

Inverse-Type Sampling Procedure for Estimating the Number of Multinomial Classes

Hokwon A. Cho

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

ABSTRACT

A procedure with inverse-type sampling is proposed for the estimation of the number of classes (or cells) in a given multinomial distribution with a devised stopping rule that satisfies a preassigned P^* -condition and controls the probability of a correct decision $P\{CD\}$. Dirichlet Integrals of Type II are adapted for developing the procedure, which is based on a decision-theoretic approach. We assume that we know neither the number of classes (or cells) nor the cell probabilities (parameters) of the given multinomial but we do assume the minimum cell probability $\varepsilon (> 0)$. Finally, the Monte Carlo experimentation is carried out in order to illustrate the theoretical results and the behavior of proposed procedure.

Key Words: Inverse-type sampling; Number of classes in multinomial distribution; Probability of correct decision; Dirichlet Integrals Type II; Minimum cell probability.

1 INTRODUCTION

Suppose we have a multinomial population which is partitioned into an unknown number of classes, say k . Then a natural question will be to ask how many classes (or categories, cells) there are. The following examples are rather common in practical situations and have given in the literature.

- (1) Biologists, ecologists and botanists may be interested in estimating the number of different species in a population of fish, animals or plants. Such estimates are essential to obtain indices of diversity or the rates of extinction which play an important role in the study of ecosystems or in studying environmental conditions of the ecological diversity of species (Mann (1991)).
- (2) Numismatists may be concerned with estimating the total number of dies used to produce coins to study ancient economies. Suppose the researcher assumes that each die has been used to manufacture approximately the same number of coins (Marchand and Schroeck (1982), Arnold and Beaver (1988)).
- (3) Linguists may be interested in estimating the size of an author's vocabulary (McNeil (1973), Efron and Thisted (1976)).

In addition to these applications, estimating the number of distinct records in a filing system such as the social security card file can be an important issue, where many records are suspected to be duplicated.

An early paper in the above area was written by Goodman (1949). Since then, many approaches have been proposed and developed. Bunge and Fitzpatrick (1993) provided a comprehensive review of the existing literature according to methods and approaches with an extended bibliography.

Even though there has been more work done on these problems, both frequentist and Bayesian methods (Boender and Rinnooy Kan (1983)) use rather restrictive assumptions; namely, equal cell probabilities or uniform prior. Here, we consider a more flexible inverse-type sampling procedure, which has not been considered so far, using multiple decision-theoretic methods.

2 FORMULATION

Let X_1, X_2, \dots be a sequence of independent observations drawn from a multinomial population with an unknown integer number of classes (or cells) k , with corresponding cell probabilities p_i , where $p_i > 0$, $i = 1, 2, \dots, k$ and $\sum_{i=1}^k p_i = 1$. Then, there is a simple relation between the unknown k and the smallest and largest values, namely,

$$1/p_{[k]} \leq k \leq 1/p_{[1]}, \quad (2.1)$$

where $p_{[i]}$ represents the i -th ordered cell probability. For moderate and large numbers of observations, the number of cells with positive frequency in the sequential sample at stopping time will be an important part of our statistic for estimating the true number of cells k . However we need an appropriate stopping rule that will provide a confidence level on our estimate of k .

From a practical point of view, we fix a sufficiently small positive value ε and say that cells with probability less than this either do not exist or are not of any practical interest. We refer this condition as to the ε -condition as $\Omega(\varepsilon) \subset \Omega$ (the extended parameter space). From (2.1) it follows that $k \leq 1/\varepsilon$; thus if $\varepsilon = 0.1$, then $k \leq 10$ and if $\varepsilon = 0.05$, then $k \leq 20$.

2.1 Stopping Rule of the Procedure

We propose the following stopping and decision rule for our procedure \mathcal{R} ; If the smallest positive frequency among the observed cells reaches a tabled positive integer $c \equiv c(\varepsilon)$, then we stop sampling. That is, the stopping rule for the total (random) number N_c of observations required is

$$N_c = \inf \left\{ n \ni \underset{1 \leq i \leq K_n}{\text{Min}} f_i = c \right\}, \quad (2.2)$$

where $n = \sum_{i=1}^{K_n} f_i$, $n \geq 1$ and K_n is the number of distinct cells seen in n observations, and the f_i , $i = 1, 2, \dots, K_n$ are corresponding frequencies of each observed cell. The value of c , which is called the stopping value, will be determined by a P^* -condition and this in turn will determine the random variable N_c . We then use the statistic K_n as an estimate of the true number of cells k . This value K_n is clearly a lower bound on k and $K_n \rightarrow k$ as $c \rightarrow \infty$.

Since we do not know at stopping time whether we have seen all the cells or not, we define a correct decision (CD) as observing all of the cells with $p_i \geq \varepsilon$ by stopping time. We show below that for the procedure \mathcal{R} the $P\{CD|\mathcal{R}\} \geq P^*$ where P^* is a prespecified level of confidence such that $1/k \leq P^* < 1$; the tabled value of the stopping integer c depends on the prespecified P^* , as well as on ε . Equivalently the probability of a wrong decision $P\{WD|\mathcal{R}\}$, i.e., of stopping without seeing all the cells with $p_i \geq \varepsilon$, will be less than $1 - P^*$.

Under our ε -condition the proposed sequential procedure \mathcal{R} terminates with probability one with a finite number of observations since the value of c will always be finite regardless of how small a positive value is prespecified for ε .

2.2 The Configurations of the Cell Probabilities and Least Favorable configuration

The procedure \mathcal{R} we use is constructed so that it satisfies the usual P^* -requirement, namely $P\{CD|\mathcal{R}\} \geq P^*$ for any configurations of cell probabilities, vector \mathbf{p} , in the parameter space $\Omega(\varepsilon)$ in which every component is at least ε .

This leads us to find the so-called *least favorable configuration* (LFC) (Gibbons, Olkin, and Sobel (1977)). That is, we must look for the configuration which infimizes (or minimizes) the $P\{CD|\mathcal{R}\}$ over all vectors $\mathbf{p} \in \Omega(\varepsilon)$. Thus if such a limit configuration \mathbf{p}_{LFC} exists and is found, then we will be able to confine our attention solely to the LFC rather than the entire space of parameters Ω . However, we do not focus on finding the LFC in this article. We want to satisfy the basic requirement that $P\{CD|\mathcal{R}\} \geq P^*$ for all $\mathbf{p} \in \Omega(\varepsilon)$.

