Inequalities in Statistics and Probability
IMS Lecture Notes–Monograph Series Vol. 5 (1984), 206–210

SELECTING THE t BEST CELLS OF A MULTINOMIAL USING INVERSE SAMPLING

BY PINYUEN CHEN and MILTON SOBEL[1]
Syracuse University and University of California at Santa Barbara

An inverse sampling procedure R is proposed for selecting the t "best" cells (i.e., cells with the largest cell probabilities) from a multinomial distribution with k cells ($1 \leq t < k$). Two different formulations of this selection problem are considered and the measure of distance in both formulations is the ratio of the largest and second largest cell probabilities. One formulation is of the usual type based on an empty indifference zone; in the other (new) formulation any collection of t cells from the union of the preference zone (for selection) and the indifference zone is called a correct selection. Type 2-Dirichlet integrals are used (i) to express the probability of correct selection as an integral with parameters only in the limits of integration, and (ii) to prove that the least favorable configuration for each of the formulations under R is the so-called slippage configurations with $k-t$ equal cell probabilities and t cell probabilities slipped to the right by a common amount.

1. Introduction. One of the important applications of ranking and selection techniques is to select (without respect to order) the t best cells of a multinomial distribution with k cells. For the special case $t = 1$ the fixed sample size problem was first considered by Bechhofer, Elmaghraby and Morse (1959) and the inverse sampling procedure was first considered by Cacoullos and Sobel (1966). We are presently discussing fixed subset size problems and not considering the random subset size problem which was considered by Gupta and Nagel (1971) and more recently by Hu (1982). It is well known by people working in this area that the generalization of the fixed subset size problem to arbitrary t ($1 < t < k$) presents some serious difficulties (cf. the work of Lee (1975) and Hwang, Hsuan and Parned (1980) on this topic). In this paper we consider the corresponding problem for general $t \geq 1$ with an inverse sampling procedure.

Actually we consider two different formulations of the ranking and selection problem. The measure of distance in both formulations is the ratio of cell probabilities as in the previous references. Let

$$p_{[1]} \leq p_{[2]} \leq \ldots \leq p_{[k-t]} \leq p_{[k-t+1]} \leq \ldots \leq p_{[k]} \tag{1.1}$$

denote the ordered cell probabilities which sum to one. Let $\delta^* > 1$ and $P^*(\binom{k}{t}^{-1} < P^* < 1)$ denote specified constants. In the usual (or first) formulation we require a procedure R such that

$$P\{CS|R\} > P^* \text{ whenever } \delta \geq \delta^*, \tag{1.2}$$

where $\delta = p_{[k-t+1]}/p_{[k-t]}$.

Actually we need only consider configurations (1.1) with $p_{[k-t]} < p_{[k-t+1]}$ and in this case the definition of correct selection (CS) is clear, namely that we select the t cells with largest p-values.

We shall say the p-value is in the indifference zone (IZ) if it lies strictly between $p_{[k-t+1]}/\delta^*$ and $p_{[k-t+1]}$. The p-values $\geq p_{[k-t+1]}$ will be said to lie in the preference zone (PZ).

[1] This paper was supported by National Science Foundation Grant Number MCS82–02247.
AMS 1980 subject classifications. Primary, 62F07; secondary, 62E15.
Key words and phrases: Ranking and selection, inverse sampling, multinomial distribution, Type 2-Dirichlet integral.

Under the alternative (or second) formulation we consider any combination of the t cell probabilities, each $\geq p_{[k-t+1]}/\delta^*$, as being a correct selection and we use the terminology CSA for any such combination. Thus if we take any t cells from those in the union of the PZ and IZ as our selected subset, we call this a correct selection (CSA) under the second formulation.

Our goal is to show that the so-called least favorable configuration is the one with t cell probabilities slipped to the right by a common amount. We show below that under inverse sampling this is least favorable for both of the above formulations.

The main tool used in this paper is the fact that the P(CS) and also the P(CSA) can be expressed exactly 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 integration.

For both formulations we use the same sampling and the same decision procedure specified by a positive integer r to be determined (with the help of Dirichlet Tables).

