Proceedings of the 1988 Winter Simulation Conference
M. Abrams, P. Haigh, and J. Comfort (eds.)

On selecting the best of K systems: An expository survey of subset-selection multinomial procedures

Pinyuen Chen
Department of Mathematics
Syracuse University
Syracuse, NY 13244-1150

ABSTRACT

This paper studies the subset-selection approach of the ranking and selection procedures of choosing among k arbitrary systems or alternatives. Ranking and selection problems have customarily been treated using two different approaches, namely, the indifference-zone approach and the subset-selection approach. An expository survey of indifference-zone approach for selecting the best of k systems has been given in Goldsman (1984a). In this paper, we present a number of fixed-sample-size and sequential procedures based on subset-selection approach.

1. INTRODUCTION

Goldsman (1984a) presented an expository survey of procedures for selecting that multinomial cell which has the largest underlying probability. It was shown in that paper that the problem of finding the 'best' one of k arbitrary competing systems or alternatives can be formulated as a problem of finding that one category of a k-nomial distribution with the highest underlying 'success' probability. Since most real-life systems do not follow the 'usual' probability distributions, such multinomial selection procedures are seen to be very useful. Furthermore, Goldsman (1984b) discussed various important roles that simulation plays in the development of such procedures. Motivated by the discussion in the above-mentioned articles, we are presenting multinomial selection procedures in another approach, namely, the subset-selection approach. The goal of the subset-selection approach is to select a nonempty random size subset of the k multinomial cells so that the 'best' is included in the subset with a guaranteed minimum probability $P^*\left(\frac{1}{k} < P^* < 1\right)$. There are many practical situations where it may be better to select more than one population. For example, we may need to divide a number of competing systems into two groups--one that we feel certain the 'best' system and one that we feel certain mostly inferior systems. Under this circumstance, the experimenter may decide to adopt the subset-selection approach. There are several fundamental differences between the indifference-zone approach and subset-selection approach. For a more complete discussion concerning these two approaches, the readers are referred to Gupta and Panchapakesan (1979) and Gibbons, Olkin, and Sobel (1977). In Section 2, we give a summary of the pertinent notation and terminology for multinomial selection problems. In Section 3, we review several existing fixed-sample-size procedures and sequential procedures and give examples to illustrate the use of tables for these procedures. We conclude this paper by proposing two sequential procedures in Selection 4.

2. NOTATION AND TERMINOLOGY

A multinomial distribution with k cells $\pi_1, \pi_2, \ldots, \pi_k$ is given; let the ordered values of the unknown cell probabilities $p_i \geq 0$ $(1 \leq i \leq k)$ with $\sum_{i=1}^{k} p_i = 1$ be denoted by $p_{[1]} \leq p_{[2]} \leq \cdots \leq p_{[k]}$, and the corresponding cell be denoted by $\pi_{(1)}, \pi_{(2)}, \ldots, \pi_{(k)}$. It is assumed that the values of the p_i and $p_{[j]}$ $(1 \leq i, j \leq k)$ are unknown, and the parings of the π_i with $p_{[j]}$ are completely unknown. The goal of the experimenter is to select a random size subset containing the cell $\pi_{(k)}$. A correct selection (CS) is defined as the selection of any subset of the k cells which

contains the cell $\pi_{(k)}$. If more than one cell has a p-value equal to $p_{[k]}$, then one of the cells with the largest value is considered 'tagged' and the selection is correct if the 'tagged' cell is in the selected subset. Let $p^*\left(\frac{1}{k} < p^* < 1\right)$ denote a specified constant. We require a procedure R which guarantees that for all $\underset{\sim}{p} = (p_1, p_2, \ldots, p_k)$ we have

$$P(CS|R) \geq P^* . \qquad (1)$$

Let n denote the largest number of vector-observations that the experimenters will be allowed to take. The value of n may have been based on economic considerations. By stage m ($m \leq n$), we shall mean that a total of m vector-observations have already been taken. Let the random variables $Z_{i,m}$ ($1 \leq i \leq k$, $1 \leq m \leq n$) denote the frequency in cell π_i through stage m. We also use the notations $Z_{[1],m} \leq Z_{[2],m} \leq \cdots \leq Z_{[k],m}$ to denote the ordered frequencies at stage m. For fixed-sample size procedure with size n, the experimenter may take n observations at once and thus $m \equiv n$. For sequential procedures, observations are taken one-at-a-time (up to a limit n) until one call has 'significantly more' successes than the other cells.

Two important configurations of $p_{[i]}$'s are usually used for comparison among procedures:

(1) Slippage configuration (SC):

$p_{[k]} = \theta^* p_{[i]}$, $i=1,\ldots,k-1$ where $\theta^* > 1$ is a fixed constant.

