

Statistical Identification in Multinomial Models with Sequential Sampling

Hokwon A. Cho

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

In Memory of Professor Milton Sobel, 1919-2002

Abstract

We propose an inverse-type sequential method of statistical identification in multinomial models having unequal cell probabilities. Using the indifference-zone formulation and based on the likelihood ratio of decision vectors, a stopping rule is devised that controls the probability of a correct identification, $P\{CI\}$ and satisfies a preassigned probability level condition P^* . By performing a Monte Carlo experiment, the expected sample sizes are obtained and the numerical results of the proposed procedure are presented for illustration.

Key Words and Phrases: statistical identification; inverse sequential sampling; indifference-zone formulation; decision vectors; probability of correct identification.

1 Introduction

Suppose $\mathbf{X} = (X_1, X_2, \dots, X_k)$ is a vector of observations from a multinomial probability distribution function

$$g(\mathbf{x}|\boldsymbol{\theta}) = \frac{n!}{x_1!x_2! \cdots x_k!} \theta_1^{x_1} \theta_2^{x_2} \cdots \theta_k^{x_k}, \quad (1)$$

where the θ_i 's are parameters with $\sum_{i=1}^k \theta_i = 1$ and $\sum_{i=1}^k x_i = n$ with $x_i \in \{0, 1, \dots, n\}$. Let $\Omega = \{\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k) : 0 < \theta_i < 1, i = 1, 2, \dots, k\}$ denote the parameter space, and θ_i corresponds to the probability of the i -th component (or i -th category). Let $\theta_{[1]} \leq \theta_{[2]} \leq \cdots \leq \theta_{[k]}$ denote the ordered θ_i .

The random variable X_i is paired with the parameter θ_i in the sense that, for every subset of (i_1, i_2, \dots, i_s) of $(1, 2, \dots, k)$, the joint distribution of $(X_{i_1}, X_{i_2}, \dots, X_{i_s})$

depends on the set $(\theta_1, \theta_2, \dots, \theta_k)$ only through the subset $(\theta_{i_1}, \theta_{i_2}, \dots, \theta_{i_s})$. Since it is assumed that the true association (by pairing) of the x_i with the $\theta_{[i]}$ is unknown to the experimenter, the goal of the statistical identification problem is to identify the component i correctly associated with $\theta_{[i]}$. (See Bechhofer, Kiefer and Sobel 1977, and Gupta and Panchapakesan 1979).

A correct identification (CI) occurs if one correctly identifies the component i associated with $\theta_{[i]}$, $i = 1, 2, \dots, k$. Let \mathcal{R} be a valid procedure for the problem. Then, the rule \mathcal{R} is constructed so that it satisfies the P^* -requirement; namely, the probability of correct identification (PCI) will be at least P^* for specified $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$ with a pre-determined value of the distance measure δ^* (> 0). Denoting the PCI for the suggested identification procedure \mathcal{R} by $P\{CI|\mathcal{R}\}$, at the stopping time we require that

$$P\{CI|\mathcal{R}\} \geq P^* \text{ whenever } \{\boldsymbol{\theta} \in \Omega : \theta_{[j]} - \theta_{[j-1]} \geq \delta^*, j = 2, 3, \dots, k\}. \quad (2)$$

From a practical viewpoint, we fix a small positive discerning value δ^* and say that populations with $\theta_{[j]} - \theta_{[j-1]} < \delta^*$ are either indistinguishable or not of practical interest in identifying components in multinomial models. We refer to this as the δ^* -condition and the subspace $\Omega(\delta^*) = \{\boldsymbol{\theta} : \theta_{[j]} - \theta_{[j-1]} \geq \delta^*\} \subseteq \Omega$ is called the *preference-zone*, whereas its complement is called the *indifference-zone*. When $\boldsymbol{\theta}$ falls in the indifference-zone, as its name indicates, there is no probability requirement set on the PCI. The formulation we describe in the above is known as the indifference-zone formulation in Ranking and Selection Methodology (RSM). (See Gibbons, Olkin and Sobel 1977, and Gupta and Panchapakesan 1979).

1.1 The Least Favorable Configuration

The procedure \mathcal{R} is constructed so that it satisfies the usual P^* -requirement: $P\{CI\} \geq P^*$ for any configuration $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$ in the preference-zone $\Omega(\delta^*)$. In order to satisfy the inequality (2), one typically evaluates the infimum of the PCI over the region $\Omega(\delta^*)$ and seeks the smallest value of n such that

$$\inf_{\Omega(\delta^*)} P_{\boldsymbol{\theta}}\{CI|\mathcal{R}\} \geq P^*. \quad (3)$$

This leads us to find the so-called *least favorable configuration* (LFC) which minimizes the PCI over all vectors $\boldsymbol{\theta} \in \Omega(\delta^*)$. If such a limit configuration $\boldsymbol{\theta}_{LFC}$ exists and is found, then we can confine attention solely to the LFC rather than the entire space of heterogeneity Ω . However, we do not focus rigorously on finding the LFC in this paper. Instead, we examine the LFC through numerical studies in Section 4.

