Ann. Inst. Statist. Math.
37 (1985), Part A, 303-314

SUBSET SELECTION FOR THE LEAST PROBABLE MULTINOMIAL CELL

PINYUEN CHEN

(Received Oct. 28, 1983; revised Mar. 5, 1985)

Summary

An inverse sampling procedure R is proposed for selecting a random-size subset which contains the least probable cell (i.e., the cell with the smallest cell probabilities) from a multinomial distribution with k cells. Type 2-Dirichlet integrals are used (i) to express the probability of a correct selection in terms of integrals with parameters only in the limits of integration, (ii) to prove that the least favorable configuration under R is the so-called slippage configuration with k equal cell probabilities, and (iii) to express exactly the expectation of the total number of observations required and the expectation of the subset size under the procedure R.

1. Introduction

The problem of selecting among multinomial cells has been studied in many articles by using either indifference zone formulation or subset selection formulation. Under the indifference zone formulation, Bechhofer, Elmaghraby and Morse [2] studied a fixed-sample-size procedure for selecting the largest cell probability from a multinomial distribution and Cacoullos and Sobel [3] studied an inverse-sampling procedure for the same selecting goal. Under the subset selection formulation, Gupta and Nagal [4] and Panchapakesan [5], [6] considered the goal of selecting a random-size subset which contains the cell with the largest cell probability by using fixed-sample-size and inverse-sampling procedure respectively. Alam and Thompson [1] studied the selection goal for the smallest multinomial cell probability. They proposed a fixed-sample-size procedure under the indifference zone formulation and used "difference" for the measure of distance. In this paper, we propose an inverse sampling procedure and use "ratio" as the measure of distance to select

Key words and phrases: Dirichlet integral, inverse sampling, least favorable configuration, multinomial distribution, subset selection formulation.

a random-size subset which contains the cell associated with the smallest cell probability.

Let $p_1, \cdots, p_k$ be the cell probabilities in the multinomial distribution with $\sum_{i=1}^{k} p_i = 1$. The ordered values of the cell probabilities are denoted by $p_{[1]}, \cdots, p_{[k]}$ and the cell associated with $p_{[i]}$ is denoted $\Pi_{[i]}$, $i=1, \cdots, k$. The goal of the experimenter is to select a subset containing the cell $\Pi_{[1]}$. In the case of a tie, one of the cells with probability $p_{[1]}$ will be considered as the best. A correct selection (CS) is defined as the selection of any random-size subset which includes the best cell $\Pi_{[1]}$. Under the usual subset selection formulation, we seek a procedure R such that

$$\mathrm{P}(\mathrm{CS} \mid R) \geqq P^* \tag{1.1}$$

where $\mathrm{P}(\mathrm{CS} \mid R)$ denotes the probability of a correct selection using the procedure R and $P^*(1/k < P^* < 1)$ is a specified probability level.

The procedure R is defined in Section 2 and the $\mathrm{P}(\mathrm{CS} \mid R)$ is expressed exactly in terms of Type 2-Dirichlet integrals. Section 2 also discusses the infimum of the $\mathrm{P}(\mathrm{CS} \mid R)$ over the parameter space. It is shown that the infimum is attained for the configuration $(1/k, \cdots, 1/k)$, which is called the least favorable configuration. Section 3 discusses the exact results for $\mathrm{E}(N)$, the expected number of observations required to reach a decision. An exact formula for the expected subset size $\mathrm{E}(S)$ is also given in terms of Type 2-Dirichlet integrals in Section 3.

The main tool used in this paper is the fact that the $\mathrm{P}(\mathrm{CS} \mid R)$ can be written in terms of Type 2-Dirichlet integrals. This turns out to be highly useful because it is exact and because the p-values show up only in the limits of integrations.

2. The procedure R and the least favorable configuration

Procedure R: Continue sampling one-at-a-time until either

(1) One cell reaches a frequency of r, or

(2) $(k-1)$ cells reach frequencies of at least r' $(1 \leqq r' \leqq r+1)$.

As soon as (1) occurs before (2), we stop and select the cell with the frequencies x_i where $x_i < r'$. As soon as (2) occurs before (1), we stop and select the cell with the frequency $< r'$. The constants r and r' are chosen so as to satisfy the basic probability requirement (1.1).

It should be noted that we can always find a pair (r, r') which satisfies the P^*-requirement since $\mathrm{P}(\mathrm{CS} \mid R)$ increases in r' for every r and increases to one when $r' = r+1$.

Before we give the exact expression for $\mathrm{P}(\mathrm{CS} \mid R)$, we need the following probability interpretation of Dirichlet integrals.

