0} \left[k^2 \left\{ E\left(\hat{k}_N^2/k^2\right) - E\left(\hat{k}_{n^*}^2/k^2\right) \right\} - 2 \left\{ E\left(\hat{k}_N/k\right) - E\left(\hat{k}_{n^*}/k\right) \right\} \right]. \end{aligned}$$

Now, using arguments similar to those employed in the proof of Theorem 3.1, we can show that LHS = 0. This completes the proof. ■

4 NUMERICAL STUDIES

4.1 Monte Carlo Experimentation

The Monte Carlo method is used in order to investigate the finite sample behavior and to illustrate the performance of the proposed 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 (1.31), 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 $R(N) - R(n^*)$. The simulation results are based on 5000 replications for each selected value of c .

Table 1: The True Number of Classes, $k = 5$

c	\hat{k}	$E(N)$	n^*	$Risk(N)$	$Risk(n^*)$	Risk Eff	Regret
.10	4.67	13.46	15	2.191	1.746	1.255	.445
.05	4.83	16.18	17	1.240	.996	1.250	.244
.01	4.91	22.34	23	.324	.263	1.232	.061
.005	4.95	25.57	25	.176	.146	1.205	.030
.001	4.99	31.83	32	.042	.036	1.167	.006

Table 2: The True Number of Classes, $k = 10$

c	\hat{k}	$E(N)$	n^*	$Risk(N)$	$Risk(n^*)$	Risk Eff	Regret
.10	9.64	27.63	29	4.340	3.563	1.218	.777
.05	9.79	32.84	34	2.405	2.056	1.170	.349
.01	9.95	46.34	46	.595	.547	1.088	.048
.005	9.97	51.62	52	.326	.305	1.069	.021
.001	9.99	66.88	67	.080	.076	1.053	.004

Table 3: The True Number of Classes, $k = 20$

c	\hat{k}	$E(N)$	n^*	$Risk(N)$	$Risk(n^*)$	Risk Eff	Regret
.10	19.54	56.12	58	7.728	7.249	1.066	.479
.05	19.78	66.38	67	4.404	4.179	1.054	.225
.01	19.95	92.93	93	1.172	1.117	1.049	.055
.005	19.98	105.49	106	.647	.622	1.040	.025
.001	19.99	135.85	136	.160	.155	1.032	.005

Table 4: The True Number of Classes, $k = 30$

c	\hat{k}	$E(N)$	n^*	$Risk(N)$	$Risk(n^*)$	Risk Eff	Regret
.10	29.46	84.24	86	11.408	10.964	1.040	.444
.05	29.78	99.81	100	6.561	6.312	1.039	.249
.01	29.94	140.44	140	1.740	1.687	1.031	.053
.005	29.97	159.14	159	.966	.940	1.028	.026
.001	30.33	205.79	205	.237	.234	1.013	.003

From Tables 1-4, we infer that the *almost unbiased* estimate \hat{k}_n converges to the corresponding true number of classes k as $c \rightarrow 0$ for all values of k . The risk efficiency gets close to one and the regret goes to zero as $c \rightarrow 0$. This provides a substantial amount of numerical evidence, for concluding that the proposed sequential estimator \hat{k}_N performs very well.

We also see from the above tables that $E(N)$ increases as the sampling cost per observation, c , becomes smaller. However, the average risk under stopping time N decreases dramatically as $c \rightarrow 0$. Recall that the loss function we have assumed, incorporates both losses from estimation and costs from sampling. In this context, the value of c plays a role as an inflating-sample factor in the sequential procedure.

4.2 Comparison of Sequential Procedures based on \hat{k}_n and K_n

We note that \hat{k}_n converges to k faster than K_n . Overall, we also observe that the sequential procedure based on \hat{k}_n is uniformly better than the one based on K_n with respect to both criteria: namely, risk-efficiency and regret (see Cho and Govindarajulu (3)). Also, \hat{k}_N outperforms K_N with respect to the average stopping time for most cases except when k is relatively small. For example, when $k = 15$ and $c = 0.05$, the sequential procedure based on \hat{k}_n requires 50 on average for the number of observations, whereas the one based on K_n requires 54. Summary of detailed comparison between two sequential procedures is made in Table 5. This leads us to conclude that the proposed procedure achieves reduction in risk as well as in stopping time except for small k .

Table 5: Comparison of Two Sequential Procedures when $c = 0.05$

k	Estimator	\hat{k}	$E(N)$	n^*	R_N	R_{n^*}	Risk Eff	Regret
5	\hat{k}_n	4.83	16.18	17	1.240	.996	1.250	.244
	K_n	4.45	14.23	16	1.906	1.010	1.887	.896
10	\hat{k}_n	9.79	32.84	34	2.405	2.056	1.170	.349
	K_n	9.40	32.66	34	3.525	2.102	1.677	1.383
20	\hat{k}_n	19.78	66.38	67	4.404	4.179	1.054	.225
	K_n	19.46	76.45	77	5.206	4.447	1.171	.759

Note: \hat{k} = Estimate of the number of cells.