A probability configuration of parameters is said to be a δ^* -least favorable configuration (δ^* -LFC) if, for any given N , k and δ^* , the PCI becomes an infimum over all vectors $\boldsymbol{\theta} \in \Omega(\delta^*)$ under the δ^* -condition. For a given configuration, define the slippage ratio ρ_j by $\theta_{[j]}/\theta_{[j-1]}$, $j = 2, 3, \dots, k$. Then, among $(k - 1)$ slippage ratios, we can find the LFC using either $\min_j \rho_j$ or $\max_j \rho_j$. (See Gibbons, Olkin and Sobel 1977, and Cho 2003).

Remark 1.1 Since the multinomial distribution described in Eq. (1) belongs to an exponential family with a connected natural parameter, the identification problem can be thought as a solution of the ranking and selection problem. Bechhofer, Kiefer and Sobel (1968) have discussed several cases of their general results to ranking distributions of exponential family using Monte Carlo sampling results. Gupta and Panchapakesan (1979) show that a frequency function (in our case, the multinomial distribution) possesses the monotone likelihood ratio property if it satisfies the rankability condition (see p. 41 and p. 168 of Bechhofer, Kiefer and Sobel, 1968). Also, Huang and Panchapakesan (1978) considered the complete ranking problem with a subset selection goal to the means of normal distributions.

2 Stopping Rule in the Procedure

Now, we consider the sequential method of identifying components from a k -variate multinomial population. In their monograph, Bechhofer, Kiefer and Sobel seem to be the first who studied the theoretical framework of sequential identification and ranking procedures in a general scheme. Using their idea, we show the sequential identification procedure for the multinomial population.

Suppose we have n vectors of observations. Since there are k component categories, $k!$ decision vectors d_i ($i = 1, 2, \dots, k!$) are available for the terminal decision. Let $L_n(\Omega)$ denote the sum of all $k!$ possible likelihood functions in the entire decision space Ω , and let $L_n(\Omega_d)$ represent the sum of the likelihood functions with respect to the objective decision vector d that includes the specific component(s) to be identified. Since the $k!$ decision vectors have equal *a priori* probabilities, we define a likelihood ratio statistic Q_n for the procedure as

$$Q_n = \frac{\max_d L_n(\Omega)}{L_n(\Omega)} = \frac{L_n(\Omega_d)}{\sum_{i=1}^{k!} L_n(d_i)}, \quad (4)$$

where $L_n(d_i)$ represents the likelihood of i -th decision vector. To find the maximum of the likelihood functions with respect to the objective decision vector(s), $\max_d L_n(\Omega_d)$ in the numerator, we establish the following theorem.

Theorem 2.1 *Let $\mathbf{X} = (X_1, X_2, \dots, X_k)$ be the vector of observations for $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ given in model (1). Among $k!$ possible decision functions $L_n(d_i)$, $i = 1, 2, \dots, k!$, the maximum with respect to the specific decision vector d for identifying $\theta_{[j]}$ is given by the sum of the decision functions having term $(x_{[j]}, \theta_{[j]})$, the j -th ordered subscript for both x 's and θ 's in $L_n(\Omega)$.*

Proof. See Bechhofer, Kiefer and Sobel (1968).

Recall that the probability requirement for any decision for this problem is that it makes the correct identification with a minimum guaranteed probability at level P^* for specified $\theta = (\theta_{[1]}, \theta_{[2]}, \dots, \theta_{[k]})$ with $\theta_{[i]} - \theta_{[i-1]} \geq \delta$. Then, for

a specified desirable probability requirement P^* along with the likelihood ratio statistic given in Eq. (4), the stopping rule is as follows: Stop at N whenever

$$N = \inf \{n \geq n_0 : Q_n \geq P^*\}, \quad (5)$$

where $n_0 (\geq 2)$ is a initial sample size and $1/k! < P^* < 1$. If $P^* < 1/k!$, the requirement (2) is readily satisfied. We note that under our δ^* -condition, the proposed sequential procedure \mathcal{R} terminates with probability one with a finite number of observations since the value of n will always be finite regardless of how small a positive value is prespecified for δ^* .