Consider a multinomial distribution with $b+1$ cells; independent and identically distributed observations are taken. We regard the observations as falling into one of the $b+1$ cells $\Pi_0, \Pi_1, \cdots, \Pi_b$ with respective probabilities $p_0, p_1, \cdots, p_b$, $\sum_{i=0}^{b} p_i = 1$. Observations are taken one-at-a-time until cell Π_0 contains exactly r_0 observations; the first time this occurs is called the stopping time. Denote the random number of observations in cell Π_α at stopping time by X_α $(\alpha=1, \cdots, b)$ and consider at the time of stopping the compound event E defined by

$$\text{(2.1)} \quad \begin{aligned} E:\ & X_0 = r_0 \text{ at stopping time, } X_\alpha = r_\alpha \ (\alpha=1, \cdots, j), \\ & X_\alpha < r_\alpha \ (\alpha = j+1, \cdots, b) . \end{aligned}$$

In terms of multinomial sums, we clearly have

$$\text{(2.2)} \qquad \mathrm{P}(E) = p_0^{r_0} \prod_{\alpha=1}^{j} \left(\frac{p_\alpha^{r_\alpha}}{r_\alpha !} \right) \sum_{x_{j+1}=0}^{r_{j+1}-1} \cdots \sum_{x_b=0}^{x_b-1} \frac{\Gamma\left(r_0 + \sum_{\alpha=1}^{j} r_\alpha + \sum_{\alpha=j+1}^{b} x_\alpha\right)}{\Gamma(r_0)} \prod_{\alpha=j+1}^{b} \frac{p_\alpha^{x_\alpha}}{x_\alpha !} .$$

It has been shown in Sobel, Uppuluri and Frankowski [7] that the multinomial sum in (2.2) can be written in terms of Dirichlet integral $CD_{\boldsymbol{a}}^{(b-j;j)}(\boldsymbol{r}, r_0)$, i.e.,

$$\text{(2.3)} \quad \begin{aligned} \mathrm{P}(E) &= CD_{\boldsymbol{a}}^{(b-j;j)}(\boldsymbol{r}, r_0) \\ &= \frac{\Gamma(r_0+R)}{\Gamma(r_0) \prod_{\alpha=1}^{b} \Gamma(r_\alpha)} \left[\prod_{\alpha=1}^{j} \left(\frac{a_\alpha^{r_\alpha}}{r_\alpha} \right) \int_{a_b}^{\infty} \cdots \right. \\ &\qquad \left. \int_{a_{j+1}}^{\infty} \frac{\prod_{\alpha=j+1}^{b} x_\alpha^{r_\alpha - 1} dx_\alpha}{\left(1 + \sum_{\alpha=1}^{j} a_\alpha + \sum_{\alpha=j+1}^{b} x_\alpha\right)^{r_0+R}} \right] , \end{aligned}$$

where $R = \sum_{\alpha=1}^{b} r_\alpha$, $\boldsymbol{r} = (r_1, \cdots, r_b)$, $\boldsymbol{a} = (a_1, \cdots, a_b) = (p_1/p_0, \cdots, p_b/p_0)$.

When $j=0$ in the event E in (2.1), we can simply drop the superscript and write

$$\text{(2.4)} \quad E_1:\ x_0 = r_0 \text{ at stopping time, and } x_\alpha < r_\alpha \ (\alpha = 1, \cdots, b)$$

and

$$\text{(2.5)} \quad \mathrm{P}(E_1) = CD_{\boldsymbol{a}}(\boldsymbol{r}, r_0)$$

$$=\frac{\Gamma(r_0+R)}{\Gamma(r_0)\prod_{\alpha=1}^{b}\Gamma(r_\alpha)}\left[\int_{a_b}^{\infty}\cdots\int_{a_1}^{\infty}\frac{\prod_{\alpha=1}^{b}x_\alpha^{r_\alpha-1}dx_\alpha}{\left(1+\sum_{\alpha=1}^{b}x_\alpha\right)^{r_0+R}}\right]$$

with the same R, $\mathbf{r}$, $\boldsymbol{a}$ as defined in (2.3).

In the $CD_{\boldsymbol{a}}(\mathbf{r}, r_0)$ integral, consider $a_i=p_i/p_0$ as a variable (say $a_i=x$) and all the other parameters as constants, we have the following lemma.

LEMMA 2.1. (1) *The derivative of* $CD_{\boldsymbol{a}}(\mathbf{r}, r_0)$ *with respect to* x *is equal to* $-r_i/x\,\mathrm{P}(E_2)$ *where*

E_2: $x_0=r_0$ *at stopping time,* $x_i=r_i$ *and* $x_\alpha<r_\alpha$ $(\alpha=1,\cdots,b;\ \alpha\neq i)$.

(2) *The derivative of* $CD_{\boldsymbol{a}}(\mathbf{r}, r_0)$ *with respect to* $1/x$ *is equal to* $r_0x\,\mathrm{P}(E_3)$ *where*

E_3: $x_i=r_i$ *at stopping time,* $x_0=r_0$ *and* $x_\alpha<r_\alpha$ $(\alpha=1,\cdots,b;\ \alpha\neq i)$.

