

# SEQUENTIAL CONFIDENCE INTERVALS FOR THE NUMBER OF CELLS IN A MULTINOMIAL POPULATION

Z. Govindarajulu and Hokwon Cho

Department of Statistics  
University of Kentucky, Lexington, KY 40506

and

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

## Abstract

A two-sided sequential confidence interval is suggested for the number of equally probable cells in a given multinomial population with prescribed width and confidence coefficient. We establish large-sample properties of the fixed-width confidence interval procedure using a normal approximation, and some comparisons are made. Also a simulation study is carried out in order to investigate the finite sample behavior of the suggested sequential interval estimation procedure.

*Key Words:* Two-sided sequential confidence intervals; number of cells; multinomial population; normal approximation.

## 1 Introduction

Suppose we have a multinomial population with equally probable but unknown number of distinct cells (or categories)  $k$  ( $0 < k < \infty$ ). Based on a random sample of size  $n$  drawn from the population, we wish to construct a two-sided sequential confidence interval for the true number of cells.

Several practical examples fitting this model have been given in the literature. We mention a few in the following.

- (a) A problem of some archeological interest is determining how many days there were in the calendar of some ancient civilization. By noting the days marked on gravestones and treating these days as random samples from the total population of  $k$  days in the annual calendar, we might estimate  $k$  by visiting more graves and making note of days which occur more than once. (See Goodman (1953, p. 57)).
- (b) A company has offered to give away free samples of its product. A large number of requests, say  $N$  ( $N = 100,000$ ) come in and these requests have been filled. It is known that the same people often sent more than one request. From a sample of the requests, say  $n$ , we wish to estimate the distinct number of people  $k$  who have sent in the requests. It is reasonable to assume that a request made by a certain individual is  $1/k$ . (See Mosteller (1949, p. 12)).
- (c) We wish to estimate the total number of dies used to manufacture coins based on a sample (a coin hoard) of size  $n$  which has been classified according to the die used to produce each coin.

Suppose we assume that each die has been used to strike approximately the same number of coins. We can view this as drawing a random sample of size  $n$  from a population of  $k$  (the number of dies) equally likely classes. (See Arnold and Beaver, (1988, p. 413)).

In fact, estimating the number of cells in a given multinomial population can be thought of as the classical occupancy problem in which we draw a sample of size  $n$  with replacement from a population of  $k$  distinct objects. Therefore, the basic problem turns out to be equivalent to the problem of distributing  $n$  balls into  $k$  distinct bins.

Confidence interval estimation of  $k$  has been considered by Ivechenko and Timonina (1983), Arnold and Beaver (1988), and Finkelstein, Tucker and Veeh (1998) based on a fixed sample size. Here we will consider fixed-width sequential confidence interval estimation.

## 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 construct a confidence interval for the true unknown number of cells,  $k$  ( $< \infty$ ) using  $\hat{k}_n$  which is an estimate of the number of distinct cells in  $n$  observations.

Therefore we wish to set up a two-sided confidence interval for  $k$  based on  $\hat{k}_n$ . That is,

$$P \left\{ \left| \hat{k}_n - k \right| \leq d \right\} \geq 1 - \alpha,$$

where  $1 - \alpha$  ( $0 < \alpha < 1$ ) denotes the level of confidence, for specified  $d$ .  $K_n$ , the number of distinct cells in a random sample of size  $n$ , is a biased estimate of  $k$  since  $K_n \leq k$ . However, the exact probability distribution of the number of unseen cells,  $k - K_n$ , is given in Feller (1968, p. 102) (see also Jordan (1950, p. 178)):

$$P(k - K_n = m) = \binom{k}{m} \sum_{v=0}^{k-m} (-1)^v \binom{k-m}{v} \left(1 - \frac{m+v}{k}\right)^n.$$

## 1.2 Almost Unbiased Estimator for $k$

Goodman (1953) and Harris (1968) consider the best unbiased estimation of  $k$ . Since  $K_n$  is sufficient and complete for  $k$ , Harris (1968) has shown that the unique unbiased estimator of  $k$ , can be expressed as 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

$$\hat{k}_n = (n+1)/R_1, \tag{1.1}$$

where  $R_1$  is the solution of the transcendental equation

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

Letting

$$x = (n+1)/k \quad \text{and} \quad R = x(1 + \varepsilon),$$

we have

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

or

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

Thus,

$$R_1 \doteq x(1 - e^{-x}).$$

Hence,  $\hat{k}_n$  given in (1.1) becomes

$$\begin{aligned}
\hat{k}_n &= \left( \frac{n+1}{R_1} \right) (1 - e^{R_1}) \\
&\doteq (n+1)/R_1 \\
&= \frac{n+1}{[(n+1)/K_n] [1 - e^{-(n+1)/K_n}]} \\
&= \frac{K_n}{1 - e^{-(n+1)/K_n}} \\
&\doteq K_n [1 + e^{-(n+1)/K_n}].
\end{aligned} \tag{1.2}$$

We want to determine  $n$  such that

$$P \left\{ \left| \hat{k}_n - k \right| \leq d \right\} \geq 1 - \alpha,$$

for specified  $d (> 0)$  and  $\alpha$ .