Procedure R: Continue sampling one-at-a-time until t cells reach a frequency of at least r. As soon as this occurs we stop and select these t cells as being those with the t largest probabilities.

It is clear that under this procedure there can be no ties for the t-th position and hence the selected subset is well defined. Note that we are selecting the t best without respect to order, so that frequency ties present no difficulty.

2. P(CS) and Least Favorable Configuration for the First Formulation. It has been shown (Sobel, Uppuluri and Frankowski (1983)) that for a multinomial distribution the probability, when a specified cell (called the counting cell) reaches frequency m, that certain $(t-1)$ specified cells all have frequency $\geq r$ and the remaining $(k-t)$ cells all have frequency $< r$, is given exactly by the CD integral

$$\text{(2.1)} \qquad \mathrm{CD}_{\mathbf{a}}^{(t-1;\,k-t)}(r, m) = [\Gamma(m+(k-1)r)/\Gamma^{k-1}(r)\Gamma(m)] \int_0^{a_1} \cdots \int_0^{a_{t-1}} \int_{a_t}^{\infty} \cdots \int_{a_{k-1}}^{\infty} (1 + \Sigma_{i=1}^{k-1} x_i)^{-(m+(k-1)r)} \cdot \Pi_{i=1}^{k-1} x_i^{r-1} dx_i$$

where $\mathbf{a} = (a_1, \ldots a_{t-1}, a_t, \ldots, a_{k-1})$ and $a_j = p_j/p_0$ is the ratio of the j-th cell probability to that of the counting cell; here we have assumed that the first $t-1$ cells form the specified set of size $t-1$ and that the counting cell is the last cell with probability p_0.

Using this probability interpretation with $m = r$ we can write the P(CS) for the first formulation above as the sum of t terms; in the j-th term we take the cell associated with $p_{[j]}$ as the counting cell ($j = k-t+1, k-t+2, \ldots k$). Hence we obtain

$$\text{(2.2)} \qquad P(\mathrm{CS}|R) = \Sigma_{j=k-t+1}^{k} \mathrm{CD}_{\mathbf{a}_j}^{(t-1;\,k-t)}(r, r)$$

where

$$\text{(2.3)} \qquad \mathbf{a}_j = (p_{[k]}/p_{[j]}, \ldots, p_{[j+1]}/p_{[j]}, p_{[j-1]}/p_{[j]}, \ldots, p_{[1]}/p_{[j]}).$$

Thus the sum of the two superscripts is the number of components of $\mathbf{a}$ and from (2.1) we see that the first $t-1$ components are upper limits in the CD integral, while the last $k-t$ components are lower limits.

The same probability interpretation gives us another expression for the P(CS), which we need for our result, by taking any one of the second set of $k-t$ specified cells as our counting cell. Thus we can also write

$$\text{(2.4)} \qquad P(\mathrm{CS}|R) = \Sigma_{j=1}^{k-t} \mathrm{CD}_{\mathbf{a}_j}^{(t,\,k-t-1)}(r, r),$$

where $\mathbf{a}_j$ is again given by (2.3) but j runs over a different set.

Under the first formulation the condition $\delta \geq \delta^*$ is equivalent to the inequality $p_{[k-t]} < p_{[k-t+1]}/\delta^*$ or equivalently the open interval $(p_{[k-t+1]}/\delta^*, p_{[k-t+1]})$ is empty. Using (2.2) and keeping the pairwise ratio of each two of the t largest p-values fixed we now consider $p_{[k-t]}$ as a variable. Since $p_{[k-t]}$ is in the numerator of the lower limit, the $P(\mathrm{CS})$ is monotonically decreasing in $p_{[k-t]}$. Hence we can decrease the $P(\mathrm{CS})$ by increasing the value of $p_{[k-t]}$ until it reaches $p_{[k-t+1]}/\delta^*$. The increase in the value of $p_{[k-t]}$ is offset by a decrease in $p_{[j]}$, as the sum of all the p-values has to remain equal to one. Since all the lower limits of the last $k-t$ integrals are increased, the value of the $P(\mathrm{CS})$ must decrease. Note that $p_{[j]} \geq p_{[k-t+1]} > p_{[k-t+1]}/\delta^*$; hence $p_{[k-t]}$ must reach its boundary first. The same argument allows us to increase in turn $p_{[k-t+1]}, p_{[k-t+2]}, \ldots, p_{[1]}$ up to the same boundary value, namely $p_{[k-t+1]}/\delta^*$.