PROOF. (1) Consider $CD_{\boldsymbol{a}}(\mathbf{r}, r_0)$ in (2.5) as a function of $x=p_i/p_0$, we have by (2.4) and (2.5)

$$\frac{dCD_{\boldsymbol{a}}(\mathbf{r}, r_0)}{dx}=\frac{\Gamma(r_0+R)}{\Gamma(r_0)\prod_{\alpha=1}^{b}\Gamma(r_\alpha)}(-x^{r_i-1})\int_{a_b}^{\infty}\cdots\int_{\substack{a_j\\(j\neq i)}}^{\infty}\cdots\int_{a_1}^{\infty}\frac{\prod_{\substack{\alpha=1\\ \alpha\neq i}}^{b}x_\alpha^{r_\alpha-1}dx_\alpha}{\left(1+x+\sum_{\substack{\alpha=1\\ \alpha\neq i}}^{b}x_\alpha\right)^{r_0+R}}$$

$$=-\frac{r_i}{x}\mathrm{P}(E_2)\,.$$

(2) Consider $CD_{\boldsymbol{a}}(\mathbf{r}, r_0)$ in (2.5) as a function of $y=1/x=p_0/p_i$, we have

$$\text{(2.6)}\quad \frac{dCD_{\boldsymbol{a}}(\mathbf{r}, r_0)}{dy}$$

$$=\frac{\Gamma(r_0+R)}{\Gamma(r_0)\prod_{\alpha=1}^{b}\Gamma(r_\alpha)}\left(\frac{1}{y}\right)^{r_i+1}\int_{a_b}^{\infty}\cdots\int_{\substack{a_j\\(j\neq 1)}}^{\infty}\cdots\int_{a_1}^{\infty}\frac{\prod_{\substack{\alpha=1\\ \alpha\neq i}}^{b}x_\alpha^{r_\alpha-1}dx_\alpha}{\left(1+y+\sum_{\substack{\alpha=1\\ \alpha\neq i}}^{b}x_\alpha\right)^{r_0+R}}\,.$$

Let $y_\alpha=y\cdot x_\alpha$ for $\alpha=1,\cdots,b$ and $\alpha\neq i$. (2.6) becomes

$$\frac{dCD_{\boldsymbol{a}}(\mathbf{r}, r_0)}{dy}$$

$$=\frac{\Gamma(r_0+R)}{\Gamma(r_0)\prod_{\alpha=1}^{b}\Gamma(r_\alpha)}y^{r_0-1}\int_{p_b/p_i}^{\infty}\cdots\int_{p_j/p_i}^{\infty}\cdots\int_{p_1/p_i}^{\infty}\frac{\prod_{\substack{\alpha=1\\ \alpha\neq i}}^{b}y_\alpha^{r_\alpha-1}dy_\alpha}{\left(1+y+\sum_{\substack{\alpha=1\\ \alpha\neq i}}^{b}y_\alpha\right)^{r_0+R}}$$

$$=\frac{r_0}{y}\,\mathrm{P}(E_3)=r_0x\,\mathrm{P}(E_3)\qquad \text{by (2.4) and (2.5)}\,.$$

Remark 2.1. Consider the case $k=b+1$ and let x_i ($i=1,\cdots,k$) denote the number of observations in cell Π_i at stopping time. Let E_4, E_5, E_6 be the following events:

$$E_4:\ \ x_1=r \text{ ast}\,;\ x_\alpha<r_\alpha\ (\alpha=2,\cdots,k-1)\,;\ x_k<r'\,.$$

$$E_5:\ \ x_k=r' \text{ ast}\,;\ x_1<r\,;\ x_\alpha<r_\alpha\ (\alpha=2,\cdots,k-1)\,.$$

$$E_6:\ \ x_1=r \text{ ast}\,;\ x_k=r'\,;\ x_\alpha<r_\alpha\ (\alpha=2,\cdots,k-1)\,.$$

If $p_k/p_1=x$, then $d\,\mathrm{P}(E_4)/dx=-r'(p_1/p_k)\mathrm{P}(E_6)=-d\,\mathrm{P}(E_5)/dx$.

Remark 2.2. We will apply Lemma 2.1 to the differentiation of *CD* integrals in the next theorem. As we will see in the proof of the theorem and the following illustrative example, the variable x appears not only in one limit of the integral piece. When this is the case, we should consider taking derivative one integral piece at a time and keep the other integral pieces fixed like differentiating a product of two functions. These are the main properties we will use in proving the next theorem.