### 1.3 Probability Distribution of $k - K_n$ using Stirling Numbers

The distribution of the number of unseen cells,  $k - K_n$ , using Stirling numbers of the second kind was given by Jordan (1950). By the generalized Bernoulli theorem of repeated trials,

$$P(k - K_n = j) = \binom{k}{j} k^{-n} \sum_{s_1 + s_2 + \dots + s_{k-j} = n} \frac{n!}{s_1! s_2! \dots s_{k-j}!}, \quad 0 \leq j \leq k-1.$$

where the  $s_i$ 's denote the  $i$ -th cell frequencies.

$\mathcal{S}_n^m$ , the Stirling numbers of the second kind are defined by

$$\mathcal{S}_n^m = \frac{n!}{m!} \sum_{s_1! s_2! \dots s_m!} \frac{1}{s_1! s_2! \dots s_m!},$$

where the summation is all over  $s_i$  such that  $s_1 + s_2 + \dots + s_m = n$ .

Then, we have

$$\begin{aligned}
P(k - K_n = j) &= \binom{k}{j} k^{-n} (k-j)! \mathcal{S}_n^{k-j} \\
&= (k!/j!) k^{-n} \mathcal{S}_n^{k-j}.
\end{aligned}$$

For large  $n$ , Jordan (1950, p. 173) has shown that

$$\mathcal{S}_n^m \sim m^n / m!,$$

which is valid for  $m$  quite small relative to  $n$ .

## 2 Approximations to the Distribution of the Number of Unseen Cells

### 2.1 Asymptotic normality of the Distribution of $U_n = k - K_n$

Weiss (1958) established the asymptotic normality of the number of unseen cells,  $U_n = k - K_n$ , by showing that all the central moments of  $U_n$  converge to the moments of the standard normal

distribution. In particular, he has shown that if  $n$  and  $k$  are sufficiently large (independently of one another) such that  $n/k \rightarrow a$  ( $a > 0$ ) then,  $U_n$  is asymptotically normal with mean

$$\mu_n = k(1 - 1/k)^n \quad (2.1)$$

and variance

$$\sigma_n^2 = k(1 - 1/k)^n - k^2(1 - 1/k)^{2n} + k(k-1)(1 - 2/k)^n. \quad (2.2)$$

Let

$$\alpha_1 = k \log(1 - 1/k) \doteq -1 - (1/2k)$$

and

$$\alpha_2 = k \log(1 - 2/k) \doteq -2 - (1/k).$$

Then, using the approximate values for  $\alpha_1$  and  $\alpha_2$ , (2.1) and (2.2) can be written as

$$\begin{aligned} \mu_n = E(U_n) &= ke^{\alpha_1 n/k} \doteq ke^{-n/k} e^{-n/2k^2} \\ &\doteq ke^{-n/k} (1 - n/2k^2) \\ &= ky + (y/2) \ln y, \end{aligned} \quad (2.3)$$

where  $y = e^{-n/k}$ .

Similarly we obtain

$$\begin{aligned} \sigma_n^2 = \text{var}(U_n) &= ke^{\alpha_1 n/k} \left[ 1 - ke^{\alpha_1 n/k} + (k-1)e^{(\alpha_2 - \alpha_1)n/k} \right] \\ &\doteq ky \left( e^{-n/2k^2} - ye^{-n/k^2} \right) \\ &\doteq ky \left[ 1 - (n/2k^2) \right] - ky^2 \left[ 1 - (n/k^2)n \right] \\ &= ky(1 - y) + (y/2) \ln y - y^2 \ln y. \end{aligned} \quad (2.4)$$

## 2.2 Normal Approximation to $U_n = k - K_n$

Recall that the goal is to determine  $n$  such that

$$P \left\{ \left| \hat{k}_n - k \right| \leq d \right\} \geq 1 - \alpha,$$

for specified  $d$  ( $> 0$ ) and  $\alpha$ .

First, we establish the asymptotic distribution of  $\hat{k}_n - k$ , where  $\hat{k}_n = K_n(1 + e^{-n/K_n})$ , which is approximately the best unbiased estimator for  $k$ .