(2) Equal probability configuration (EPC):

$p_{[i]} = 1/k \quad i=1,\ldots,k.$

For most multinomial selection procedures (in either indifference-zone approach or subset-selection approach), SC has been shown to be the configuration that gives the minimum probability of a correct selection under the so called preference-zone:

$$PZ = \{\underset{\sim}{p} : p_{[k]} \geq \theta^* p_{[i]},\ i=1,2,\ldots,k-1\}. \qquad (2)$$

For most procedures in subset-selection approach, EPC has been shown to be the configuration that gives the minimum probability of a correct selection in the entire parameter space. The expected selected subset size E(S) is a natural criterion on the efficiency of a subset selection procedure which satisfies the probability requirement (1). Another criterion for comparing sequential procedures is E(N), the expected sample size. For fixed-sample-size procedure with size n, it is clear that $E(N) \equiv n$. We will use E(S) and E(N) under SC and EPC to compare several selection procedures in next section.

3. SOME SELECTION PROCEDURES

In this section, we state four existing procedures for selecting a random size subset that contains the multinomial cell which has the largest cell probability. Procedure R_{GN} is a fixed-sample-size procedure proposed by Gupta and Nagel (1967). Procedure R_{BC} is a curtailed version of R_{GN} studied by Bechhofer and Chen (1987). Procedure R_P is an inverse sampling procedure proposed by Panchapakesan (1971). Procedure P_C is a truncated version of R_P proposed by Chen (1987).

Procedure R_{GN}: A fixed-sample-size procedure A total of n observations is taken in a single stage. Include in the selected subset the cell π_i with the observed frequency $z_{i,n}$ if and only if $z_{i,n} \geq z_{[k],n} - D$ where D is a predetermined non-negative integer.

Remark 3.1: Gupta and Nagel (1967) proved that the configuration of the parameters that minimizes the probability of a correct selection in the entire parameter space (which is called the least favorable configuration, abbreviated as LFC) is of the form

$$\underset{\sim}{p} = (0,0,\ldots,0,s,p,p,\ldots,p) \text{ where } 0 \leq s \leq p. \qquad (3)$$

They have done numerical computation to show that $s \neq 0$ in only one case for the cases $D = 0(1)4$, $k = 2(1)10$ and $n = 2(1)15$, namely, for the case k=3, n=6 and D=4. Notice that (3) can be simplified to EPC when s=0. Tables of P(CS) and E(S)/k of various combinations of D, k, and n were provided in their paper for EPC and for SC with $\theta^* = 3$ and 5.

Example 3.1: Suppose that k=4 and $P^* = .75$. Table 1 is abstracted from Gupta and Nagel (1967):

Table 1: $P(CS)$ and $E(S)/k$ for $\theta^* = 1$

n\D	0	1	2
3	.4375	.6719	.9531
4	.3555	.6367	.8945
5	.3379	.6309	.8359
6	.3672	.5869	.7957

Notice that $P(CS) = E(S)/k$ for $\theta^* = 1$ under EPC. It can be seen that $P(CS)$ is not an increasing function in n, but it is increasing in D. For all the fixed-sample size $n = 3,4,5$, and 6, the experimenter would have to use $D = 2$ in order that $P^* = .75$ is guaranteed. However, $n = 6$ gives the smallest $E(S)/k$. Thus if the experimenter can choose among $n = 3,4,5$, and 6, he will use $n = 6$ to keep $E(S)/k$ as small as possible.

Procedure R_{BC}: A curtailed sequential procedure. Observations are taken one at a time. Stop sampling at the first stage m at which there exists a cell π_i such that $z_{i,m} > z_{j,m} + n - m + D$ for all $j \neq i$ $(i,j = 1,2, \ldots,k)$. Having stopped, include in the selected subset the cell π_i with frequency $z_{i,m}$ if and only if $z_{i,m} \geq z_{[k],m} - D$ where D is a predetermined non-negative integer.

Remark 3.2: Bechhofer and Chen (1987) proved that both R_{GN} and R_{BC} select the same subset of the k cells if both use the same D. The result is uniform in $(p_1, p_2, \ldots, p_k)$. As a consequence of the above result, the LFC is the same for both procedures uniformly in n and k. However the sequential procedure R_{BC} accomplishes the same $P(CS)$ and $E(S)/k$ with a smaller expected number of observations $E(N)$ than required by R_{GN}.

Example 3.2: Suppose that $k = 4$ and that we specify $P^* = .90$ and $\theta^* = 3.0$. Table 2 gives $P(CS)$ and Table 3 gives respective $E(S)/k$. Both tables are abstracted from Gupta and Nagel (1967).

Table 2: $P(CS|R_{GN})$

n\D	0	1	2
9	.8578	.9351	.9747
10	.8734	.9398	.9761
13	.9064	.9546	.9804
14	.9161	.9585	.9818

Table 3: $E(S|R_{GN})/k$

n\D	0	1	2
9	.2894	.3911	.5159
10	.2870	.3714	.4837
13	.2738	.3331	.4111
14	.2719	.3146	.3926

It is clear that $\{n = 13, D = 0\}$ is the best combination to achieve $P(CS) \geq P^*$ since it gives the smallest $E(S)/k$. Now we look at $E(N)$ for the procedure R_{BC}. The following table is abstracted from Bechhofer and Chen (1987).