THEOREM 2.1. *The least favorable configuration under the procedure R is given by*

$$p_{[1]}=\cdots=p_{[k]}=\frac{1}{k}\,. \tag{2.7}$$

PROOF. We only have to consider the case $r'\leqq r$. The case that $r'=r+1$ will always give $\mathrm{P}(\mathrm{CS}\,|\,R)=1$. Consider the configuration

$$p_{[1]}\leqq p_{[2]}\leqq\cdots\leqq p_{[k]}\qquad \text{where } \sum_{i=1}^{k}p_{[i]}=1\,. \tag{2.8}$$

Let $S_n=\{p_{[1]},\cdots,p_{[n]}\}$, $T_n=\{p_{[n+1]},\cdots,p_{[k]}\}$ be two sets. Fix all the ratios of the elements in S_n and T_n respectively and let $p_{[n+1]}/p_{[n]}=x$ be the only variable in any of the *CD*-integrals under the configuration (2.8). We now prove that $\mathrm{P}(\mathrm{CS}\,|\,R)$ is a non-decreasing function of x for every $n\in\{2,3,\cdots,k\}$.

Let x_i denote the frequency in the cell with cell probability $p_{[i]}$, $i=1,\cdots,k$.

Then the probability of a correct selection under procedure R can be written as

$$\mathrm{P}(\mathrm{CS}\,|\,R)=\sum_{i=2}^{k}\mathrm{P}\,(x_i=r \text{ ast; at least one of } \{x_2,\cdots,x_k\} \tag{2.9}$$

is less than r'; $x_1<r'$)$+\sum_{i=2}^{k}$P($x_i=r'$ ast;
$r'\leqq x_j<r(j\neq i,\ j\neq 1)$; $x_1<r'$).

For $i, j=2,\cdots,k$ $(i\neq j)$, let $A_{i,j}$, B_i, $B_{i,j}$ be the following events:

$A_{i,j}$: $x_i=r$ ast; $x_j<r'$; $x_1<r'$; $x_m<r$ for all $m\neq 1, i, j$.

B_i: $x_i=r'$ ast; $x_j<r$ for $j\neq i$, $j\neq 1$; $x_1<r'$.

$B_{i,j}$: $x_i=r'$ ast; $x_j<r'$; $x_1<r'$; $x_m<r$ for all $m\neq 1, i, j$.

Then (2.9) can be written as

$$\text{(2.10)}\qquad \mathrm{P}(\mathrm{CS}\,|\,R)=\sum_{i=2}^{k}\mathrm{P}\Big(\bigcup_{\substack{j\neq i\\ j\neq 1}}A_{i,j}\Big)+\sum_{i=2}^{k}\mathrm{P}\Big(B_i-\bigcup_{\substack{j\neq i\\ j\neq 1}}B_{i,j}\Big)\,.$$

By the inclusion-exclusion principle, we have

$$\text{(2.11)}\qquad \mathrm{P}(\cup A_{i,j})=\sum\mathrm{P}(A_{i,j})-\sum_{j_\alpha\neq j_\beta}\mathrm{P}(A_{i,j_\alpha}\cap A_{i,j_\beta})+\cdots$$
$$+(-1)^{k-1}\mathrm{P}(\cap A_{i,j})\,,$$

$$\text{(2.12)}\qquad \mathrm{P}(B_i-\cup B_{i,j})=\mathrm{P}(B_i)-\Big[\sum_{j\neq i}\mathrm{P}(B_{i,j})-\sum_{j_\alpha\neq j_\beta}\mathrm{P}(B_{i,j_\alpha}\cap B_{i,j_\beta})$$
$$+(-1)^{k-1}\mathrm{P}(\cap B_{i,j})\Big]\,.$$

Notice that the terms on the right hand side of (2.11) are always in the form of E_4 in Remark 2.1 and the terms on the right hand side of (2.12) are always in the form of E_5 in Remark 2.1. For every negative term that appears in the derivative of P(CS$|R$) with respect to x, we can always find a positive term in the derivative with the same absolute value. For example, the term in the derivative of P$(A_{i,j_1}\cap A_{i,j_2}\cap\cdots\cap A_{i,j_h})$ corresponding to the integral with lower limit $p_{j_1}/p_i=p_{[n]}/p_i\cdot p_{j_1}/p_{[n+1]}\cdot x$ (where $i\in S_n$, $j_1\in T_n$) is equal to the negative of the term in the derivative of P$(B_{j_1,j_2}\cap B_{j_1,j_3}\cap\cdots\cap B_{j_1,j_n})$ corresponding to the integral with lower limit $p_i/p_{j_1}=p_i/p_{[n]}\cdot p_{[n+1]}/p_{j_1}\cdot 1/x$. From the fact that the cells $\pi_{[2]},\cdots,\pi_{[k]}$ serve as counting cells with either frequency r or frequency r' at stopping time in the P(CS$|R$), we can cancel all the CD-integrals (which are in the form of E_6) in the derivative of the P(CS$|R$) with respect to x except those terms that come from the derivatives of the integrals with lower limit $p_{[1]}/p_{[\alpha]}$ $(\alpha=2,\cdots,k)$. The derivatives of those terms are clearly non-negative. So P(CS$|R$) is a non-decreasing function of x and we can decrease the P(CS$|R$) by changing the configuration (2.8) to

$$p_{[1]}\leqq\cdots\leqq p_{[n]}=p_{[n+1]}\leqq\cdots\leqq p_{[k]}\,.$$

The above argument is true for every n with $2 \leqq n \leqq k$. Thus we can apply the same argument $(k-1)$ times and this completes the proof of this theorem.