Letting

$$f(x) = x \left( 1 + e^{-n/x} \right),$$

we have

$$\begin{aligned} f'(k) &= 1 + e^{-n/k} + (n/k) e^{-n/k} \\ &= 1 + y - y \log y > 0 \end{aligned}$$

where  $y = e^{-n/k}$ .

Then, we are led to the following result.

**Result 2.1** *Let  $f(x) = x(1 + e^{-n/x})$  for  $x > 0$  and let  $K_n$  denote the number of distinct cells that have been observed in a random sample of size  $n$  from a multinomial distribution having  $k$  equally probable cells. Let  $\hat{k}_n = f(K_n)$  and  $U_n = k - K_n$ . Then*

$$k - \hat{k}_n \stackrel{d}{\simeq} N \left( B, [f'(k) \sigma_n]^2 \right)$$

where

$$B = \mu_n f'(k) - ke^{-n/k}. \quad (2.5)$$

**Proof.** One can write

$$\begin{aligned} k - \hat{k}_n &= k - f(k) + [f(k) - f(K_n)] \\ &\doteq -ke^{-n/k} + U_n f'(k) \\ &= (U_n - \mu_n) f'(k) + \mu_n f'(k) - ke^{-n/k}. \end{aligned} \quad (2.6)$$

It follows from (2.6) that

$$\frac{k - \hat{k}_n}{f'(k) \sigma_n} = \frac{U_n - \mu_n}{\sigma_n} + \frac{B}{f'(k) \sigma_n}.$$

That is,

$$\frac{k - \hat{k}_n - B}{f'(k) \sigma_n} = \frac{U_n - \mu_n}{\sigma_n}$$

Hence, we have

$$P \left\{ \frac{k - \hat{k}_n - B}{f'(k) \sigma_n} \leq t \right\} = P \left\{ \frac{U_n - \mu_n}{\sigma_n} \leq t \right\} \doteq \Phi(t)$$

after using the result of Weiss (1958) pertaining to the asymptotic normality of  $U_n$  when suitably standardized, and  $\Phi$  denotes the standard normal distribution function. ■

Thus, from Result 2.1 we have

$$\hat{k}_n \stackrel{d}{\simeq} N \left( k - B, [f'(k) \sigma_n]^2 \right). \quad (2.7)$$

Now, it follows from (2.5)

$$\begin{aligned} B_n &= \mu_n f'(k) - ke^{-n/k} \\ &= [ky + (y/2) \ln y] (1 + y - \ln y) - ky \\ &= ky^2 (1 - \ln y) + [(y/2) \ln y] (1 + y - \ln y) \\ &\doteq ky^2 (1 - \ln y) + (y/2) \ln y \\ &\doteq ky^2 (1 - \ln y). \end{aligned}$$

Note that the bias in  $\hat{k}_n$  is  $-B$  which is small when  $y (= e^{-n/k})$  is small.

Further, assuming that  $y$  is small, the variance of  $\hat{k}_n$  is

$$\begin{aligned} [f'(k) \sigma_n]^2 &= [ky(1-y) + (y/2) \ln y - y^2 \ln y]^2 (1 + y - y \ln y)^2 \\ &\doteq (1 + y - y \ln y)^2 [ky(1-y) + (y/2) \ln y] \\ &\doteq ky(1 + 2y - 2y \ln y) [1 - y + (\ln y)/2k] \\ &= ky \left[ 1 + y + (\ln y)/2k - y(\ln y)^2/k \right] \end{aligned}$$

after ignoring inferior terms.

Using Result (2.1), we can write

$$\begin{aligned}
P \left\{ \left| \hat{k}_n - k \right| \leq d \right\} &= P \left\{ -d \leq \hat{k}_n - k \leq d \right\} \\
&= P \left\{ -d + B \leq (\hat{k}_n - k) + B \leq d + B \right\} \\
&= P \left\{ \frac{-d + B}{f'(k) \sigma_n} \leq \frac{(\hat{k}_n - k) + B}{f'(k) \sigma_n} \leq \frac{d + B}{f'(k) \sigma_n} \right\} \\
&\doteq \Phi \left[ \frac{d + B}{f'(k) \sigma_n} \right] - \Phi \left[ \frac{-d + B}{f'(k) \sigma_n} \right] \\
&\geq 2\Phi \left[ \frac{d - B}{f'(k) \sigma_n} \right] - 1 \\
&\geq 1 - \alpha,
\end{aligned}$$

which implies

$$\frac{d - B}{f'(k) \sigma_n} \geq z_{\alpha/2} = \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right),$$

where  $z_{\alpha/2}$  is the upper  $(\alpha/2)$ th quantile of the  $\Phi$ .