Using the stopping rule (5), we now consider two types of identification for a cell configuration given in the multinomial model. One is to correctly identify a component that has the largest probability given in the model. We denote this as correct identification of the largest cell (CIL). The other is to correctly identify all components (or cells) in the model. We denote this as correct identification of all components (CIA). For instance, the first type of identification, the CIL can be applicable to any multiple decision problem that a dominant (or best) category has to be identified (or selected) in the model.

2.1 Identification of the Largest Component

In this section, we consider the case where one wishes to identify only the largest component in a given multinomial population with minimum probability P^* . Since $0 < \theta_i < 1$, we reparameterize for convenience (to maximize). Let $\eta_i = -\log \theta_i, i = 1, 2, \dots, k$ and $\eta_{[1]} \leq \eta_{[2]} \leq \dots \leq \eta_{[k]}$ denote the ordered η_i , and define $y_{[1]} = \min(x_1, x_2, \dots, x_k), \dots, y_{[k]} = \max(x_1, x_2, \dots, x_k)$. Then, from Theorem 2.1, the numerator should be the sum of likelihood functions that have the same pair in k -th largest order. After the cancellation of constant terms, the likelihood ratio statistic is then

$$Q_n = \frac{L_n(\Omega_d)}{L_n(\Omega)} = \frac{\exp \left\{ \sum_{i=1}^{k-1} \sum_{\substack{j=1 \\ i \neq j}}^{k-1} y_{[i]} \eta_{[j]} + y_{[k]} \eta_{[k]} \right\}}{\sum_{i=1}^{k!} L_n(d_i)}. \quad (6)$$

More specifically, the denominator in Eq. (6) can be written as

$$\sum_{i=1}^{k!} L_n(d_i) = \sum_{\beta \in S_k} \exp \left\{ \sum_{j=1}^k y_{[j]} \theta_{[\beta(j)]} \right\},$$

where $\beta = (\beta(1), \dots, \beta(k))$ and S_k is the set of permutations of $(1, 2, \dots, k)$.

To implement the procedure, we use Eq. (6) and the prespecified desirable probability level P^* to find the smallest value of n that satisfies Eq. (5) for identifying the largest component in the suggested procedure \mathcal{R} .

Example 2.1 Suppose we have a multinomial population with $k = 3$ and configuration $\theta = (1/2, 1/3, 1/6)$. One wishes to identify (correctly) only the

cell with the largest probability by $\delta^* = 1/6$ in the population. Since $k = 3$, there are $3! = 6$ likelihood functions available. Then, the denominator of the likelihood ratio statistic in Eq. (4) will be the sum of six likelihood functions:

$$\begin{aligned} L_n(\Omega) = & \exp\left(y_{[1]}\eta_{[1]} + y_{[2]}\eta_{[2]} + y_{[3]}\eta_{[3]}\right) + \exp\left(y_{[1]}\eta_{[1]} + y_{[2]}\eta_{[3]} + y_{[3]}\eta_{[2]}\right) \\ & + \exp\left(y_{[1]}\eta_{[2]} + y_{[2]}\eta_{[3]} + y_{[3]}\eta_{[1]}\right) + \exp\left(y_{[1]}\eta_{[2]} + y_{[2]}\eta_{[1]} + y_{[3]}\eta_{[3]}\right) \\ & + \exp\left(y_{[1]}\eta_{[3]} + y_{[2]}\eta_{[2]} + y_{[3]}\eta_{[1]}\right) + \exp\left(y_{[1]}\eta_{[3]} + y_{[2]}\eta_{[1]} + y_{[3]}\eta_{[2]}\right). \end{aligned}$$

However, since the goal is to identify the largest cell $\theta_{[3]}$, the numerator of the likelihood ratio statistic is given by the sum the likelihood functions that have $y_{[3]}\theta_{[3]}$. There can be two possible terminal decisions among six decision vectors, namely

$$L_n(\Omega_d) = \exp\left(y_{[1]}\theta_{[1]} + y_{[2]}\theta_{[2]} + y_{[3]}\theta_{[3]}\right) + \exp\left(y_{[1]}\theta_{[2]} + y_{[2]}\theta_{[1]} + y_{[3]}\theta_{[3]}\right).$$

After reparameterization through taking a logarithm, finally we have the likelihood ratio statistic

$$Q_n = \frac{\sum_{i=1, j=1, i \neq j}^2 \exp\left(y_{[j]}\eta_{[i]} + y_{[i]}\eta_{[j]} + y_{[3]}\eta_{[3]}\right)}{\sum_{i=1}^3 \sum_{j=1}^3 \sum_{k=1}^3 \substack{\exp\left(y_{[i]}\eta_{[1]} + y_{[j]}\eta_{[2]} + y_{[k]}\eta_{[3]}\right) \\ i \neq j \neq k}}. \quad (7)$$

The numerical implementation using Eq. (7) with the stopping rule given in Eq. (5) for the example is presented in Subsection 4.1.