Illustrative example. For $k=3$, we can write

$$\begin{aligned}
&\mathrm{P}(\mathrm{CS}\,|\,R)\\
&\quad=\mathrm{P}(x_2=r,\ \mathrm{ast};\ x_3<r',\ x_1<r')+\mathrm{P}(x_3=r,\ \mathrm{ast};\ x_2<r',\ x_1<r')\\
&\qquad+\mathrm{P}(x_2=r',\ \mathrm{ast};\ x_3<r,\ x_1<r')-\mathrm{P}(x_2=r',\ \mathrm{ast};\ x_3<r',\ x_1<r')\\
&\qquad+\mathrm{P}(x_3=r',\ \mathrm{ast};\ x_2<r,\ x_1<r')-\mathrm{P}(x_3=r',\ \mathrm{ast};\ x_2<r',\ x_1<r')\,.
\end{aligned}$$

Here, we only consider the case $r' \leqq r$, since $\mathrm{P}(\mathrm{CS}\,|\,R,\ r'=r+1) \equiv 1$ for any configuration. At first, consider $n=2$ and take $x=p_{[3]}/p_{[2]}$ and $p_{[2]}/p_{[1]}$ a fixed constant. By Lemma 2.1,

$$\begin{aligned}
(2.14)\qquad \frac{d\,\mathrm{P}(\mathrm{CS}\,|\,R)}{dx}=&\Big[-\frac{r'}{x}\mathrm{P}(x_2=r\ \mathrm{ast};\ x_3=r',\ x_1<r')\\
&+\frac{r}{x}\mathrm{P}(x_2=r'\ \mathrm{ast};\ x_3=r,\ x_1<r')\\
&-\frac{r}{x}\mathrm{P}(x_2=r',\ \mathrm{ast};\ x_3=r,\ x_1<r')\\
&+\frac{r'}{x}\mathrm{P}(x_2=r',\ \mathrm{ast};\ x_3=r',\ x_1<r')\\
&+\frac{r'}{x}\mathrm{P}(x_2=r,\ \mathrm{ast};\ x_3=r',\ x_1<r')\\
&-\frac{r'}{x}\mathrm{P}(x_2=r'\ \mathrm{ast};\ x_3=r',\ x_1<r')\Big]\\
&+\Big[\frac{r}{x}\mathrm{P}(x_1=r'\ \mathrm{ast};\ x_3=r,\ x_2<r')\\
&+\frac{r'}{x}\mathrm{P}(x_1=r'\ \mathrm{ast};\ x_3=r',\ x_2<r)\\
&-\frac{r'}{x}\mathrm{P}(x_1=r'\ \mathrm{ast};\ x_3=r',\ x_2<r')\Big]\,.
\end{aligned}$$

The first six terms in (2.14) come from the derivatives of the integral pieces with power limits $p_{[3]}/p_{[2]}$ and $p_{[2]}/p_{[3]}$ and the last three terms come from the derivatives of the integral pieces with lower limits $p_{[1]}/p_{[3]}$. It is clear that the first six terms sum up to 0 and the last three terms sum up to

$$\begin{aligned}
(2.15)\qquad &\frac{1}{x}[r\,\mathrm{P}(x_1=r'\ \mathrm{ast};\ x_3=r,\ x_2<r')\\
&\qquad+r'\,\mathrm{P}(x_1=r'\ \mathrm{ast};\ x_3=r',\ r'\leqq x_2<r)]
\end{aligned}$$

which is always non-negative.

Secondly, we consider $n=1$. Then by letting $y=p_{[2]}/p_{[1]}$ and consider $p_{[3]}/p_{[2]}$ as a fixed constant in (2.13), we have

$$\begin{aligned}(2.16)\qquad \frac{d\,\mathrm{P}(\mathrm{CS}\,|\,R)}{dy}=&\frac{r}{y}\,\mathrm{P}(x_1=r',\ \text{ast};\ x_3<r',\ x_2=r)\\ &+\frac{r}{y}\,\mathrm{P}(x_1=r',\ \text{ast};\ x_2<r',\ x_3=r)\\ &+\frac{r'}{y}\,\mathrm{P}(x_1=r',\ \text{ast};\ x_3<r,\ x_2=r')\\ &-\frac{r'}{y}\,\mathrm{P}(x_1=r',\ \text{ast};\ x_3<r',\ x_2=r')\\ &+\frac{r'}{y}\,\mathrm{P}(x_1=r',\ \text{ast};\ x_2<r,\ x_3=r')\\ &-\frac{r'}{y}\,\mathrm{P}(x_1=r'\ \text{ast};\ x_2<r',\ x_3=r')\\ =&\frac{r}{y}[\mathrm{P}(x_1=r',\ \text{ast};\ x_3<r',\ x_2=r)\\ &+\mathrm{P}(x_1=r',\ \text{ast};\ x_2<r',\ x_3=r)]\\ &+\frac{r'}{y}[\mathrm{P}(x_1=r',\ \text{ast};\ r'\leqq x_3<r,\ x_2=r')\\ &+\mathrm{P}(x_1=r'\ \text{ast};\ r'\leqq x_2<r,\ x_3=r')]\end{aligned}$$

which is always non-negative.