A probability configuration of parameters is said to be a ε -least favorable configuration (ε -LFC) if for any given N , k and ε the probability of a correct decision $P\{CD\}$ becomes a minimum (or infimum) over all vectors $\mathbf{p} \in \Omega(\varepsilon)$ under ε -condition. We note that the slippage ratio of cell probabilities ρ , defined as $p_{[j]}/p_{[j-1]}$, $j = 2, 3, \dots, k$, is related to find the LFC.

2.3 Basic Scheme of the Procedure

In the first stage of our solution we take $\varepsilon = 0.1$ and P^* to be the traditional value 0.95 to give a specific example of the use of the procedure \mathcal{R} . Then the goal is to find the smallest integers c under the stopping rule \mathcal{R} which satisfies the P^* -requirement.

Firstly let us consider the case where only one cell has been observed, i.e., $t = 1$. Using our assumption where only one cell has been observed, there may be a second cell; in the worst scenario suppose its cell probability values are $p_{[1]} = 0.1$ and $p_{[2]} = 0.9$. Also suppose $P^* = 0.95$. Since $P\{CD|\mathcal{R}\} + P\{WD|\mathcal{R}\} = 1$, we want to find integer c such that $P\{WD\} \leq 1 - P^*$. In order to find $c(0.1)$ for this situation, we set up the equation in n :

$$0.9^n + 0.1^n \leq 0.05 \tag{2.3}$$

By solving this equation we find n to be 28.4 and the smallest integer above this is 29. In this case c is the value of n that satisfies Eq. (2.3) under the condition $\varepsilon = 0.1$, namely 29. Therefore the number of observation required is 29 observations in the one cell observed in order to have $P\{CD|\mathcal{R}\} \geq 0.95$; the decision in this case would be $k = 1$. That is, if we observe one cell at least 29 times in a row, then we conclude with probability 0.95 that there is only one class (or kind) in the population.

In the next section we apply the same analogy to extend to the case where we observe more than one cell, i.e., $t \geq 2$.

3 PROBABILITY OF CORRECT DECISION AND DIRICHLET INTEGRALS

3.1 Multinomial Events and C and D Functions

Let two events E_1 and E_2 represent the minimum frequency and maximum frequency respectively among the cells at stopping time (a.s.t.) and the stopping rule is of the type used

in inverse sampling. Then applying the results shown in Olkin and Sobel (1965), we have

$$\begin{aligned}
P\{E_1\} &= P\{f_i \geq r, i = 1, 2, \dots, b \text{ a.s.t. when } f_{b+1} = m \text{ for the first time}\} \\
&= \int_0^{a_b} \int_0^{a_{b-1}} \dots \int_0^{a_1} f^{(b)}(\mathbf{x}; \mathbf{r}, m) \prod_{i=1}^b dx_i \\
&\equiv C_{\mathbf{a}}^{(b)}(\mathbf{r}, m)
\end{aligned} \tag{3.1}$$

and

$$\begin{aligned}
P\{E_2\} &= P\{f_i < r, i = 1, 2, \dots, b \text{ a.s.t. when } f_{b+1} = m \text{ for the first time}\} \\
&= \int_{a_b}^{\infty} \int_{a_{b-1}}^{\infty} \dots \int_{a_1}^{\infty} f^{(b)}(\mathbf{x}; \mathbf{r}, m) \prod_{i=1}^b dx_i \\
&\equiv D_{\mathbf{a}}^{(b)}(\mathbf{r}, m)
\end{aligned} \tag{3.2}$$

where

$$f^{(b)}(\mathbf{x}; \mathbf{r}, m) = \frac{\Gamma(m+R)}{\Gamma(m) \prod_{i=1}^b \Gamma(r_i)} \frac{\prod_{i=1}^b x_i^{r_i-1} dx_i}{\left(1 + \sum_{i=1}^b x_i\right)^{m+R}} \tag{3.3}$$

which is the (b -variate) Dirichlet-type II distribution, f_i denotes the frequency in the i -th cell, and $a_i = p_i/p_{b+1}$, $i = 1, 2, \dots, b$.

Consider a multinomial model with $1+b+j$ cells and the corresponding cell probabilities are p_0 and $\mathbf{p} = (p_1, p_2, \dots, p_b : p_{b+1}, \dots, p_{b+j})$, where $p_0 + \sum_{i=1}^{b+j} p_i = 1$. Let c be an integer such that $1 \leq c \leq b$. Suppose that we stop sampling when the frequency f_0 reaches the value of m for the first time. Let E denote the compound multinomial event that the frequencies for $i = 1, 2, \dots, c$ are all at least r , the frequencies f_i for $i = c+1, c+2, \dots, b$ are all less than r and the frequencies for $i = b+1, b+2, \dots, b+j$ are all exactly equal to r at stopping time. Then the probability of the compound multinomial event E is given by the CD -integrals in Sobel, Uppuluri and Frankowski (1985):

$$\begin{aligned}
CD_a^{(c, b-c; j)}(r; m) &= \frac{\Gamma(m+R)}{\Gamma(m)\Gamma^{b+j}(r)} \left(\prod_{i=b+1}^{b+j} \frac{a_i^r}{r} \right) \\
&\cdot \int_0^{a_1} \dots \int_0^{a_c} \int_{a_{c+1}}^{\infty} \dots \int_{a_b}^{\infty} \frac{\prod_{i=1}^b x_i^{r-1} dx_i}{\left(1 + \sum_{i=b+1}^{b+j} a_i + \sum_{i=1}^b x_i\right)^{m+R}},
\end{aligned} \tag{3.4}$$

where $\mathbf{a} = \mathbf{p}/p_0 = (p_1, p_2, \dots, p_b : p_{b+1}, \dots, p_{b+j})$ and $R = (b+j)r$. We note that c denotes the number of integrals of type C in the expression (3.4). Furthermore if $j = 0$, we are able to eliminate the j and simplify (3.4) as follows:

$$CD_a^{(c, b-c)}(r; m) = \frac{\Gamma(m+R)}{\Gamma(m)\Gamma^b(r)} \int_0^{a_1} \dots \int_0^{a_c} \int_{a_{c+1}}^{\infty} \dots \int_{a_b}^{\infty} \frac{\prod_{i=1}^b x_i^{r-1} dx_i}{\left(1 + \sum_{i=1}^b x_i\right)^{m+R}} \tag{3.5}$$

By letting $c = b$, this CD -integral reduces to the usual C -integral in Eq. (3.1) and letting $c = 0$, it reduces to the usual D -integral in Eq. (3.2). In addition, if we have a common value for the a_{α} , then we can use a without specifying the vector notations on the left-hand side of both Eqs. (3.4) and (3.5).