2.2 Identification of All Components

Suppose that one wishes to identify all components $\theta_i, i = 1, 2, \dots, k$ in a given multinomial model. As mentioned in the previous section, there are $k!$ equally likely decision functions. Therefore, the denominator of the likelihood ratio statistic will be unchanged. However, since the goal is to identify all the θ 's, the objective vector must be the one with complete association among the x 's and θ 's. Then, by applying Theorem 2.1, we have

$$Q_m = \frac{\max_d L_n(\Omega)}{L_n(\Omega)} = \frac{\exp\left(\sum_{j=1}^k y_{[j]}\eta_{[j]}\right)}{\sum_{\beta \in S_k} \exp\left(\sum_{j=1}^k y_{[i]}\eta_{[\beta(i)]}\right)} \quad (8)$$

where $\beta = (\beta(1), \dots, \beta(k))$ and S_k is the set of permutations of $(1, 2, \dots, k)$.

Similarly, using Eq. (8) with a prespecified probability level P^* , the goal is to find the stopping time n that satisfies Eq. (5) for identifying all the components in a given multinomial distribution.

Example 2.2 Consider the same configuration $(1/2, 1/3, 1/6)$ given in the previous example. Suppose we want to identify (correctly) all of the components

in the multinomial model with $k = 3$. One wishes to take the cell difference $\delta^* = 1/6$ in the population with the minimum probability P^* . Again, there are $3! = 6$ frequency functions, the denominator in the likelihood ratio statistic stays the same as in Example 2.1. Since that the goal is to identify all three cells, the numerator of the likelihood ratio statistic is (after reparameterization)

$$\max_d L_n(\Omega) = \exp\left(y_{[1]}\eta_{[1]} + y_{[2]}\eta_{[2]} + y_{[3]}\eta_{[3]}\right).$$

Thus, the statistic for the procedure we have

$$Q_n = \frac{\exp\left(\sum_{i=1}^3 \eta_{[i]} y_{[i]}\right)}{\sum_{i=1}^3 \sum_{\substack{j=1 \\ i \neq j}}^3 \sum_{k=1}^3 \exp\left(\eta_{[1]} y_{[i]} + \eta_{[2]} y_{[j]} + \eta_{[3]} y_{[k]}\right)}. \quad (9)$$

The numerical results using Eq. (9) for this example are given in Subsection 4.2.

3 Dirichlet Integrals and Multinomial Events

There are two types of Dirichlet distributions, whose incomplete integrals are useful tools for analyzing multinomial events. A b -variate random vector $\mathbf{x} = (x_1, x_2, \dots, x_b)$ is said to have a Dirichlet-type II distribution (also called the inverted beta distribution) with parameters $(\mathbf{r}; m) = (r_1, r_2, \dots, r_b; m)$ if the joint probability density is given by

$$f^{(b)}(\mathbf{x}; \mathbf{r}, m) = \frac{\Gamma(m + br)}{\Gamma(m) \prod_{i=1}^b \Gamma(r_i)} \frac{\prod_{i=1}^b x_i^{r_i - 1} dx_i}{\left(1 + \sum_{i=1}^b x_i\right)^{m+br}} \quad (10)$$

over the b -dimensional positive orthant $\mathfrak{S}^b = \{(x_1, x_2, \dots, x_b) : x_i \geq 0, i = 1, 2, \dots, b\}$ and zero outside \mathfrak{S}^b .

For the identification problem in a given multinomial population, our main concern in the procedure is that the largest cell reaches the stopping time m in sampling. We define an event E to be the maximum frequency among the cells at stopping time (a.s.t.). Let $b = k - 1$ and denote the frequency of the i -th cell by $q_i, i = 1, 2, \dots, b$. Let the cell probabilities be p_0 for the counting cell and p_i for the blue cells. Then, using the results shown in Olkin and Sobel (1965), we have

$$\begin{aligned} P\{E\} &= P\{q_i < r \text{ a.s.t. when } q_{b+1} = m \text{ for the first time}\} \\ &= \int_{a_b}^{\infty} \int_{a_{b-1}}^{\infty} \dots \int_{a_1}^{\infty} f^{(b)}(\mathbf{x}; \mathbf{r}, m) \prod_{i=1}^b dx_i \\ &\equiv D_{\mathbf{a}}^{(b)}(\mathbf{r}; m) \end{aligned} \quad (11)$$

where $\mathbf{a} = (a_1, a_2, \dots, a_b)$ and $a_i = p_i/p_0$.