3. Expectations

Let N denote the total number of observations required by the procedure R. Using the same notation as in Section 2, we can write the expectation of N as follows:

$$\begin{aligned}(3.1)\qquad \mathrm{E}(N)=\sum_{i=1}^{k}(x_1+\cdots+x_k)\Big[&\sum_{m=2}^{k-1}\sum_{i,m}{}^{*}\,\mathrm{P}(x_i=r\ \text{ast};\ x_{i1}<r',\cdots,x_{im}<r';\\ &r'\leqq x_j<r\ \text{for}\ j\in\{1,\cdots,k\}-\{i,\,i1,i2,\cdots,im\})\\ &+\sum_{\substack{m=1\\ m\neq i}}^{k}\mathrm{P}(x_i=r'\ \text{ast};\ x_m<r';\ r'\leqq x_j<r\\ &\text{for}\ j\in\{1,\cdots,k\}-\{i,\,m\})\Big]\end{aligned}$$

where $\sum^*_{i,m}$ is over all the possible combinations $\{i1,\cdots,im\}$ of size m from $\{1,2,\cdots,k\}-\{i\}$.

It is easy to see from the definition of CD integral (2.3) that for any i,

$$(3.2)\qquad (x_1+\cdots+x_k)\,\mathrm{P}(x_i=r \text{ ast};\ x_j=r_j,\ j\neq i) = \frac{r}{p_i}\,\mathrm{P}(x_i=r+1 \text{ ast};\ x_j=r_j,\ j\neq i)\ .$$

By (3.1) and (3.2), we can write $\mathrm{E}(N)$ under the configuration (2.7) as

$$(3.3)\qquad \begin{aligned} \mathrm{E}(N|(2.7))=rk^2 \sum_{m=2}^{k-1} \sum_{k,m}{}^{*} \mathrm{P}(x_k=r+1 \text{ ast};\ &x_{k1}<r',\cdots,x_{km}<r';\\ &r'\leqq x_j<r \text{ for } j\in\{1,\cdots,k\}-\{k,k1,\cdots,km\})\\ &+r'k^2(k-1)\,\mathrm{P}(x_k=r'+1 \text{ ast};\ x_1<r';\ r'\leqq x_j<r\\ &\text{for } j\in\{1,\cdots,k\}-\{1,k\})\ . \end{aligned}$$

But in (3.3),

$$(3.4)\qquad \begin{aligned} &\sum_{m=2}^{k-1} \sum_{k,m}{}^{*} \mathrm{P}(x_k=r+1 \text{ ast};\ x_{k1}<r',\cdots,x_{km}<r';\ r'\leqq x_j<r\\ &\qquad \text{for } j\in\{1,\cdots,k\}-\{k,k1,\cdots,km\})\\ &\quad =\sum_{m=1}^{k-1} \sum_{k,m}{}^{*} \mathrm{P}(x_k=r+1 \text{ ast};\ x_{k1}<r',\cdots,x_{km}<r';\ r'\leqq x_j<r\\ &\qquad \text{for } j\in\{1,\cdots,k\}-\{k,k1,\cdots,km\})\\ &\quad\ -(k-1)\,\mathrm{P}(x_k=r+1 \text{ ast};\ x_1<r';\ r'\leqq x_j<r\\ &\qquad \text{for } j\in\{1,\cdots,k\}-\{1,k\}) \end{aligned}$$

where $\sum_{k,m}^{*}$ is the same as we defined in (3.1).

By using tne inclusion-exclusion principle as we did in Section 2 and using (3.4), we can rewrite (3.3) as

$$(3.5)\qquad \begin{aligned} \mathrm{E}(N|(2.7))=rk^2 &\sum_{i=1}^{k-1}(-1)^{i+1}\binom{k-1}{i}CD_{(1,\ldots,1)}(\boldsymbol{r}_i,\ r+1)\\ &-rk^2(k-1)\sum_{i=1}^{k-2}(-1)^{i+1}\binom{k-2}{i}CD_{(1,\ldots,1)}(\boldsymbol{r}_i,\ r+1)\\ &+r'k^2(k-1)\sum_{i=1}^{k-2}(-1)^{i+1}\binom{k-2}{i}CD_{(1,\ldots,1)}(\boldsymbol{r}_i,\ r'+1) \end{aligned}$$

where $\boldsymbol{r}_i=(r,\cdots,r,r',\cdots,r'\ i \text{ times})$.