Consider the inequality (where we write  $z = z_{\alpha/2}$ )

$$d - B > z f'(k) \sigma_n,$$

which implies

$$(d - B)^2 > z^2 [f'(k) \sigma_n]^2. \text{ (Since } d > B.)$$

Thus

$$\begin{aligned}
&d^2 - 2dB > z^2 k y [1 + y + (\ln y) / 2k] \\
\Leftrightarrow &d^2 > z^2 k y [1 + y + (\ln y) / 2k] + 2dk y^2 (1 - \ln y).
\end{aligned}$$

First, solve the inequality  $d^2 > z^2 k y$ , i.e.,

$$y < d^2 / z^2 k \equiv y_0 \text{ (say).}$$

Let

$$g(y) = d^2 - z^2 k y + z^2 k y [y + (\ln y) / k] - 2dk y^2 (1 - \ln y).$$

Then, Newton-Raphson method gives the next approximation  $y_1$  as

$$y_1 = y_0 - [g(y_0) / g'(y_0)],$$

where

$$g(y_0) = z^2 k y_0 [y_0 + (\ln y_0) / k] - 2dk y_0^2 (1 - \ln y_0)$$

and

$$g'(y) = -z^2 k + z^2 k [2y + (\ln y) / k + (1/k)] - 2dk y (1 - 2 \ln y).$$

Also

$$\ln y_0 \doteq \ln (d^2 / z^2 k) - \ln k = 2 \ln (d/z) - \ln k$$

and

$$y_0 \ln y_0 = (d^2 / z^2 k) [\ln (d^2 / z^2 k) - \ln k].$$

Hence

$$\begin{aligned}
g(y_0) &= d^2 \left[ \frac{d^2}{z^2 k} + \frac{2 \ln(d/z)}{k} - \frac{\ln k}{k} \right] - 2dk \left( \frac{d^2}{z^2 k} \right)^2 \left[ 1 - 2 \ln \left( \frac{d}{z} \right) + \ln k \right] \\
&= \frac{d^2}{k} \left[ \frac{d^2}{z^2} + 2 \ln \left( \frac{d}{z} \right) - \ln k \right] - \frac{2d^5}{z^2 k} \left[ 1 - 2 \ln \left( \frac{d}{z} \right) + \ln k \right] \\
&= \frac{d^4}{z^2 k} (1 - 2d) + \frac{2d^2}{k} \ln \left( \frac{d}{z} \right) \left( 1 + \frac{2d^3}{z^2} \right) - \frac{d^2}{k} (\ln k) \left( 1 + \frac{2d^3}{z^2} \right)
\end{aligned}$$

and

$$\begin{aligned}
g'(y_0) &= -z^2 k + z^2 k \left[ \frac{2d^2}{z^2 k} + \frac{1 + 2 \ln(d/z) - \ln k}{k} \right] - 2d \frac{d^2}{z^2} \left[ 1 - 4 \ln \left( \frac{d}{z} \right) + 2 \ln k \right] \\
&= 2d^2 + z^2 k \left[ 2 \ln \left( \frac{d}{z} \right) - \ln k \right] - \frac{2d^3}{z^2} \left[ 1 - 4 \ln \left( \frac{d}{z} \right) + 2 \ln k \right] \\
&\doteq z^2 k \left[ 2 \ln \left( \frac{d}{z} \right) - \ln k \right] + 2d^2 + \frac{4d^3}{z^2} \left[ 2 \ln \left( \frac{d}{z} \right) - \ln k \right] \\
&\doteq \left( z^2 k + \frac{4d^3}{z^2} \right) \left[ 2 \ln \left( \frac{d}{z} \right) - \ln k \right] + 2d^2 \\
&\doteq z^2 k [2 \ln(d/z) - \ln k] + 2d^2.
\end{aligned}$$

Thus

$$\begin{aligned}
g(y_0) &\doteq (2d^2/k) \ln(d/z) - (d^2/k) \ln k \\
&= (d^2/k) [2 \ln(d/z) - \ln k].
\end{aligned}$$

Hence

$$\begin{aligned}
y_1 &= y_0 - [g(y_0) / g'(y_1)] \\
&= \frac{d^2}{z^2 k} - \frac{(d^2/k) [2 \ln(d/z) - \ln k]}{z^2 k [2 \ln(d/z) - \ln k] + 2d^2} \\
&\doteq \frac{d^2}{z^2 k} - \frac{d^2}{z^2 k^2} \\
&= (d/zk)^2 (k - 1).
\end{aligned}$$