It should be noted that since there is a common value of r in each partition, it is convenient to use one C and one D in the left hand side of the expressions (3.4) and (3.5).

3.2 Expressions of $P\{WD\}$ based with C and D Functions

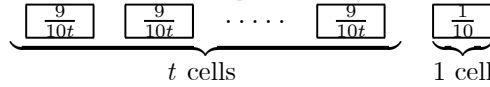
Let k be the true number of cells in the state of nature. Let T denote the observed number of cells at the stopping time. Also we assume in the present stage that every cell probability $p_i \geq 1/10$, $i = 1, 2, \dots, k$. Consider a configuration such as $\{9/10t, 9/10t, \dots, 9/10t, 1/10\}$, where $9/10t \geq 1/10$, $t = 1, 2, \dots, 9$. In this configuration the slippage ratio of cell probabilities ρ is a constant and $9/t$. Note that if $t = 10$, then we immediately stop sampling since under our ε -condition all p_i 's are equally probable spontaneously, and we call this *equal parameter configuration* (EPC).

We now are going to focus on the case of omitting exactly one cell.

3.2.1 $P\{WD\}$ for the case of Omitting One Cell: $T = k - 1$

In order to take the case of omitting one cell into consideration, first we need to investigate the structure of the cells as shown in Figure 3.1.

Figure 3.1 Cell structure and probabilities for omitting one cell



Since the total number of cells is the sum of the observed cells and the missing cell, the total number of cells from the above diagram must be $k = t + 1$ where $t = 1, 2, \dots, 9$.

According to the cell structure, there can be three possibilities in omitting one cell. Hence the $P\{WD\}$ can be completely considered on the basis of the three cases. Denoting by L the cell of size $9/10t$ and by S the cell of size $1/10$, the probability of a wrong decision $P\{WD\}$ for omitting one cell is composed of three parts by the cell structure, that is,

$$P\{WD\} = t \langle L; S \rangle + t(t-1) \langle L; L \rangle + t \langle S; L \rangle, \quad (3.6)$$

where $t \langle L; S \rangle$, for instance, represents the fact that there are t possible cases in which one of the L is the missing cell and S is the stopping cell.

Then by Eq. (3.4) we can express $P\{WD\}$ in terms of multiple of C and D -integrals as follows:

$$\begin{aligned} P\{WD\} &= tDC_{(\frac{t}{9}, 1)}^{(1, t-1)} + tDC_{(\frac{9}{t}, \frac{9}{t})}^{(1, t-1)}(1, c; c) + t(t-1)DCC_{(\frac{t}{9}, 1)}^{(1, 1, t-2)}(1, c, c; c) \\ &= t \frac{\Gamma(1+tc)}{\Gamma^t(c)} \int_{t/9}^{\infty} \int_0^1 \dots \int_0^1 \frac{dx \prod_{i=1}^{t-1} y_i^{c-1} dy_i}{\left(1+x+\sum_{i=1}^{t-1} y_i\right)^{1+tc}} \\ &\quad + t \frac{\Gamma(1+tc)}{\Gamma^t(c)} \int_{\frac{9}{t}}^{\infty} \int_0^{\frac{9}{t}} \dots \int_0^{\frac{9}{t}} \frac{dx \prod_{i=1}^{t-1} y_i^{c-1} dy_i}{\left(1+x+\sum_{i=1}^{t-1} y_i\right)^{1+tc}} \\ &\quad + t(t-1) \frac{\Gamma(1+tc)}{\Gamma^t(c)} \int_1^{\infty} \int_0^{\frac{t}{9}} \int_0^1 \dots \int_0^1 \frac{dx (y^{c-1} dy) \prod_{i=1}^{t-2} z_i^{c-1} dz_i}{\left(1+x+y+\sum_{i=1}^{t-2} z_i\right)^{1+tc}}. \end{aligned} \quad (3.7)$$

After some algebra and integrations of the D -type integrals we write Eq. (3.7) in the form

$$\begin{aligned}
P\{WD\} &= t \left(\frac{9}{9+t} \right)^c \frac{\Gamma(tc)}{\Gamma^t(c)} \int_0^{\frac{9}{9+t}} \cdots \int_0^{\frac{9}{9+t}} \frac{\prod_{i=1}^{t-1} w_i^{c-1} dw_i}{\left(1 + \sum_{i=1}^{t-1} w_i\right)^{tc}} \\
&+ t \left(\frac{t}{9+t} \right)^c \frac{\Gamma(tc)}{\Gamma^t(c)} \int_0^{\frac{9}{9+t}} \cdots \int_0^{\frac{9}{9+t}} \frac{\prod_{i=1}^{t-1} w_i^{c-1} dw_i}{\left(1 + \sum_{i=1}^{t-1} w_i\right)^{tc}} \\
&+ t(t-1) \left\{ \frac{1}{2^c} \frac{\Gamma(tc-c)}{\Gamma^{t-1}(c)} \int_0^{\frac{1}{2}} \cdots \int_0^{\frac{1}{2}} \frac{\prod_{i=1}^{t-2} w_i^{c-1} dw_i}{\left(1 + \sum_{i=1}^{t-2} w_i\right)^{tc-c}} \right. \\
&- \left(\frac{t}{9} \right)^{c-1} \left(\frac{9}{18+t} \right)^{2c-1} \frac{\Gamma(tc+1)}{\Gamma^2(c)\Gamma^{t-2}(c)} \int_0^{\frac{9}{18+t}} \cdots \int_0^{\frac{9}{18+t}} \\
&\times \frac{\prod_{i=1}^{t-2} w_i^{c-1} dw_i}{\left(1 + \sum_{i=1}^{t-2} w_i\right)^{tc-1}} - \left(\frac{t}{9} \right)^{c-2} \left(\frac{9}{18+t} \right)^{2c-2} \\
&\times \frac{\Gamma(tc+2)}{\Gamma(c)\Gamma(c-1)\Gamma^{t-2}(c)} \int_0^{\frac{9}{18+t}} \cdots \int_0^{\frac{9}{18+t}} \frac{\prod_{i=1}^{t-2} w_i^{c-1} dw_i}{\left(1 + \sum_{i=1}^{t-2} w_i\right)^{tc-2}} \\
&\dots \\
&- \left(\frac{9}{18+t} \right)^c \frac{\Gamma(tc+c)}{\Gamma(c)\Gamma(1)\Gamma^{t-2}(c)} \\
&\times \left. \int_0^{\frac{9}{18+t}} \cdots \int_0^{\frac{9}{18+t}} \frac{\prod_{i=1}^{t-2} w_i^{c-1} dw_i}{\left(1 + \sum_{i=1}^{t-2} w_i\right)^{tc-c}} \right\}. \tag{3.8}
\end{aligned}$$