Now we let S denote the subset size of the selected subset $\mathcal{S}$ for the procedure R. For any configuration, it is clear that

$$\mathrm{E}(S|R)=\sum_{i=1}^{k}(\Pi_i\in\mathcal{S}|R)\ .$$

Under the configuration (2.7), we can write

$$(3.6)\qquad \mathrm{E}(S|(2.7))=k\,\mathrm{P}(\Pi_{[1]}\in\mathcal{S}|(2.7))=k\,\mathrm{P}(\mathrm{CS}|(2.7))\ .$$

By (2.10), (2.11), (2.12), we can write $\mathrm{E}(S|(2.7))$ in (3.6) as

$$\text{(3.7)}\quad \mathrm{E}(S|(2.7))=k(k-1)\Bigg[\sum_{i=1}^{k-2}\binom{k-2}{i}(-1)^{i+1}CD_{(1,\dots,1)}(\boldsymbol{r}_i, r) \\ +CD_{(1,\dots,1)}(\boldsymbol{r}_1, r')+\sum_{i=2}^{k-1}\binom{k-2}{i-1}(-1)^{i+1}CD_{(1,\dots,1)}(\boldsymbol{r}_i, r')\Bigg]$$

where $\boldsymbol{r}_i$ is defined as in (3.5).

4. Comparison with a fixed-sample size procedure

In this section, we will make some analytical comparisons between our procedure R and the fixed-sample size procedure T that was proposed in Gupta and Nagel [4] for the special case $k=2$. The procedure T for selecting a subset containing the cell with the smallest probability is as follows:

Procedure T: Select the ith cell iff

$$x_i \leqq \min(x_1, \cdots, x_k)+c$$

where x_i $(i=1, \cdots, k)$ is the frequency of the ith cell and c is a given non-negative integer. Two meaningful criteria for evaluating the performance of a subset selection procedure are the smallness of $\mathrm{E}(S)$, the expected subset size and $\mathrm{E}(N)$, the expected sample size. When $k=2$, $\mathrm{E}(S|R)$ is either 1 (when $r' \leqq r$) or 2 (when $r'=r+1$). To make a suitable comparison, we take $r'=r$ for the procedure R and $c=0$ for the procedure T to make $\mathrm{E}(S|R)$ and $\mathrm{E}(S|T)$ as close as possible. For any configuration $(\mathrm{P}_{[1]} \cdot \mathrm{P}_{[2]})$,

$$\text{(4.1)}\qquad \mathrm{P}(\mathrm{CS}|R, r=r')=\mathrm{P}_{[2]}^{r}\sum_{\alpha=0}^{r-1}\binom{r+\alpha-1}{\alpha}\mathrm{P}_{[1]}^{\alpha}$$

and

$$\mathrm{P}(\mathrm{CS}|T, c=0, N=n)=\sum_{\alpha=[n/2]^+}\binom{n}{\alpha}\mathrm{P}_{[2]}^{\alpha}\mathrm{P}_{[1]}^{n-\alpha}$$

where $[n/2]^+$ is the smallest integer greater than $r/2$.

We also have

$$\text{(4.2)}\qquad \mathrm{E}(S|R, r=r')\equiv 1$$

$$\mathrm{E}(S|T, c=0, N=n)=\begin{cases} 1+\binom{n}{n/2}\mathrm{P}_{[2]}^{n/2}\mathrm{P}_{[1]}^{n/2} & \text{when } n \text{ is even}, \\ 1 & \text{when } n \text{ is odd}. \end{cases}$$

and

$$\mathrm{E}(N|T, c=0, N=n)=n\ .$$

We consider the following two cases separately.

Case 1. When $n=2r-1$,

$$\text{(4.3)}\qquad \mathrm{P}(\mathrm{CS}|R, r=r')=\mathrm{P}_{[2]}^{r}\sum_{\alpha=0}^{n-r}\binom{r+\alpha-1}{\alpha}\mathrm{P}_{[1]}^{\alpha}=\sum_{\alpha=r}^{n}\binom{n}{\alpha}\mathrm{P}_{[2]}^{\alpha}\mathrm{P}_{[1]}^{n-\alpha}$$
$$=\mathrm{P}(\mathrm{CS}|T, c=0, N=2r-1)\ .$$

The second equality for the sum of the upper and lower tails of the binomial series can be found in (2.3) and (2.4) of Sobel, Uppuluri and Frankowski [7]. From (4.2), we have

$$\mathrm{E}(S|R, r=r')=\mathrm{E}(S|T, c=0, N=2r-1)=1$$

and

$$\mathrm{E}(N|R, r=r')<2r-1=\mathrm{E}(N|T, c=0, N=2r-1)\ .$$

Thus for any configuration, the procedure R save the expected cost of sampling for this case.