We want the probability that when a (specified) counting cell reaches m and all the remaining cells have frequency $< r$. The D function in (11) can be expressed in two forms, which are given in Sobel, Uppuluri and Frankowski (1985);

$$D_{\mathbf{a}}^{(b)}(\mathbf{r}; m) = \frac{1}{(1+ba)^m} \sum_{x_1 < r} \cdots \sum_{x_b < r} \binom{m-1 + \sum_{i=1}^b x_i}{m-1, x_1, \dots, x_b} \left(\frac{a}{1+ba} \right)^{x_1 + \cdots + x_b} \quad (12)$$

and, to obtain recurrence relations we introduce a new parameter j ($0 \leq j \leq b$)

$$D_{\mathbf{a}}^{(b,j)}(\mathbf{r}; m) = \frac{\Gamma(m+br)}{\Gamma(m) \prod_{i=1}^b \Gamma(r_i)} \left(\frac{a^r}{r} \right)^j \int_{a_b}^{\infty} \int_{a_{b-1}}^{\infty} \cdots \int_{a_1}^{\infty} \frac{\prod_{i=j+1}^b x_i^{r_i-1} dx_i}{\left(1 + ja + \sum_{i=j+1}^b x_i\right)^{m+br}}. \quad (13)$$

Then, we modify Eqs. (12)-(13) for our purposes and taking transformations $y_i = x_i/a_i$ we obtain

$$D_{a,a_0}^{(b,j)}(\mathbf{r}, \mathbf{y}; m) = \left[\frac{1}{1 + ja_0 + (b-j)a} \right]^m \sum_{\substack{0 \leq x_{j+1} < r \\ \vdots \\ 0 \leq x_{b-j} < r}} \binom{(m-1) + \sum_{i=1}^j y_i + x_0 + \sum_{i=1}^{b-1} x_i}{(m-1), \sum_{i=1}^j y_i, x_0, x_1, \dots, x_{b-1}} \\ \times \left[\frac{a_0}{1 + ja_0 + (b-j)a} \right]^{y_1 + \cdots + y_j} \left[\frac{a}{1 + a_0 + (b-1)a} \right]^{x_{j+1} + \cdots + x_{b-1}} \quad (14)$$

and

$$D_{\mathbf{a}}^{(b)}(\mathbf{r}; m) = \frac{\Gamma \left[m + \sum_{i=1}^j y_i + (b-j)r \right]}{\Gamma(m) \Gamma^{b-1}(r) \prod_{i=1}^j \Gamma(y_i)} \left(\frac{a_0^{y_1 + \cdots + y_j}}{y_1 \cdots y_j} \right) \\ \times \underbrace{\int_a^{\infty} \int_a^{\infty} \cdots \int_a^{\infty}}_{(b-j) \text{ integrals}} \frac{\prod_{\alpha=1}^{b-1} x_{\alpha}^{r-1} dx_{\alpha}}{\left(1 + ja_0 + \sum_{i=1}^{b-1} x_i\right)^{m+r_0+(b-1)r}}, \quad (15)$$

respectively where $\mathbf{y} = (y_1, y_2, \dots, y_j)$. We will use Eqs. (14)-(15) to express the probability of identification in multinomial events.

3.1 P{CIL} USING DIRICHLET INTEGRALS

In this section, we study the expression of $P\{CIL\}$ in the procedure using incomplete Dirichlet integrals. As in Example 2.1, we consider the case where $k = 3$ with the configuration $(1/2, 1/3, 1/6)$. Suppose that one wishes to identify the largest cell with $P^* = 0.95$ and $\delta^* = 1/6$.

Then, the goal is to find the smallest integer (sample number) m such that $P\{CIL\} \geq P^*$. Since our job is to identify the only largest cell, we must find the smallest integer m such that the probability that the frequency associated with $\theta_{[3]}$ is larger than the frequencies associated with $\theta_{[2]}$ and $\theta_{[1]}$ at the stopping time is greater than or equal to P^* . This is the same as the probability that when one specified cell reaches the frequency m , the frequencies of the remaining two cells (i.e., inferior cells) are smaller than m .