After combining similar terms and reshuffling the remaining terms in terms of the C -function, this can be written as

$$\begin{aligned}
P\{WD\} &= t \left\{ \left(\frac{9}{9+t} \right)^c + \left(\frac{t}{9+t} \right)^c \right\} C_{\frac{9}{9+t}}^{(t-1)}(c; c) + \frac{t(t-1)}{2^c} \left\{ C_{\frac{1}{2}}^{(t-2)}(c; c) \right. \\
&- \left. \sum_{i=1}^c \left(\frac{t}{18+t} \right)^{c-i} \left(\frac{18}{18+t} \right)^c \binom{2c-1-i}{c-1} C_{\frac{9}{18+t}}^{(t-2)}(c; 2c-i) \right\}. \tag{3.9}
\end{aligned}$$

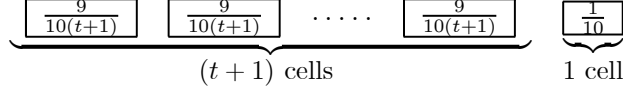
From Eq. (3.9), we are able to obtain values of c corresponding to different values of t , ($t = 1, 2, \dots, 9$). The values are given in Table 3.1 when P^* is specified as 0.95 under $\varepsilon = 0.1$. It is most important to note that $P\{WD\} \leq 0.05$ uniformly. It should also be noted that $P\{WD, T = k-1\}$ in Table 3.1 is not monotonic in either t or c . A reasonable explanation for this is that both the least favorable configuration and the c values are changing. After c settles (to 4 in the table) and the least favorable configuration becomes the one with exactly one cell probability slipped to the left and thereafter the results in Table 3.1 are monotonic.

As we see in the table, the maximum value t can take is 9 since the number of missing cells is 1 and total number of cells $k = t + 1$, and the probabilities of omitting one cell are uniformly smaller than $1 - P^*$ which is 0.05. Therefore the suggested procedure is well justified by guaranteeing P^* -condition under the assumption.

3.2.2 $P\{WD\}$ for the case of Omitting Two Cells: $T = k - 2$

In order to apply the same analogy of the case of omitting one cell, let us consider the cell structure for the case of missing two cells.

Figure 3.2 Cell structure and probabilities for omitting two cells



Since the total number of cells is the sum of the observed cells and the missing cells, the noticeable difference from the case of omitting one cell is that the total number of cells as shown in Figure 3.2 must be $k = (t + 1) + 1 = t + 2$ where $t = 1, 2, \dots, 8$. By the cell structure, there can be three possibilities in omitting two cells. Hence the $P\{WD\}$ can be completely considered on the basis of the following three cases. Denoting L by the cell of size $9/10t$ and S by the cell of size $1/10$, the probability of wrong decision $P\{WD\}$ when we have two unobserved cells is given by

$$P\{WD\} = \binom{t+1}{2} \langle L, L; S \rangle + t(t+1) \langle L, S; L \rangle + (t+1) \binom{t}{2} \langle L, L; L \rangle, \quad (3.10)$$

where $t(t+1) \langle L, S; L \rangle$, for instance, represents that there are $t(t+1)$ possible cases of which one of L and S are missing cell and L is the stopping cell. Then we can express $P\{WD\}$ in terms of multiple of C and D -integrals by Eq. (3.6).

$$\begin{aligned} P\{WD\} &= \binom{t+1}{2} D_{\frac{9}{t+1}}^{(1)} D_{\frac{9}{t+1}}^{(1)} C_{\frac{9}{t+1}}^{(1)}(1, 1, c; c) \\ &\quad + (t+1)t D_1^{(1)} D_{\frac{9}{t+1}}^{(1)} C_1^{(t-1)}(1, 1, c; c) \\ &\quad + (t+1) \binom{t}{2} D_1^{(1)} D_1^{(1)} C_1^{(t-2)} C_{\frac{9}{t+1}}^{(1)}(1, 1, c, c; c). \end{aligned} \quad (3.11)$$

Then we can express $P\{WD\}$ in terms of multiple of C and D -integrals by Eq. (3.6) as follows.