4.3 Approximation of the Risk and Examples

When $k \geq 25$, $\hat{\alpha}_1 \sim -1$, $\hat{\alpha}_2 \sim -2$, and $\hat{\theta}_1 \sim -\log c$, the optimal fixed-sample size required when k is known (see (1.29)) becomes

$$n^* = k \left\{ \hat{\theta}_1 - e^{\log c} \left[-2\hat{\theta}_1 - 2 \left(1 + \hat{\theta}_1 \right) - 2(k-1) + 2k \right] \right\} \\ \doteq k \{ -\log c - 4c \log c \}. \quad (4.1)$$

The risk $R_{n^*}(c)$ from (1.30) assumes the simpler form (after setting $\alpha_1 = -1$ and $\alpha_2 = -2$)

$$R_{n^*}(c)/k = [1 + (1 + \theta) e^{-\theta}]^2 [e^{-\theta} + (k-1) e^{-2\theta}] \\ + k [e^{-2\theta} - 2 \{1 + (1 + \theta) e^{-\theta}\} e^{-2\theta}] + c\theta \quad (4.2)$$

which simplifies (after some algebra) to

$$R_{n^*}(c)/k = e^{-\theta} \left[1 + 2(1 + \theta) e^{-\theta} + (1 + \theta)^2 e^{-2\theta} \right] - e^{-2\theta} \\ - 2(1 + \theta) e^{-3\theta} + (k-1)(1 + \theta)^2 e^{-4\theta} + c\theta. \quad (4.3)$$

Letting $y = e^{-n^*/k} = e^{(1+4c) \log c} \doteq c^{(1+4c)} \approx c$ and $\theta = n^*/k$, we obtain (after ignoring c^3 and higher order terms)

$$R_{n^*}(c)/k = c(1+c) - c \log c = c(1+c - \log c). \quad (4.4)$$

In the following two examples, we will compare the n^* and $R_{n^*}(c)$ given by (4.1) and (4.4) respectively with the true values of n^* and $R_{n^*}(c)$ given by (1.29) and (1.20).

Example 4.1 Suppose that $k = 20$, $c = 0.005$. Then, from (4.1) and (4.4) we have $n^* = 108.09$ and the corresponding risk $R_{n^*} = 0.630$, whereas the average values based on 5000 replications are $n = 105.49$ and $R_n = 0.622$ respectively. So the values obtained using the approximation are in good agreement with the true values.

Example 4.2 Suppose that $k = 30$, $c = 0.01$. Then, using (4.1) and (4.4) we obtain $n^* = 143.68$ and the corresponding risk $R_{n^*} = 1.685$, whereas the average values based on 5000 replications are $n = 140.44$ and $R_n = 1.687$ respectively. So the values obtained using the approximation are fairly close to the true values.

Hence, we recommend the simpler adaptive stopping time given by (1.32) when we suspect that k is large, say $k \geq 25$.

References

- [1] Arnold, B.; Beaver, R. Estimation of the number of classes in a population. *Biometrical Journal*, 1988, 30, 413-424.
- [2] Bunge, J.; Fitzpatrick, M. Estimating the number of species: a review. *Journal of the American Statistical Association*, 1993, 88, 364-373.
- [3] Cho, H.; Govindarajulu, Z. Sequential estimation of number of multinomial cells. *American Journal of Mathematical and Management Sciences*, 2002, 22, 15-30.
- [4] Chow, Y. S.; Martinsek A. T. Bounded regret of a sequential procedure for estimation of the mean. *The Annals of Statistics*, 1982, 10, 909-914.
- [5] Chow, Y. S.; Robbins, H. On the asymptotic theory of fixed width sequential confidence intervals for the mean. *The Annals of Mathematical Statistics*, 1965, 36, 457-462.
- [6] Finkelstein, M.; Tucker, H.; Veeh, J. Confidence intervals for the number of unseen types. *Statistics and Probability Letters*, 1998, 37, 423-430.

- [7] Goodman, L. On the estimation of the number of classes in a population. The Annals of Mathematical Statistics, 1949, 20, 572-579.
- [8] Goodman, L. Sequential sampling tagging for population size problem. The Annals of Mathematical Statistics, 1953, 24, 56-69.
- [9] Govindarajulu, Z. The Elements of Sampling Theory and Methods. Prentice-Hall, Inc., New Jersey, 1999.
- [10] Harris, B. Statistical inference in the classical occupancy problem unbiased estimation of the number of classes. Journal of the American Statistical Association, 1968, 63, 837-847.
- [11] Weiss, I. Limiting Distributions in some Occupancy Problems. The Annals of Mathematical Statistics, 1958, 29, 878-884.