Since $k = 3$, $b = 2$, define two ratios in Eq. (14) as follows:

$$a_0 = \frac{\theta_{[1]}}{\theta_{[3]}} = \frac{1}{3}, \quad a_1 = \frac{\theta_{[2]}}{\theta_{[3]}} = \frac{2}{3}.$$

Then, using Eq. (14) we have $P\{CIL\}$ of the form

$$P\{CIL\} = \sum_{x=1}^{m-1} D_{\frac{1}{3}, \frac{2}{3}}^{(2,1)}(x, x; m) + \sum_{x=0}^{m-1} \left[D_{\frac{1}{3}, \frac{2}{3}}^{(2,1)}(x+1, x; m) - D_{\frac{1}{3}, \frac{2}{3}}^{(2,1)}(x, x; m) \right]. \quad (16)$$

After simplifying the compound D -functions in Eq. (16), the following expression is obtained:

$$P\{CIL\} = \sum_{r=1}^{m-1} \binom{m+r-1}{r} \left(\frac{2}{5}\right)^r \left(\frac{3}{5}\right)^m D_{\frac{1}{5}}^{(1)}(r, m+r). \quad (17)$$

Finally, after some algebra (for computational purpose), it follows from Eq. (16) that

$$P\{CIL\} = 2^{-m} \sum_{j=1}^{m-1} \left(\frac{1}{3}\right)^j \sum_{i=0}^{j-1} \binom{m-1+i+j}{m-1, i, j} \left(\frac{1}{6}\right)^i + 2^{-m} \sum_{j=0}^{m-1} \binom{m-1+2j}{m-1, j, j} \left(\frac{1}{3} \cdot \frac{1}{6}\right)^j, \quad (18)$$

where m represents the integer needed for stopping time, r is a dummy variable for $\theta = 1/3$ and s is a dummy variable for $\theta = 1/6$.

Tabled values for the D -function in Eq. (17) are available in Sobel, Uppuluri and Frankowski (1985). We use Eq. (18) to solve for m , the smallest integer value of observations m which satisfies $P\{CIL\} \geq P^*$. The values of the optimal stopping time m^* are presented in Table 3.1.

P^*	0.75	0.80	0.85	0.90	0.95	0.975	0.990	0.995
m^*	13	17	22	31	43	55	74	86

4 Numerical Studies

Monte Carlo experimentation is carried out to illustrate the behavior and performance of the stopping rule in the proposed sequential procedure. We use the configuration $(1/2, 1/3, 1/6)$ in a multinomial model with two cases: (a) to identify the largest (or dominant) component in the model; and (b) to identify all components (or cells) in a given configuration.

4.1 Identifying a Dominant Component

Using the stopping rule in Eq. (5), we find the expected stopping time for given observations (x_1, x_2, x_3) by solving the inequality $Q_n \geq P^*$ for n , which will be the smallest (integer) value that satisfies the inequality. The results of the Monte Carlo experimentation are presented in Table 4.1. In the table, every value in the rows is based on 10,000 independent replications with various nominal probabilities P^* . The average stopping time is denoted by \bar{n} , the coverage probability CP, and the standard error S.E. are shown for each of pre-determined value P^* .

Table 4.1 Identifying Largest Cell for $k = 3$
Configuration $(1/2, 1/3, 1/6)$, $\delta^* = 1/6$

P^*	\bar{n}	CP	S.E.
0.75	12.23	0.7829	0.0817
0.80	17.53	0.8367	0.1265
0.85	24.23	0.8920	0.1747
0.90	31.21	0.9247	0.2223
0.95	43.67	0.9607	0.3149
0.975	56.20	0.9814	0.3783
0.990	72.59	0.9923	0.4358
0.995	83.61	0.9969	0.4924
0.999	107.69	0.9990	0.5736

From Table 4.1, depending upon the desirable probability level P^* , we can construct a $(1 - \alpha) 100\%$ confidence interval for the true number of required samples as $\bar{n} \pm z_{\alpha/2} S.E.$. Suppose that we want to have a 90% confidence interval with $P^* = 0.95$. Then, the 90% confidence interval is given by $43.67 \pm 1.645 (0.3149) = (43.15, 44.19)$. So, with 90% confidence we need a sample size between 43 and 45 to correctly identify the dominant component in the model. We also note that the values of \bar{n} in the above table are consistent with the optimal values of stopping time m^* shown in Table 3.1.

The following table presents various cell configurations when the dominant cell exists in the model.