$$\begin{aligned} P\{WD\} &= \frac{t(t+1) \Gamma(tc+2)}{2 \Gamma^t(c)} \\ &\quad \cdot \int_{\frac{9}{t+1}}^{\infty} \int_{\frac{9}{t+1}}^{\infty} \int_0^{\frac{9}{t+1}} \cdots \int_0^{\frac{9}{t+1}} \frac{\prod_{i=1}^{t-1} x_i^{c-1} dx_i dy_1 dy_2}{\left(1 + y_1 + y_2 + \sum_{i=1}^{t-1} x_i\right)^{tc+2}} \\ &\quad + t(t+1) \frac{\Gamma(2+tc)}{\Gamma^t(c)} \\ &\quad \cdot \int_1^{\infty} \int_{\frac{t+19}{9}}^{\infty} \int_0^1 \cdots \int_0^1 \frac{\prod_{i=1}^{t-1} x_i^{c-1} dx_i dy_1 dy_2}{\left(1 + y_1 + y_2 + \sum_{i=1}^{t-1} x_i\right)^{tc+2}} \\ &\quad + \frac{(t-1)t(t+1) \Gamma(tc+2)}{2 \Gamma^t(c)} \\ &\quad \cdot \int_1^{\infty} \int_1^{\infty} \int_0^1 \cdots \int_0^1 \int_0^{\frac{t+1}{9}} \frac{y_3^{c-1} dy_3 \prod_{i=1}^{t-2} x_i^{c-1} dx_i dy_1 dy_2}{\left(1 + y_1 + y_2 + \sum_{i=1}^{t-2} x_i + y_3\right)^{tc+2}}. \end{aligned} \quad (3.12)$$

After considerable algebra and integrations, we obtain

$$\begin{aligned}
P\{WD\} &= \frac{t(t+1)}{2} \left(\frac{t+1}{t+19}\right)^{tc} \frac{\Gamma(tc)}{\Gamma(c)\Gamma^{t-1}(c)} \int_0^{\frac{9}{t+19}} \cdots \int_0^{\frac{9}{t+19}} \frac{\prod_{i=1}^{t-1} y_i^{c-1} dy_i}{\left(1 + \sum_{i=1}^{t-1} y_i\right)^{tc}} \\
&+ t(t+1) \left(\frac{9}{t+19}\right)^c \frac{\Gamma(tc)}{\Gamma(c)\Gamma^{t-1}(c)} \int_0^{\frac{t}{t+19}} \cdots \int_0^{\frac{t}{t+19}} \frac{\prod_{i=1}^{t-1} y_i^{c-1} dy_i}{\left(1 + \sum_{i=1}^{t-1} y_i\right)^{tc}} \\
&+ \frac{t(t-1)(t+1)}{2} \left\{ \frac{1}{3^c} \frac{\Gamma(tc-c)}{\Gamma(c)\Gamma^{t-2}(c)} \int_0^{\frac{1}{3}} \cdots \int_0^{\frac{1}{3}} \frac{\prod_{i=1}^{t-2} y_i^{c-1} dy_i}{\left(1 + \sum_{i=1}^{t-2} y_i\right)^{tc-c}} \right. \\
&- \sum_{i=1}^c \left(\frac{t+1}{9}\right)^{c-i} \left(\frac{9}{28+t}\right)^{2c-i} \frac{\Gamma(2c-i)}{\Gamma(c)\Gamma(c-i+1)} \frac{\Gamma(tc-i)}{\Gamma^{t-2}(c)\Gamma(2c-i)} \\
&\left. \times \int_0^{\frac{9}{28+t}} \cdots \int_0^{\frac{9}{28+t}} \frac{\prod_{i=1}^{t-2} y_i^{c-1} dy_i}{\left(1 + \sum_{i=1}^{t-2} y_i\right)^{tc-i}} \right\}. \tag{3.13}
\end{aligned}$$

Using Eq. (3.1) this can be written in terms of C -functions as

$$\begin{aligned}
P\{WD\} &= \frac{t(t+1)}{2} \left(\frac{t+1}{t+19}\right)^c C_{\frac{9}{t+19}}^{(t-1)}(c; c) + t(t+1) \left(\frac{9}{t+19}\right)^c C_{\frac{9}{t+19}}^{(t-1)}(c; c) \\
&+ \frac{t(t-1)(t+1)}{2 \cdot 3^c} \left\{ C_{\frac{1}{3}}^{(t-2)}(c; c) - \sum_{i=1}^c \left(\frac{t+1}{28+t}\right)^{c-i} \left(\frac{27}{28+t}\right)^c \right. \\
&\left. \times \binom{2c-1-i}{c-1} C_{\frac{9}{28+t}}^{(t-2)}(c; 2c-i) \right\}. \tag{3.14}
\end{aligned}$$

Since the number of missing cells is 2 and total number of cells $k = t + 2$, the maximum value t can take is up to 8. The values of the stopping time and $P\{WD\}$ are calculated and given in Table 3.1. We need to note importantly that three components in above expression have the identical C -functions when $t = 8$ except coefficients of t . We know that if $t = 8$, the configuration of cell structure becomes EPC.

3.2.3 Total Probability of Wrong Decision

Since the events that exactly 1 cell is missing, exactly 2 cells are missing, ..., are mutually exclusive and $P\{WD\}$ for more than two cells missing are extremely small by using the same analysis as in Subsections 3.3.1 and 3.3.2, the total $P\{WD\}$ is

$$\begin{aligned}
P\{WD\} &= \sum_{j=1}^{k-1} P\{WD, t = k - j\} \\
&\approx P\{WD, k = t - 1\} + P\{WD, k = t - 2\}. \tag{3.15}
\end{aligned}$$

Therefore we can approximate the total $P\{WD\}$ as being close to the sum of column (1) and column (2) in the Table 3.1 and indeed we note that the results are uniformly below 0.05, which satisfies P^* -condition in the procedure, for each value of k . Therefore the proposed

procedure is well justified by guaranteeing the P^* -condition under the assumption.

Table 3.1 Total $P\{WD\}$ for $\varepsilon = 0.1$, $P^* = 0.95$

| k | t | $c(\varepsilon)$ | $P\{WD, t=k-1\}$ | $P\{WD, t=k-2\}$ | Total $P\{WD\}$ |
|-----|-----|------------------|------------------|--------------------------|-----------------|
| | | | (1) | (2) | (1) + (2) |
| 2 | 1 | 29 | 0.04710 | 0 | 0.04710 |
| 3 | 2 | 13 | 0.04510 | 1.1754×10^{-10} | 0.04510 |
| 4 | 3 | 8 | 0.04181 | 1.1724×10^{-6} | 0.04182 |
| 5 | 4 | 6 | 0.03234 | 5.9684×10^{-5} | 0.03240 |
| 6 | 5 | 5 | 0.02490 | 2.3480×10^{-4} | 0.02514 |
| 7 | 6 | 4 | 0.03348 | 4.5431×10^{-4} | 0.03393 |
| 8 | 7 | 4 | 0.02503 | 1.6251×10^{-3} | 0.02665 |
| 9 | 8 | 4 | 0.02117 | 1.1589×10^{-3} | 0.02233 |
| 10 | 9 | 4 | 0.01921 | 9.1749×10^{-4} | 0.02013 |

As we see in Table 3.1, and the probabilities of omitting two cells are uniformly smaller than $1 - P^*$ which is 0.05. We note that for two cells missing the probabilities are relatively very smaller than the probability of one missing cell and have a negligible effect on raising the probability of wrong decision.