Since  $y_1 = e^{-n/k}$ ,

$$e^{-n/k} \leq (d/zk)^2 (k - 1). \quad (2.8)$$

Taking logarithms on both sides of (2.8) we have,

$$n \geq k [2 \ln(zk/d) - \ln(k - 1)] \quad (2.9)$$

Hence, we take the optimal fixed-sample size  $n^*$  to be the smallest positive integer satisfying (2.9). Since the true number of cells  $k$  is unknown, no fixed-sample size procedure is available. Therefore the adaptive sequential rule is; Stop at  $N$  where

$$N = \inf \left\{ n \geq n_0 : n \geq \hat{k}_n \left[ 2 \ln \left( z \hat{k}_n / d \right) - \ln \left( \hat{k}_n - 1 \right) \right] \right\} \quad (2.10)$$

where  $n_0$  ( $n_0 \geq 2$ ) is the initial sample size,  $\hat{k}_n$  is given by (1.2).

**Example 2.1** (a) A computer algorithm was used to generate multinomial data for  $k = 9$ . We wish to construct a 95% confidence interval for the true number of cells,  $k$  with  $d = 1.0$ . The observations are

$$\begin{aligned} &7, 1, 4, 9, 7, 5, 6, 1, 9, 5, 7, 7, 1, 6, 1, 7, 2, 9, 3, 8 \\ &4, 1, 3, 6, 3, 9, 6, 9, 9, 4, 1, 4, 6, 8 \end{aligned}$$

where the first number 7, for instance, indicates that it belonged to seventh cell. The proposed sequential procedure (starts with  $n_0 = 3$ ) stops at  $n = 34$  yielding  $K_n = 9$ , the estimated the number of cells,  $\hat{k}_n = 9.1842$ ; so the 95% confidence interval for  $k$  is given by  $(8.1842, 10.1842)$ .

(b) We also generated data for  $k = 20$ , and suppose we wish to set up a 90% confidence interval with  $d = 2.0$ . Our sequential procedure stops at  $n = 50$  yielding  $K_n = 18$ , and  $\hat{k} = 19.1192$ . Therefore, the resultant 90% confidence interval is  $(17.1192, 21.1192)$ .

### 3 Asymptotics for the Procedure

In this section, we investigate the asymptotic behavior of the proposed sequential procedure.

#### 3.1 Finite Sure Termination

We now show the fundamental property that the proposed stopping rule terminates finitely almost surely.

**Theorem 3.1** *Let  $N$  denote the stopping time associated with the sequential procedure, then  $\lim_{n \rightarrow \infty} P(N = \infty) = 0$ .*

**Proof.** For the stopping rule given by (2.10) we have after using (1.2),

$$P(N > n) \leq P \left\{ n \leq K_n \left( 1 + e^{-(n+1)/K_n} \right) \left[ 2 \ln \left( z K_n \left( 1 + e^{-(n+1)/K_n} \right) / d \right) - \ln \left( K_n \left( 1 + e^{-(n+1)/K_n} \right) - 1 \right) \right] \right\} \quad (3.1)$$

Taking limit on both sides in (3.1), we obtain

$$\begin{aligned} \lim_{n \rightarrow \infty} P(N > n) &\leq \lim_{n \rightarrow \infty} P \left\{ n \leq K_n \left( 1 + e^{-(n+1)/K_n} \right) \left[ 2 \ln \left\{ z K_n \left( 1 + e^{-(n+1)/K_n} \right) / d \right\} \right. \right. \\ &\quad \left. \left. - \ln \left\{ K_n \left( 1 + e^{-(n+1)/K_n} \right) - 1 \right\} \right] \right\} \\ &= 0 \end{aligned}$$

due to the fact that the sequence of estimators  $K_n$  is bounded above by  $k$  ( $k < \infty$ ) as  $n \rightarrow \infty$ . This establishes the finite sure termination of the proposed sequential procedure. ■

#### 3.2 First Order Asymptotics

The stopping rule given in (2.10) can be rewritten as

$$N = \inf \left\{ n \geq n_0 : \frac{n}{k} \geq \frac{\hat{k}_n}{k} \left[ 2 \ln \left( z \hat{k}_n / d \right) - \ln \left( \hat{k}_n - 1 \right) \right] \right\}. \quad (3.2)$$