Table 4.2 Identifying Largest Cell (bold faced) in a Multinomial Population

Cell Configuration ($\theta_1, \theta_2, \theta_3$)	δ^*	$P^* = 0.90$		0.95		0.99	
		CP	\bar{n}	CP	\bar{n}	CP	\bar{n}
(a) (1/2 , 1/3, 1/6)	1/6	0.9247	31.21	0.9607	43.67	0.9921	71.73
(b) (4/7 , 2/7, 1/7)	1/7	0.9329	12.16	0.9720	17.25	0.9931	24.99
(c) (1/2 , 5/14, 1/7)	1/7	0.9166	41.15	0.9558	57.67	0.9929	97.09
(d) (5/8 , 2/8, 1/8)	1/8	0.9457	7.56	0.9773	10.60	0.9911	14.34
(e) (1/2 , 3/8, 1/8)	1/8	0.9062	52.08	0.9593	80.45	0.9908	124.34

As seen in Table 4.2, from the facts that the coverage probability in configuration (c) is lower than the one in (b), and the expected sample size in configuration (c) is larger than the one in (b), we can say that configuration (c) is less favorable than configuration (b) under the same distance measure δ^* . Similarly, for distance measure $\delta^* = 1/8$, configuration (e) is less favorable than configuration (d) in the identification procedure. It also needs to note that $\min \rho^{(e)} = 1.33 < \min \rho^{(d)} = 2.5$. In fact, configuration (e) seems to be the δ^* -least favorable configuration (LFC) in identifying the dominant cell among all possible configurations with $\delta^* = 1/8$.

Example 4.1 In an ecological study, suppose that the occupancy rates of three species in an area are known to be 50%, 37.5%, and 12.5%, respectively. A researcher wants to know that how many samples are required to be 90% confident that the occupancy rate of the dominant species is 50% by a difference of at least 12.5% from the other species. The configuration of the species is (1/2, 3/8, 1/8), as in configuration (a) of Table 4.2 with $\delta^* = 1/8$. Then, the researcher needs approximately at least 53 samples (on average) to achieve 90% confidence.

4.2 Identifying All Components

To identify all components given in the model for observations (x_1, x_2, x_3) , we use the stopping rule given in Eq. (5) to find the expected stopping time. The results of the Monte Carlo experimentation are presented in Table 4.3. In the table, every value in the rows is based on 10,000 independent replications with various nominal probabilities P^* .

Table 4.3 Identifying all cells for $k = 3$
 Configuration $(1/2, 1/3, 1/6)$, $\delta^* = 1/6$

P^*	\bar{n}	CP	S.E.
0.75	23.54	0.7696	0.1477
0.80	31.10	0.8324	0.2069
0.85	36.79	0.8712	0.2501
0.90	45.03	0.9172	0.3133
0.95	59.10	0.9589	0.4023
0.975	72.23	0.9832	0.4920
0.990	85.34	0.9916	0.5562
0.995	97.06	0.9954	0.5959
0.999	120.17	0.9992	0.6756

From the above table, by constructing a $(1 - \alpha)$ 100% confidence interval for the true number of samples, $\bar{n} \pm z_{\alpha/2} S.E.$, we can make a decision of the number of samples required to identify all three components correctly with our desired P^* . Suppose that one wishes to have the probability level $P^* = 0.95$, the 95% confidence interval for true number of sample sizes is given by $59.10 \pm 1.96(0.4023) = (58.31, 59.89)$. That is, the approximate number of required samples is between 58 and 60 to identify all three cells for the given configuration.

In the next experiment, we consider five different configurations for $k = 3$ in Π_i ; (a) $(1/2, 1/3, 1/6)$, (b) $(4/7, 2/7, 1/7)$, (c) $(1/2, 5/14, 1/7)$, (d) $(5/8, 2/8, 1/8)$, and (e) $(1/2, 3/8, 1/8)$. We calculated the coverage probability (CP) and the average stopping time \bar{n} for each of required probability levels $P^* = 0.90, 0.95$ and 0.99 . The results of the Monte Carlo simulation are summarized in Table 4.4 below.

Table 4.4 Identifying All Components in Population $\Pi_i, i = 3$

Cell Configuration $(\theta_1, \theta_2, \theta_3)$	δ^*	$P^* = 0.90$		0.95		0.99	
		CP	\bar{n}	CP	\bar{n}	CP	\bar{n}
(a) $(1/2, 1/3, 1/6)$	1/6	0.9172	45.03	0.9589	59.10	0.9916	85.34
(b) $(4/7, 2/7, 1/7)$	1/7	0.9288	31.56	0.9626	39.68	0.9928	55.10
(c) $(1/2, 5/14, 1/7)$	1/7	0.9112	48.23	0.9506	64.07	0.9926	100.89
(d) $(5/8, 2/8, 1/8)$	1/8	0.9314	31.02	0.9674	40.62	0.9910	57.18
(e) $(1/2, 3/8, 1/8)$	1/8	0.9112	57.09	0.9604	81.48	0.9898	126.18

From the above table, we observe that all of the coverage probabilities are greater than the nominal probability P^* . By comparing two configurations (e.g. (b) and (c), or (d) and (e)) with the same distance measure δ^* , we can surmise that if there are more dominant components (or cells) in the configuration with the same distance measure δ^* , then they produce an increase in the number of observations for identification. On the contrary, the subdominant components contribute to identifying the difference sooner.