Case 2. When $n=2r$,

$$\text{(4.4)}\qquad \mathrm{P}(\mathrm{CS}|R, r=r')=\mathrm{P}_{[1]}^{r}\sum_{\alpha=0}^{r-1}\binom{r+\alpha-1}{\alpha}\mathrm{P}_{[2]}^{\alpha}$$

and

$$\mathrm{P}(\mathrm{CS}|T, c=0, N=2r)=\sum_{\alpha=r}^{n}\binom{n}{\alpha}\mathrm{P}_{[2]}^{\alpha}\mathrm{P}_{[1]}^{n-\alpha}=\mathrm{P}_{[1]}^{r}\sum_{\alpha=0}^{r}\binom{r+\alpha-1}{\alpha}\mathrm{P}_{[2]}^{\alpha}\ .$$

Here, for the last equality, we use the same result as in (4.3).

From (4.4) and (4.2), we obtain

$$\text{(4.5)}\qquad \mathrm{P}(\mathrm{CS}|T, c=0, N=2r)-\mathrm{P}\left(\mathrm{CS}|R, r=r'=\frac{n}{2}\right)=\binom{2r-1}{r}\mathrm{P}_{[1]}^{r}\mathrm{P}_{[2]}^{r}\ ,$$

and

$$\text{(4.6)}\qquad \mathrm{E}(S|T, c=0, N=2r)=\mathrm{E}\left(S|R, r=r'=\frac{n}{2}\right)=\binom{2r}{r}\mathrm{P}_{[1]}^{r}\mathrm{P}_{[2]}^{r}$$

respectively. Thus by using the procedure R with a stopping frequency $r=N/2$, we shall expect a smaller selected subset and a smaller probability of correct selection than by using the procedure T. However, the differences $\binom{2r}{r}\mathrm{P}_{[1]}^{r}\mathrm{P}_{[2]}^{r}$ and $\binom{2r-1}{r}\mathrm{P}_{[1]}^{r}\mathrm{P}_{[2]}^{r}$ approach to 0 when r (or n) is large. The most important fact is that the expected sample size $\mathrm{E}(N|R, r=r')<2r-1$ is always less than $2r$, the fixed sample size for T.

5. Concluding remarks

$\mathrm{P}(\mathrm{CS}|R)$ the probability of correct selection, $\mathrm{E}(N)$ the expected number of observations required and $\mathrm{E}(S)$ the expected subset size can all be written in terms of the CD-integrals as we can see from (3.5), (3.6) and (3.7). The computation of these values is complicated due to the fact that a compound stopping rule is involved in the procedure R. (Thus both r and r' appear in r_i and this makes the computation difficult.) Actually, the table for $r=r'$ can be found in (8) for various arguments involved. However, it is believed that a compound stopping rule is necessary for the selecting goal for the least possible cell. We hope that the technique and procedure we present in this paper can be used in developing multinomial selection procedures.

Acknowledgements

I wish to thank Milton Sobel for suggesting this delightful research topic. Special thanks are due to the referees for their careful reading and correction, for suggesting Section 4 and Remark 2.2.

Syracuse University

References

[1] Alam, K. and Thompson, J. R. (1971). On selecting the least probable multinomial event, *Ann. Math. Statist.*, **43**, 1983-1990.

[2] Bechhofer, R. E., Elmaghraby, S. and Morse, N. (1959). A single-sample multiple-decision procedure for selecting multinomial event which has the highest probability, *Ann. Math. Statist.*, **30**, 102-119.

[3] Cacoullos, T. and Sobel, M. (1966). An inverse-sampling procedure for selecting the most probable event in a multinomial distribution, *Multivariate Analysis* (ed. P. R. Krishnaiah), Academic Press, New York, 423-455.

[4] Gupta, S. S. and Nagel, K. (1967). On selection and ranking procedures and order statistics from the multinomial distributions, *Sankhyā*, Ser. B, **29**, 1-34.

[5] Panchapakesan, S. (1971). On a subset selection prodedure for the most probable event in a multinomial distribution, *Statistical Decision Theory and Related Topics* (eds. S. S. Gupta and J. Yackel), Academic Press, New York, 275-298.

[6] Panchapakesan, S. (1971). On a subset selection procedure for the best multinomial cell and related problems, *Abstract, Bull. Inst. Math. Statist.*, **2**, 112-113.

[7] Sobel, M., Uppuluri, V. R. R. and Frankowski, K. (1977). Dirichlet Distribution-Type 1, *Selected Tables in Mathematical Statistics*, Vol. **4**, American Mathematical Society, Providence, Rhode Island.

[8] Sobel, M., Uppuluri, V. R. R. and Frankowski, K. Dirichlet Integrals of Type 2 and Their Applications, (to appear).