We now use (2.4) with $p_{[j]} = p_{[k-t+1]}/\delta^*$ for $j = 1, 2, \ldots, k-t$. Consider $p_{[k-t+2]}$ as a variable with all the largest p-values as fixed, except that $p_{[k-t+1]}$ and the boundary $p_{[k-t+1]}/\delta^*$ can still vary. The $P(\mathrm{CS})$ has now been decreased to the value $P_1(\mathrm{CS})$ where

$$P_1\{\mathrm{CS}|R\} = (k-t)\mathrm{CD}_{\mathbf{a}}^{(t,\,k-t-1)}(r, r) \tag{2.5}$$

where

$$\mathbf{a} = (p_{[k]}/p_{[k-t+1]})\,\delta^*, \ldots, (p_{[k-t+2]}/p_{[k-t+1]})\delta^*, \delta^*, 1, \ldots, 1)$$

and the last $k-t-1$ components are all exactly 1. Since $p_{[k-t+2]}$ appears as the numerator of an upper limit of integration in (2.5), it follows that we can further decrease the $P(\mathrm{CS})$ by decreasing $p_{[k-t+2]}$ to $p_{[k-t+1]}$; actually $p_{[k-t+1]}$ was increasing so equality has to occur. Similarly we decrease all the larger p-values until they reach $p_{[k-t+1]}$.

The above argument proves the following

THEOREM 1. *The least favorable configuration for the first formulation is given by*

$$\begin{aligned} p_{[1]} = p_{[2]} = \ldots = p_{[k-t]} &= 1/(k-t+t\delta^*), \\ p_{[k-t+1]} = \ldots = p_{[k]} &= \delta^*/(k-t+t\delta^*). \end{aligned} \tag{2.6}$$

3. P(CSA) and the Least Favorable Configuration for the Second Formulation. Consider the general configuration for the second formulation as follows:

$$p_{[1]} \leq \ldots \leq p_{[k-t-r]} \leq p_{[k-t+1]}/\delta^* < p_{[k-t-r+1]} \leq \ldots \leq p_{[k-t]} \leq \ldots \leq p_{[k]} \tag{3.1}$$

where r is the number of cell probabilities in the IZ.

The probability of correct selection under the second formulation can be written in the following Dirichlet integral form:

$$P(\mathrm{CSA}|R) = \Sigma^* \Sigma_{j=1}^{t} \mathrm{CD}_{\mathbf{a}_j}^{(t-1,\,k-t)}(r, r), \tag{3.2}$$

where Σ^* is over all possible subsets $\{p_{s_1}, p_{s_2}, \ldots, p_{s_t}\}$ of size t that can be taken from the set $\{p_{[k-t-r+1]}, \ldots, p_{[k]}\}$ of size $t+r$ and

$$\mathbf{a}_j = (p_{s_j})^{-1}(p_{s_1}, \ldots, p_{s_{j-1}}, p_{s_{j+1}}, \ldots, p_{s_t}, p_{[1]}, \ldots, p_{[k]}). \tag{3.3}$$

It should be noted that there are only $k-t$ components in (3.3) after p_{s_t}/p_{s_j} and that the numerators of these are taken from the set $\{p_{[1]}, p_{[2]}, \ldots, p_{[k]}\} - \{p_{s_1}, \ldots, p_{s_t}\}$ so that $\mathbf{a}_j$ in (3.3) has a total of $k-1$ components.

Using (3.2) and keeping the pairwise ratio of each of the $t+r$ largest p-values fixed, we now consider $p_{[k-t-r]}$ as a variable. Since $p_{[k-t-r]}$ is a numerator among the lower limits in (3.2), the $P(\mathrm{CSA})$ is monotonically decreasing in $p_{[k-t-r]}$. Hence we can decrease the $P(\mathrm{CSA})$ by increasing the values of $p_{[k-t-r-1]}, \ldots, p_{[1]}$ up to the common boundary value, namely $p_{[k-t+1]}/\delta^*$.