Now, rule (3.2) is of the form

$$N = N(t) = [\min n \geq n_0 : Y_n \leq f(n)/t].$$

where

$$\begin{aligned} f(n) &= n, \\ t &= k [2 \ln (zk/d) - \ln (k-1)], \end{aligned}$$

and

$$Y_n = \frac{\hat{k}_n \left[ 2 \ln \left( z\hat{k}_n/d \right) - \ln \left( \hat{k}_n - 1 \right) \right]}{k \left[ 2 \ln (zk/d) - \ln (k-1) \right]}.$$

Then clearly,  $Y_n$  is a sequence of random variables such that  $Y_n > 0$ ,  $\lim_{d \rightarrow 0} Y_n = 1$  almost surely (a.s.) since  $\hat{k}_n/k \rightarrow 1$  as  $n \rightarrow \infty$ . Also  $\lim_{n \rightarrow \infty} f(n) = \infty$ ,  $\lim_{n \rightarrow \infty} f(n)/f(n-1) = 1$ .

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 asymptotics for the proposed sequential procedure.

**Result 3.1**

- (i)  $\lim_{d \rightarrow 0} N = \infty$  a.s.,  $\lim_{d \rightarrow 0} E(N) = \infty$
- (ii)  $\lim_{d \rightarrow 0} N/n^* = 1$  a.s.,  $\lim_{d \rightarrow 0} E(N)/n^* = 1$
- (iii)  $\lim_{d \rightarrow 0} P \left( \hat{k}_N - d \leq k \leq \hat{k}_N + d \right) = 1 - \alpha$

where  $n^*$  is the optimal fixed-sample size given by (2.9).

**Proof.** (i) is easily verified. By the definition of the stopping rule in (2.10) and  $N(d)$  increases a.s. in  $d$ , we obtain  $N = N(d) \rightarrow \infty$  as  $d \rightarrow 0$ . Then by the monotone convergence theorem, we get  $E(N) \rightarrow \infty$  as  $d \rightarrow 0$ .

(ii) follows from the facts that for  $n^*$  given by (2.9),  $\lim_{d \rightarrow 0} N/n^* = 1$  a.s., and  $\lim_{d \rightarrow 0} E(N)/n^* = 1$ , since  $\hat{k}_n = K_n(1 + e^{-(n+1)/K_n})$ ,  $\sup_n \hat{k}_n \leq k(1 + e^{-1/K_n}) < 2k$ , and

$$E \left[ \sup_n \hat{k}_n \right] \leq E \left[ K_n \left( 1 + e^{-1/K_n} \right) \right] \leq 2k < \infty.$$

(iii) follows from Anscombe (1952) theorem provided we can verify his condition on uniform continuity in probability of  $\hat{k}_n$ . Towards this consider, with  $Y_n = k - \hat{k}_n$  and  $U_n = k - K_n$ ,

$$\begin{aligned} &P \{ Y_{n'} - Y_n \leq \text{for all } n < n' \leq (1+c)n \} \\ &= P \{ Y_{n'} = Y_n \text{ for all } n < n' \leq (1+c)n \}, \text{ for sufficiently small } \varepsilon \\ &= P \{ f(K_{n'}) = f(K_n) \text{ for all } n < n' \leq (1+c)n \} \\ &= P \{ K_{n'} = K_n \text{ for all } n < n' \leq (1+c)n \} \\ &= P \{ \text{no new cells are discovered in trials } n' = n+1, (1+c)n \} \\ &= P \{ U_{n'} = U_n \text{ for all } n < n' \leq (1+c)n \} \\ &= \sum_{j=0}^{k-1} P \{ K_{n'} = j \text{ for all } n < n' \leq (1+c)n | U_n = j \} P(U_n = j) \\ &= \sum_{j=0}^{k-1} \left( \frac{k-j}{k} \right)^{[cn]} P(U_n = j) \\ &\geq \sum_{j=0}^{k-1} \left( 1 - \frac{j}{k} \right)^{cn} P(U_n = j) \\ &= \sum_{j=0}^{k-1} \exp(-jcn/k) P(U_n = j), \text{ for sufficiently large } k \\ &= E \left[ \exp \left( \frac{-cn}{k} U_n \right) \right] \end{aligned}$$

$$\begin{aligned}
&\geq \exp \left[ \frac{-cn}{k} E(U_n) \right], \text{ by Jensen's inequality} \\
&= \exp \left[ -cn \{ e^{-n/k} + o(1) \} \right] \\
&\geq 1 - cn \exp(-n/k) \\
&= 1 - \eta
\end{aligned}$$

where  $\eta$  is small when  $c$  is small and  $n = k^{1+\delta}$ , for some  $0 < \delta < 1$ .