Also additionally, we find the stopping values and calculate the lower bound (LB) for the expected number of observations $E(N)$ given t , i.e., minimum total number of observations for the case of $P^* = 0.99$. Then we define the c -value of Table 3.1 to be the stopping values of the procedure \mathcal{R} and obtain the following table:

Table 3.2 Values of the stopping constant c_ε as a function of t under procedure \mathcal{R} , $\varepsilon = 0.1$

| t | $P^* = 0.95$ | | $P^* = 0.99$ | |
|-----|--------------|------------------|--------------|------------------|
| | c | LB of $E(N t)$ | c | LB of $E(N t)$ |
| 1 | 29 | 29 | 44 | 44 |
| 2 | 13 | 26 | 20 | 40 |
| 3 | 8 | 24 | 13 | 39 |
| 4 | 6 | 24 | 9 | 36 |
| 5 | 5 | 25 | 7 | 35 |
| 6 | 4 | 24 | 6 | 36 |
| 7 | 4 | 28 | 5 | 35 |
| 8 | 4 | 32 | 5 | 40 |
| 9 | 4 | 36 | 5 | 45 |

We note that we do not use any c -value for $t = 10$ since we simply assert there must be exactly 10 cells under the ε -condition.

Illustration 3.1 Suppose we observe 4 distinct cells ($t = 4$). Then from Table 3.2 we obtain $c = 6$ for $P^* = 0.95$. If we have observed at least $c = 6$ from each of the $t = 4$ cells, we stop sampling and assert that there are exactly 4 cells. Under our ε -condition, all p_i 's are ≥ 0.1 , the probability that our decision is correct is at least 0.95.

However, if we wish to have $P\{CD\} \geq 0.99$, we get $c = 9$ according to the table. This implies that there should be at least $c = 9$ observations from each of the $t = 4$ cells under the assumption $\varepsilon = 1/10$ to keep the level of confidence at least 0.99, which is guaranteed by the proposed procedure \mathcal{R} .

4 EXPECTATIONS OF THE NUMBER OF OBSERVATIONS IN THE PROCEDURE

4.1 Expected Total Number of Observations

Let N_c denote the total number of observations at stopping time needed by the procedure \mathcal{R} , which is an inverse sampling procedure. Now we denote the total number of true cells by b (in fact, b stands for blue cell). The stopping rule says that sampling is terminated when the minimum positive frequency reaches a tabled value c which depends upon the observed number of different cells t and prespecified P^* . Clearly the total number of observations is random but its moments are functions of the true number of cells and P^* ; they do not depend on c or t .

Before we derive an explicit expression for the expectation of N_c , we introduce the general representation for g -th ascending factorial moment of the waiting time until every one of the b cells reaches its quota in a simpler setting than our problem; the quota for cell i is r_i ($i = 1, 2, \dots, b$). Sobel, Uppuluri and Frankowski (1985, p. 62 Equation 5.8) have shown that the γ -th ascending factorial moment of the waiting time until the every one of the b cells reaches its quota is given by

$$\mu^{[\gamma]} = \sum_{\alpha=1}^b \frac{\Gamma(r_\alpha + \gamma)}{\Gamma(r_\alpha) p_\alpha^\gamma} C_{\mathbf{p}_{<-\alpha>/p_\alpha}}^{(b-1)}(\mathbf{r}_{<-\alpha>; r_\alpha + \gamma); \quad (4.1)$$

here α refers to the cell which is the last one to reach its quota, where $\mathbf{p}_{<-\alpha>/p_\alpha} = \left(\frac{p_1}{p_\alpha}, \frac{p_2}{p_\alpha}, \dots, \frac{p_{\alpha-1}}{p_\alpha}, \frac{p_{\alpha+1}}{p_\alpha}, \dots, \frac{p_b}{p_\alpha} \right)$; $\mathbf{r}_{<-\alpha>}$ denotes the vector of parameters (r -values) with the component r_α absent, i.e., $(r_1, r_2, \dots, r_{\alpha-1}, r_{\alpha+1}, \dots, r_b)$.

Now we consider a certain useful compound event and in order to apply the above factorial moment to our situation. In the notations we have been using, we have $N_c = f_1 + f_2 + \dots + f_b$, where f_i (Below we also use x_i for the value of f_i) is the observed cell frequency in the i -th cell at stopping time, $i = 1, 2, \dots, b$.

Lemma 4.1 *Let $\mathbf{X} = (X_1, X_2, \dots, X_b)$ be the b -variate multinomial vector of frequencies with corresponding cell probabilities $\mathbf{p} = (p_1, p_2, \dots, p_b)$ and $\sum_{i=1}^b p_i = 1$. Define an event $E = \{ \text{at stopping time (a.s.t.) } x_\alpha = c \text{ for the first time: } x_i \geq c \text{ (} i = 1, 2, \dots, \alpha - 1) \text{ and } x_i = 0 \text{ (} i = \alpha + 1, \alpha + 2, \dots, b) \}$ which can occur under our inverse sampling procedure \mathcal{R} . Let N_c be the contribution to the total number of observations at stopping time, $c + x_1 + x_2 + \dots + x_{\alpha-1} = c + x_1 + x_2 + \dots + x_{\alpha-1} + x_{\alpha+1} + \dots + x_b$, contributed by the event E . Then, writing $E(N_c)$ for this contribution, the total expected sample size (a.s.t.) is a sum of terms of the form*

$$E(N_c) = \sum_{\alpha=1}^b \left(\frac{c}{p_\alpha} \right) C_{\mathbf{p}_{<-\alpha>/p_\alpha}}^{(b-1)}(\mathbf{r}_{<-\alpha>; c + 1), \quad (4.2)$$

where $\mathbf{p}_{<-\alpha>/p_\alpha} = \left(\frac{p_1}{p_\alpha}, \frac{p_2}{p_\alpha}, \dots, \frac{p_{\alpha-1}}{p_\alpha}, \frac{p_{\alpha+1}}{p_\alpha}, \dots, \frac{p_b}{p_\alpha} \right)$ and we use the scalar c when all the r 's are equal to c which is the minimum positive observed frequency at stopping time (our stopping constant).