Thus we can restrict ourselves to the following configuration:

$$p_{[1]} = \ldots = p_{[k-t-r]} = p_{[k-t+1]}/\delta^* < p_{[k-t-r+1]} \leqslant \ldots \leqslant p_{[k]} \tag{3.4}$$

for minimizing $P(\mathrm{CSA}|R)$ in (3.2).

It is clear from (3.4) that

$$p_{[1]}/p_{[k-t+1]} = \ldots = p_{[k-t+r]}/p_{[k-t+1]} = (\delta^*)^{-1}.$$

Let $p_{[j]}/p_{[k-t+1]} = a_j$ for $j = k-t-r+2, \ldots, k$ be kept as constants and let $p_{[k-t+1]}/p_{[k-t-r+1]} = x$ be the only variable in $P\{\mathrm{CSA}|R\}$ in (3.4) with the obvious restrictions that $\Sigma_{i=1}^{k} p_{[i]} = 1$ and $a_{k-t+1} = p_{[k-t+1]}/p_{[k-t+1]} = 1$. Then from (3.4) we can write

$$\begin{aligned} P\{\mathrm{CSA}|R\} &= \Sigma 1_1^* \Sigma_{\alpha=1}^{t} \mathrm{CD}_{\mathbf{a}_\alpha}^{(t-1;k-t)}(r, r) \\ &+ \Sigma_2^* [\Sigma_{\beta=1}^{k-t-r} \mathrm{CD}_{\mathbf{a}_\beta}^{(t;k-t-1)}(r, r) + \Sigma_{\substack{\gamma=k-t-r+1 \\ \gamma \neq s_1, \ldots, s_t}}^{k} \mathrm{CD}_{\mathbf{a}_\gamma}^{(t;k-t-1)}(r, r)] \end{aligned} \tag{3.5}$$

where Σ_1^* is over the subsets $(p_{s_1}, \ldots, p_{s_t})$ of size t which do not include $p_{[k-t-r+1]}$ and Σ_2^* is over those that do include $p_{[k-t-r+1]}$. In the former case (i.e., under Σ_1^*) the structure of $\mathbf{a}_\alpha$ is

$$\begin{aligned} \mathbf{a}_\alpha = (&p_{s_1}/p_{s_\alpha}, \ldots, p_{s_{\alpha-1}}/p_{s_\alpha}, p_{s_{\alpha+1}}/p_{s_\alpha}, \ldots, p_{s_t}/p_{s_\alpha}, 1/(\delta^* a_{s_\alpha}), \ldots, \\ &1/(\delta^* a_{s_\alpha}), 1/(x a_{s_\alpha}), a_{k-t-r+2}/a_{s_\alpha}, \ldots, a_k/a_{s_\alpha}) \end{aligned} \tag{3.6}$$

where x appears in exactly one component and we are holding all the other ratios fixed. In the latter case, (i.e., under Σ_2^*) we use the alternative form (2.4) to write the relevant probabilities (i.e., the counting cells are taken from the set that is not selected) and we separate this sum into two parts according to whether the counting cell is among $p_{[1]}, \ldots, p_{[k-t-r]}$ or is in the difference of the two sets $\{p_{[k-t-r+1]}, \ldots, p_{[k]}\} - \{p_{s_1}, \ldots, p_{s_t}\}$. In the first of these two parts we write $\mathbf{a}_\beta$ and its structure is

$$\mathbf{a}_\beta = (p_{s_1}/p_\beta, \ldots, \delta^*/x, \ldots, p_{s_t}/p_\beta, 1/(\delta^* a_\beta), \ldots, p_{\beta-1}/p_\beta, p_{\beta+1}/p_\beta, \ldots, a_k/a_\beta) \tag{3.7}$$

where δ^*/x comes from the ratio $p_{[k-t-r+1]}/p_\beta$ and the remaining ratios are all fixed. In the second of these two parts we write $\mathbf{a}_\gamma$ and its structure is

$$\mathbf{a}_\gamma = (p_{s_1}/p_\gamma, \ldots, 1/(x a_\gamma), \ldots, p_{st}/p_\gamma, 1/(\delta^* a_\gamma), \ldots, a_k/a_\gamma), \tag{3.8}$$

where $(x a_\gamma)^{-1}$ comes from the ratio $p_{[k-t-r-1]}/p_r$ and the remaining ratios are all fixed.