An analogous inequality holds for  $(1-c)n \leq n' < n$ . Thus Anscombe's condition holds. ■

## 4 Relation Between Equally Probable Multinomial Cell Model and the Efron and Thisted Model

Efron and Thisted (1976) considered the problem of estimating the number of words in Shakespeare's vocabulary. They realized that this was equivalent to estimating the number of unseen species. They assumed that members of each species enter a trap according to a Poisson process, the process for species  $s$  having expectation  $\lambda_s$  per unit of time.

Assuming that  $(\lambda_1, \lambda_2, \dots, \lambda_k)$  is a random sample from an unknown distribution  $G(\lambda)$ , they obtained an empirical Bayes estimate of the number of unseen species or equivalently the number of words Shakespeare knew. In the following we will show that the model used by Efron and Thisted (1976) is equivalent to an equally probable number of cells model when the number of cells  $k$  is large.

It is well known that if  $Y_1, Y_2, \dots, Y_n$  are independent Poisson random variables with means  $\lambda_1, \lambda_2, \dots, \lambda_k$ , then the conditional distribution of  $Y_1, Y_2, \dots, Y_n$  for given  $Y_1 + Y_2 + \dots + Y_n$  is multinomial distribution,  $M(n, p_i = \lambda_i / \sum_{i=1}^k \lambda_i)$ ,  $i = 1, 2, \dots, k$ , and conversely. Further, if  $(\lambda_1, \lambda_2, \dots, \lambda_k)$  is a random sample from  $G(\lambda)$ , having mean  $\mu$ , variance  $\sigma^2$  and a finite third moment, then

$$\begin{aligned}
E \left( \frac{\lambda_i}{\sum_{j=1}^k \lambda_j} \right) &= E \left[ \frac{\mu + (\lambda_i - \mu)}{k\mu + \sum_{j=1}^k (\lambda_j - \mu)} \right] \\
&= E \left\{ \left[ \frac{1}{k} + \frac{\lambda_i - \mu}{k\mu} \right] \left[ 1 + \sum_{j=1}^k \frac{(\lambda_j - \mu)}{k\mu} \right]^{-1} \right\} \\
&= \frac{1}{k} E \left\{ \left[ 1 + \frac{\lambda_i - \mu}{\mu} \right] \left[ 1 + \frac{1}{k} \sum_{j=1}^k \frac{(\lambda_j - \mu)}{\mu} \right]^{-1} \right\} \\
&= \frac{1}{k} E \left\{ \left[ 1 + \frac{\lambda_i - \mu}{\mu} \right] \left[ 1 - \frac{1}{k} \sum_{j=1}^k \frac{(\lambda_j - \mu)}{\mu} + o(k^{-1}) \right] \right\} \\
&= \frac{1}{k} E \left\{ 1 - \frac{1}{k} E \left[ \frac{(\lambda_i - \mu)^2}{\mu} \right] + \frac{1}{k} \frac{\sigma^2}{\mu^2} + o(k^{-1}) \right\} \\
&= \frac{1}{k} \left\{ 1 - \frac{\sigma^2}{k\mu^2} + \frac{1}{k} \frac{\sigma^2}{\mu^2} + o(k^{-1}) \right\} \\
&= 1/k - O(k^{-2}).
\end{aligned}$$

## 5 Simulation Studies

### 5.1 Monte Carlo Experimentation

Monte Carlo experimentation is carried out in order to illustrate the behavior and the performance of the stopping rule in the proposed sequential procedure. The results of the Monte Carlo simulation are summarized in the following tables, which show the average estimate  $\hat{k}$ , specified size of the error  $d$ , the average stopping time  $E(N)$ , and the average observed coverage probability (CP) in the experiments. Each row in the table corresponds to 5,000 independent experiments with the initial sample size  $n_0 = 3$  used.

Table 5.1  $k = 5$

$1 - \alpha$	90%			95%			99%		
$d$	$\hat{k}$	$E(N)$	CP	$\hat{k}$	$E(N)$	CP	$\hat{k}$	$E(N)$	CP
.1	4.99	37.96	.99	4.99	38.96	.99	5.00	41.97	.99
.2	4.98	30.85	.98	4.98	32.86	.98	4.99	34.94	.99
.5	4.89	21.21	.88	4.92	23.34	.91	4.96	25.65	.95
1.0	4.60	13.12	.84	4.67	15.19	.89	4.84	18.89	.94

Table 5.2  $k = 10$

$1 - \alpha$	90%			95%			99%		
$d$	$\hat{k}$	$E(N)$	CP	$\hat{k}$	$E(N)$	CP	$\hat{k}$	$E(N)$	CP
.1	9.99	80.93	.99	9.99	83.96	.99	10.00	89.97	.99
.2	9.98	66.78	.97	9.99	69.86	.98	9.99	75.93	.99
.5	9.94	48.35	.90	9.96	51.50	.92	9.97	57.64	.95
1.0	9.65	33.25	.88	9.80	37.18	.92	9.91	43.03	.96