Proof. Let denote x_i be the observed frequency at stopping time for i -th cell, $i = 1, 2, \dots, b$. There are now $(b - 1)$ blue cells and one stopping cell. Since $N_c = c + \sum_{i=1, i \neq \alpha}^b x_i$ at stopping time, the expectation of N_c becomes as follows:

$$E(N_c) = \sum_{\alpha=1}^b (x_1 + x_2 + \dots + x_b) \cdot P \{ \text{a.s.t. } X_\alpha = c, X_i \geq c, X_j = 0 \} \quad (4.3)$$

where $i = 1, 2, \dots, \alpha - 1$, and $j = \alpha + 1, \alpha + 2, \dots, b$. Then we obtain

$$\begin{aligned}
E(N_c) &= \sum_{\alpha=1}^b \left[\sum_{x_1=c}^{\infty} \sum_{x_2=c}^{\infty} \cdots \sum_{x_{\alpha-1}=c}^{\infty} \left\{ \frac{(c + \sum_{j=1, j \neq \alpha}^b x_j)!}{c! \prod_{j=1, j \neq \alpha}^b x_j!} p_{\alpha}^c \prod_{l=1, l \neq \alpha}^b p_l^{x_l} \right\} \right] \\
&= \sum_{\alpha=1}^b \left(\frac{c}{p_{\alpha}} \right) \left[\sum_{x_1=c}^{\infty} \sum_{x_2=c}^{\infty} \cdots \sum_{x_{\alpha-1}=c}^{\infty} \right. \\
&\quad \left. \times \left\{ \frac{\Gamma[(c+1) + \sum_{j=1, j \neq \alpha}^b x_j]}{\Gamma(c+1) \prod_{j=1, j \neq \alpha}^b \Gamma(x_j+1)} p_{\alpha}^{c+1} \prod_{l=1, l \neq \alpha}^b p_l^{x_l} \right\} \right]. \tag{4.4}
\end{aligned}$$

Using the exact correspondence between the cumulative multinomial distribution and Dirichlet-Type II (i.e., the C -integral) in Olkin and Sobel (1965) and Sobel, Uppuluri and Frankowski (1985), we obtain (4.2). ■

We note that we use the scalar c when all the r 's are equal to c which is the minimum positive observed frequency at stopping time (our stopping constant). That is, from the multinomial interpretation of C -integrals (in fact, D -integrals have vanished due to some cells which have not appeared before the stopping time) and the expression in Eq. (3.6), the above result exactly coincides with the right hand side of Eq. (4.1).

Example 4.1 Consider the case of $k = 2$. The configuration of ε -LFC for $k = 2$ is $\{0.9, 0.1\}$. Since the maximum number of missing cells is at most 1, the probability of the wrong decision will be exact in this case. From the Table 3.1, $P\{CD\} = 1 - P\{WD\} = 1 - 0.04710 = 0.95290$. From Eq. (4.2) we obtain

$$\begin{aligned}
E(N_c) &= 0.9529 \left[\left\{ \frac{13(1)}{0.9} C_{\frac{1}{9}}^{(1)}(13, 14) + \frac{13(1)}{0.1} C_9^{(1)}(13, 14) \right\} \right. \\
&\quad \left. - \left\{ \frac{29(1)}{0.9} D_{\frac{1}{9}}^{(1)}(1, 30) + \frac{29(1)}{0.1} D_9^{(1)}(1, 30) \right\} \right] \\
&\quad + 0.04710 \left[\frac{29(1)}{0.9} D_{\frac{1}{9}}^{(1)}(1, 30) + \frac{29(1)}{0.1} D_9^{(1)}(1, 30) \right] \\
&= 122.6396.
\end{aligned}$$

This $E(N_c|k=2)$ is quite agreeable with the result we obtain from Monte Carlo simulation which is 122.91 in Table 5.1.

Example 4.2 Consider the case of $k = 4$. The configuration of ε -LFC for $k = 4$ is $\{0.3, 0.3, 0.3, 0.1\}$. The number of missing cells can be more than 1, so we approximate the wrong decision based on one or two missing. From the Table 3.1, $P\{CD\} = 1 - 0.04181 = 0.95818$. Similarly, from (4.2) we obtain

$$\begin{aligned}
E(N_c) &= 0.95818 \left[\left\{ \frac{6(1)}{0.3} C_{\frac{1}{3}}^{(1)} C_1^{(1)}(6, 6; 7) + \frac{6(1)}{0.1} C_3^{(3)}(6; 7) \right\} \right. \\
&\quad \left. - \left\{ \frac{8(3)}{0.3} D_{\frac{1}{3}}^{(1)} C_1^{(2)}(1, 8; 9) + \frac{8(6)}{0.3} D_1^{(1)} C_1^{(1)} C_{\frac{1}{3}}^{(1)}(1, 8, 8; 9) \right. \right. \\
&\quad \left. \left. + \frac{8(3)}{0.1} D_3^{(1)} C_3^{(2)}(1, 8; 9) \right\} \right] + 0.04710 \left[\frac{8(3)}{0.3} D_{\frac{1}{3}}^{(1)} C_1^{(2)}(1, 8; 9) \right. \\
&\quad \left. + \frac{8(6)}{0.3} D_1^{(1)} C_1^{(1)} C_{\frac{1}{3}}^{(1)}(1, 8, 8; 9) + \frac{8(3)}{0.1} D_3^{(1)} C_3^{(2)}(1, 8; 9) \right] \\
&= 56.9194.
\end{aligned}$$

This $E(N_c|k=4)$ is comparable with the result we obtain from Monte Carlo simulation which is 57.15 in Table 5.2.