Moreover, since the coverage probability in configuration (c) is lower than the one in (b) and the expected sample size in configuration (c) is larger than the

one in (b), we can say that configuration (c) is less favorable than configuration (b) under the same distance measure δ^* . For the slippage ratio, we also observe that $\min \rho_{(c)} = 1.4 < \min \rho_{(b)} = 2$. Under the discerning value $\delta^* = 1/8$, the configuration (e) is less favorable than configuration (d) in the identification problem. When $\delta^* = 1/8$ in a given model, configuration (e) seems to be the least favorable configuration (LFC) in identifying all three cells.

From Tables 4.2 and 4.4, we observe that the coverage probability (CP) is uniformly (slightly) higher than the nominal probability level P^* . We also see the average stopping time \bar{n} tends to be large as the probability requirement level P^* increases. From these facts, we conclude that the simulation results provide substantial numerical evidence to support that the proposed sequential procedure \mathcal{R} performs satisfactorily.

5 Concluding Remarks

We have considered statistical identification procedures using sequential sampling for multinomial models with a pre-determined discerning value δ^* using the indifference-zone approach. We devised the procedure \mathcal{R} using likelihood ratios and found the optimal sample sizes from a decision-theoretic point of view. We also used the Dirichlet integral to obtain the required sample sizes for the proposed procedure. The proposed method is well verified through Monte Carlo experimentation. We conclude that the proposed procedure is reliable (in terms of P^*) and easily implemented (in terms of the likelihood ratio statistic Q). The procedure can be extended to other settings of configurations in multinomial models for $k \geq 4$ as well.

For future study, the identification procedure can be extended to multivariate data. For instance, suppose that we have k populations (denoted by $\Pi_i, i = 1, 2, \dots, k$) with (unknown) means $\theta_1, \theta_2, \dots, \theta_k$, respectively, and a common (known) variance σ^2 . The k population can be regarded as k component, and denote $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k$ to be the sample means from these populations. Then, by taking distance measure to be $\mu_{[j]} - \mu_{[j-1]} \geq \delta^*, j = 2, 3, \dots, k$, the problem of identifying the pair (\bar{X}_i, θ_i) as being associated with the population $\Pi_{[i]}$ will turn out to be equivalent to the conventional classification or clustering procedure.

Acknowledgment

The author wishes to thank anonymous referees for their valuable suggestions and helpful comments that improved the paper.

References

- [1] Bechhofer, R. E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *The Annals of Mathematical Statistics* 25, 16-39.
- [2] Bechhofer, R. E., Dunnett, C. W. and Sobel, M. (1954). A two-sample multiple decision procedure for ranking means of normal populations with common unknown variances. *The Annals of Mathematical Statistic* 25, 16-39.
- [3] Bechhofer, R. E., Kiefer, J. and Sobel, M. (1968). *Sequential Identification and Ranking Procedures*. The University of Chicago Press, Chicago.
- [4] Cho, H. (2003). Inverse-type sampling procedure for estimating the number of multinomial classes. *Sequential Analysis* 22, 307-324.
- [5] Gibbons, J., Olkin, I., and Sobel, M. (1977). *Selecting and Ordering Populations: A New Statistical Methodology*. John Wiley & Sons, New York.
- [6] Gupta, S. S. (1956). On a decision rule for ranking means. *Institute of Statistics. Mimeo Series No. 150*, University of North Carolina, Chapel Hill.
- [7] Gupta, S. S. and Huang, W. T. (1981). On mixtures of distributions: A survey and some new results on ranking and selection. *Journal of the Indian Statistical Association B* 43, 245-290.
- [8] Gupta, S. S. and Panchapakesan, S. (1979). *Multiple Decision Procedure: Theory and Methodology of Selecting and Ranking Populations*. John Wiley & Sons, New York.
- [9] Huang, D. and Panchapakesan, S. (1978). The complete ranking problem with a subset selection goal. *Journal of the Chinese Statistical Association* 16 (1), 5801-5810.
- [10] Olkin, I. and Sobel, M. (1965). Integral expression for tail probabilities of the multinomial and negative multinomial distributions, *Biometrika* 52, 167-179.
- [11] 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.