Table 5.3  $k = 15$

$1 - \alpha$	90%			95%			99%		
$d$	$\hat{k}$	$E(N)$	CP	$\hat{k}$	$E(N)$	CP	$\hat{k}$	$E(N)$	CP
.1	15.00	125.98	.99	15.00	130.97	.99	15.00	139.98	.99
.2	14.99	104.87	.98	14.99	110.87	.98	15.00	118.92	.99
.5	14.96	77.32	.90	14.98	82.55	.93	14.98	91.69	.96
1.0	14.84	56.13	.91	14.88	61.35	.94	14.95	70.14	.97

Table 5.4  $k = 20$

$1 - \alpha$	90%			95%			99%		
$d$	$\hat{k}$	$E(N)$	CP	$\hat{k}$	$E(N)$	CP	$\hat{k}$	$E(N)$	CP
.1	19.99	172.97	.99	20.00	179.98	.99	20.00	190.98	.99
.2	19.99	145.84	.98	19.99	152.91	.99	20.00	163.95	.99
.5	19.97	109.25	.90	19.98	115.59	.93	19.99	126.70	.96
1.0	19.86	80.82	.91	19.94	88.38	.94	19.95	98.95	.97

From Tables 5.1-5.4, we infer that the estimates  $\hat{k}$  approach  $k$  and achieve the required level of confidence  $(1 - \alpha)$  as  $d$  decreases. It is also noticeable that the expected stopping time,  $E(N)$  increases consistently as the error size  $d$  decreases.

Thus, the numerical results indicate the small sample behavior and lend support to the asymptotic behavior of the proposed sequential procedure as the size of error,  $d$  goes to zero.

**Acknowledgements:** We thank the referee and Professor Olaf Bunke, Editor, for their helpful comments.

## References

- [1] F. Anscombe (1952) Large sample theory of sequential estimation. *Proc. Cambridge Philos. Soc.*, **48**, 600 - 607.
- [2] B. Arnold, R. Beaver (1988) Estimation of the number of classes in a population. *Biometrical Journal*, 30, 413-424.
- [3] D. Basu (1950) On sampling with and without replacement. *Sankhyā*, **20**, 287 - 294.
- [4] Y. Chow, H. Robbins (1965) On the asymptotic theory of fixed width sequential confidence intervals for the mean. *The Annals of Mathematical Statistics*, **36**, 457 - 462.
- [5] Y. Chow, H. Teicher (1988) *Probability Theory*, 2nd Ed. Springer-Verlag, New York.
- [6] B. Efron, R. Tibshirani (1976) Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, **63**, 435 - 447.
- [7] W. Feller (1968) *An Introduction to Probability Theory and its Applications* Vol. 1, 3rd Ed. John Wiley, New York.
- [8] M. Finkelstein, H. Tucker, J. Veeh (1998) Confidence intervals for the number of unseen types. *Statistics and Probability Letters*, **37**, 423 - 430.
- [9] L. Goodman (1949) On the estimation of the number of classes in a population. *The Annals of Mathematical Statistics*, **20**, 572 - 579.
- [10] L. Goodman (1953) Sequential sampling tagging for population size problems. *The Annals of Mathematical Statistics*, **24**, 56 - 69.
- [11] Z. Govindarajulu (1999) *The Elements of Sampling Theory and Methods*. Prentice-Hall, Inc. New Jersey.
- [12] B. Harris (1968) Statistical inference in the classical occupancy problem unbiased estimation of the number of classes. *Journal of the American Statistical Association*, **63**, 837 - 847.
- [13] G. Ivezhenko, E. Timonina (1983) Estimating the size of a finite population, *Theory of Probability and Its Applications*, **27**, 403 - 406.
- [14] C. Jordan (1950) *Calculus of Finite Differences*, 2nd Ed., Chelsea Publishing Co., New York.
- [15] L. Moser, M. Wyman (1958) Stirling numbers of the second kind, *Duke Mathematical Journal*, **25**, 29 - 44.
- [16] F. Mosteller (1949) Questions and answers - number of different kinds of elements in a population, *The American Statistician*, **3**, 12.
- [17] I. Weiss (1958) Limiting Distributions in some Occupancy Problems. *The Annals of Mathematical Statistics*, **29**, 878 - 884.
- [18] M. Woodroffe (1977) Second order approximation for sequential point and interval estimation. *The Annals of Statistics*, **5**, 984 - 995.