Note that the total number of terms in $\Sigma_1^* \Sigma_{\alpha=1}^{t}$ is $\binom{t+r-1}{t} \cdot t = (t+r-1)!/[(t-1)!\,(r-1)!]$ and the total number of terms in the second part of Σ_2^*, namely in $\Sigma_2^* \Sigma_{\substack{\gamma=k-t-r+1 \\ \gamma \neq s_1, \ldots, s_t}}^{k}$ is $\binom{t+r-1}{t-1} \cdot r = (t+r-1)!/[(t-1)!\,(r-1)!]$ also.

Actually we can set up a 1-1 correspondance between the terms in Σ_1^* and those in the second part of Σ_2^* as follows. Each term in Σ_1^* corresponds to a selected subset of size t and one of these t cells is used as a counting cell. Say we have $p_k, \ldots, p_{k-t+1}$ and p_k is the counting cell to specify a single term in Σ_1^*. Then we take the selected subset to be $p_{[k-t-r+1]}, p_{[k-t+1]}, \ldots, p_{[k-1]}$ and use $p_{[k]}$ as a counting cell to obtain a specific term in the second part of Σ_2^*, and this illustrates the correspondance of the terms. In Σ_1^* the varying lower limit is $(x a_\alpha)^{-1}$, which in our example is $(x a_k)^{-1}$ and in the corresponding term in the second part of Σ_2^* has the varying upper limit $(x a_\gamma)^{-1}$, which in our example is $(x a_k)^{-1}$. The other limits are all the same in corresponding terms. Hence the derivatives of corresponding terms cancel. Since the only remaining terms are those from the first part of Σ_2^* and these are all negative, it follows that $P(\mathrm{CSA}|R)$ is decreasing in x. Thus we decrease $P(\mathrm{CSA}|R)$ by lowering $p_{[k-t-r+1]}$ to $p_{[k-t+1]}/\delta^*$. Similarly we decrease in turn all the $p_{[k-t-r+2]}, \ldots, p_{[k-t]}$ until they reach $p_{[k-t+1]}/\delta^*$. The above argument proves the following

THEOREM 2. *The least favorable configuration for the second formulation is given by*

(3.8) $$p_{[1]} = p_{[2]} = \ldots = p_{[k-t]} = 1/(k-t+t\delta^*),$$
$$p_{[k-t+1]} = \ldots = p_{[k]} = \delta^*/(k-t+t\delta^*),$$

exactly the same slippage configuration as in (2.6).

REFERENCES

BECHHOFER, R. E., S. A. ELMAGHRABY, and N. MORSE (1959). A single-sample multiple-decision procedure for selecting the multinomial event which has the largest probability. *Ann. Math. Statist. 30* 102–119.

CACOULLOS, T. and M. SOBEL (1966). An inverse sampling procedure for selecting the most probable event in a multinomial distribution, in *Multivariate Analysis: Proceedings of an International Symposium,* P. R. Krishnaiah, ed., Academic Press, New York, 423–455.

GUPTA, S. S. and K. NAGEL (1971). On some contributions to multiple decision theory, in *Statistical Decision Theory and Related Topics,* S. S. Gupta and J. Yackel, eds., Academic Press, New York, 79–102.

HU, S. P. (1982). Inverse sampling procedure for multinomial subset selection problems. Ph.D. Dissertation, Department of Mathematics, University of California at Santa Barbara, Santa Barbara, CA.

HWANG, F. K., F. HSUAN, and M. PARNED (1980). Some inequalities for the multinomial selection problem. Unpublished manuscript.

LEE, Y. J. (1975). On selecting the t most probable multinomial events: Single-sample procedures. Unpublished manuscript.

SOBEL, M., V. R. R. UPPULURI, and K. FRANKOWSKI (1983). *Selected Tables in Mathematical Statistics: Dirichlet Distribution - Type 2,* American Mathematical Society, Providence, RI (to appear).