4.2 Variance of the Total Number of Observations

The first ascending factorial moment ($\gamma = 1$) is clearly the mean $\mu = E(N_c)$. Since $\mu^{[2]} = E\{N_c(N_c + 1)\}$ is the second ascending factorial moment of N_c , the variance σ^2 is obtained from the relationship

$$VAR(N_c) = \mu^{[2]} - \mu(1 + \mu). \quad (4.5)$$

5 SIMULATION STUDIES

In the Monte Carlo experimentation, we used several configurations for each value of k , ($k = 2, 4, 8$) depending on the number t of cells that have been observed. We calculated the $P\{CD\}$ for each configuration and several other configurations as well.

5.1 Monte Carlo Experimentation

The results of the Monte Carlo simulation are summarized in the following tables, which show the probability of correct decision, $P\{CD\}$, the average number of observations, $E(N)$, its standard error, *s.e.*, and the number of wrong decisions in the experiments. Every row in each of Table 5.1 to 5.3 corresponds to 10,000 independent experiments.

Table 5.1 For $k = 2; P\{CD\} \geq P^*$, $P^* = 0.95$, $\varepsilon = 0.1$

| Configuration | $P\{CD\}$ | $E(N)$ | <i>s.e.</i> | #WD |
|--------------------------------|-----------|--------|-------------|-----|
| $(\frac{1}{2}, \frac{1}{2})$ | 1.0000 | 30.06 | 0.0378 | 0 |
| $(\frac{9}{10}, \frac{1}{10})$ | 0.9543 | 122.91 | 0.3837 | 457 |
| $(\frac{8}{10}, \frac{2}{10})$ | 0.9989 | 64.77 | 0.1598 | 11 |

Table 5.2 For $k = 4; P\{CD\} \geq P^*$, $P^* = 0.95$, $\varepsilon = 0.1$

| Configuration | $P\{CD\}$ | $E(N)$ | <i>s.e.</i> | #WD |
|--|-----------|--------|-------------|-----|
| $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ | 0.9993 | 34.60 | 0.0673 | 7 |
| $(\frac{3}{10}, \frac{3}{10}, \frac{3}{10}, \frac{1}{10})$ | 0.9557 | 57.15 | 0.2204 | 443 |
| $(\frac{4}{10}, \frac{4}{10}, \frac{1}{10}, \frac{1}{10})$ | 0.9912 | 73.54 | 0.2218 | 88 |
| $(\frac{7}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10})$ | 0.9975 | 80.78 | 0.2072 | 25 |

Table 5.3 For $k = 8; P\{CD\} \geq P^*$, $P^* = 0.95$, $\varepsilon = 0.1$

| Configuration | $P\{CD\}$ | $E(N)$ | <i>s.e.</i> | #WD |
|--|-----------|--------|-------------|-----|
| $(\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$ | 0.9752 | 57.01 | 0.1254 | 248 |
| $(\frac{9}{20}, \frac{9}{20}, \frac{9}{20}, \frac{9}{20}, \frac{9}{20}, \frac{9}{20}, \frac{9}{20}, \frac{1}{10})$ | 0.9769 | 58.37 | 0.1357 | 231 |
| $(\frac{1}{15}, \frac{1}{15}, \frac{1}{15}, \frac{1}{15}, \frac{1}{15}, \frac{1}{15}, \frac{1}{10}, \frac{1}{10})$ | 0.9695 | 59.46 | 0.1449 | 305 |
| $(\frac{1}{30}, \frac{1}{30}, \frac{1}{30}, \frac{1}{30}, \frac{1}{30}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10})$ | 0.9698 | 60.94 | 0.1533 | 302 |
| $(\frac{1}{20}, \frac{1}{20}, \frac{1}{20}, \frac{1}{20}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10})$ | 0.9678 | 62.48 | 0.1594 | 322 |
| $(\frac{1}{30}, \frac{1}{30}, \frac{1}{30}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10})$ | 0.9649 | 64.72 | 0.1676 | 351 |
| $(\frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10})$ | 0.9691 | 66.69 | 0.1670 | 309 |
| $(\frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10})$ | 0.9738 | 69.23 | 0.1654 | 262 |

REFERENCES

1. Arnold, B. and Beaver, R. (1988). Estimation of the number of classes in a population. *Biometrical Journal*. **30** 413-424.
2. Boender, C. and Rinnooy Kan, A. (1983). A Bayesian analysis of the number of cells of a multinomial distribution. *The Statistician*. **32** 240-248.
3. Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: a review. *J. Ameri. Statist. Assoc.* **88** 364-373.
4. Cho, H. (1997). Sequential estimation of the number of classes in a multinomial distribution. *Ph.D. dissertation*. University of California, Santa Barbara, California.
5. Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*. **63** 435-447.
6. Gibbons, J., Olkin, I., and Sobel, M. (1977). *Selecting and Ordering Population: A New Statistical Methodology*. John Wiley & Sons, New York.
7. Goodman, L. (1949). On the estimation of the number of classes in a population. *Ann. Math. Statist.* **20** 572-579.
8. Mann, C. (1991). Extinction: Are Ecologist crying wolf? *Science*. **253** 709-824.
9. Marchand, J. and Schroeck, F. Jr. (1982). On estimation of the number of equally likely classes in a population. *Commun. Statist.* **11** 1139-1146.
10. McNeil, D (1973). Estimating an author's vocabulary. *J. Ameri. Statist. Assoc.* **68** 92-96.
11. Olkin, I. and Sobel, M. (1965). Integral expression for tail probabilities of the multinomial and negative multinomial distributions, *Biometrika*. **52** 167-179.
12. Sobel, M., Uppuluri, V. and Frankowski, K. (1985). *Selected Tables in Mathematical Statistics, Vol. IX -Dirichlet Integrals of Type II and Their Applications*. Edited by IMS. American Mathematical Society, Providence, Rhode Island.

Department of Mathematical Sciences
University of Nevada, Las Vegas
Las Vegas, NV 89154-4020
E-mail: cho@unlv.nevada.edu