Table 4: $PE(N|R_{BC})$

n\D	0	1	2
9	7.961	8.459	.8725
10	8.830	9.293	9.645
13	11.276	11.859	12.281
14	12.104	12.679	13.153

The $E(N)$ value of R_{BC} for $\{n = 13, D = 0\}$ is 11.276 which is about 13% saving in the sample size over R_{GN}. If the sample size is a more important issue than the subset size for the experimenter, he may choose $\{n = 9, D = 1\}$ since it provides $P(CS|R_{BC}) = P(CS|R_{GN}) = .9351 \geq P^*$ and $E(N|R_{BC}) = 8.459$ gives 34% saving in the sample size over R_{GN}.

Procedure R_P: An inverse sampling procedure Observations are taken one at a time and the sampling is terminated at stage N when any one of the cell frequencies, say $f_{j,N}$ reaches M. Include in the selected subset the cell π_i with the observed frequency $z_{i,N}$ if and only if $z_{i,N} \geq M-D$ where M and D are predetermined integers.

Remark 3.3 M is a positive integer that determines the stopping time of the sampling and D is a non-negative integer that determines the cells that are to be included in the selected subset.

Remark 3.4: The LFC for the procedure R_P was conjectured in Panchapakesan (1971) as EPC. It was proved by Chen (1986) the conjecture is correct. The tables of M and D for various k and P* are currently being prepared.

Procedure R_C: A truncated inverse sampling procedure. Observations are taken one at a time until either (1) the frequency in any cell reaches M or (2) the total frequency reaches n. Suppose that the sampling is terminated at stage N. As soon as (1) occurs before (2), include in the selected subset the cell π_i with the observed frequency $z_{i,N}$ if and only if $z_{i,N} \geq M - D$. As soon as (2) occurs before (1), include in the selected subset the cell π_j with the observed frequency $z_{j,n}$ if and only if $z_{j,n} \geq z_{[k],n} - D$.

Remark 3.5: It was proved by Chen (1987) that the LFC for the procedure R_C is EPC.

Remark 3.6: This procedure is a composite of Procedures R_C and R_P. It is clear that R_C is exactly R_P when $n > k(M-1)$. It was proved by Chen (1987) that

$$P(CS|R_C) = P(CS|R_{GN}) \quad \text{and} \quad E(S|R_C) = E(S|R_{GN}) \tag{4}$$

uniformly in $\underset{\sim}{p} = (p_1, p_2, \ldots, p_k)$ when $2M > D + N$. It is also clear from the earlier termination of the sampling rule in R_C that

$$E(N|R_C) \leq n = E(N|R_{GN}) \tag{5}$$

Remark 3.7: Different combinations of M and D provide different P(CS) for the procedure R_C. We often choose the combination of M and D that provides the P(CS) as close as possible to the P*-requirement to keep E(N) as small as possible. The following examples illustrate that neither R_C nor R_{BC} dominates the other uniformly in n,k,P* and θ*.

Example 3.3: Suppose that $n = 6$, $k = 5$, $P^* = .80$ and $\theta^* = 1$. From Table 1A of Gupta and Nagel (1967), $D = 2$, gives $P(CS) = E(S)/k = .8054$. From Table in Bechhofer and Chen (1987), we find $E(N|R_{BC}) = 5.998$. From Table in Chen (1987), we find that {$M = 4$ and $D = 2$} gives $P(CS) = E(S)/k = .8042$. Furthermore, $E(N|R_C) = 5.959$. Thus R_C is superior to R_{BC} in both E(S) and E(N) in this example.

Example 3.4: Suppose that $n = 6$, $k = 4$, $P^* = .80$ and $\theta^* = 3$. From Table 1A of Gutpa and Nagel (1967), the smallest D (which will provide the smallest E(S)) that guarantees $P^* = .80$ is $D = 0$. The following values are found in Gupta and Nagel (1967) and Bechhofer and Chen (1987):

$P(CS|R_{BC}, D = 0) = .8125$
$E(S|R_{BC}, D = 0) = .3207$
$E(N|R_{BC}, D = 0) = 5.475$.

From Table in Chen (1987), the combination {$M = 4$, $D = 0$} gives the smallest E(N) among all the combinations of M and D that achieve $P(CS) \geq .80$. From the same table, we find

$D(CS|R_C, m = 4, d = 0) = .8125$
$E(S|R_C, M = 4, D = 0) = .3207$
$E(N|R_C, M = 4, D = 0) = 5.738$.

Thus R_{BC} is superior to R_C in E(N) in this example.

4. CONCLUDING REMARKS

Bechhofer and Goldsman (1985,1986) suggested procedures which are truncations of Bechhofer-Kiefer-Sobel (abbreviated as BKS) sequential procedure for indifference-zone approach. Numerical evidences show that their procedures give better sampling efficiency (i.e. smaller E(N)) than most of other multinomial selection procedures in indifference-zone approach. The sampling and stopping rules of BKS procedure can also be utilized in subset selection approach. Thus far no article has been published using BKS sequential procedure for multinomial selection in subset selection approach. It is desirable to study the characteristic performances of both the open and truncated versions of BKS procedure in multinomial subset selection approach.

REFERENCES

Bechhofer, R.E. and Chen, P. (1987). A curtailed sequential procedure for subset selection of multinomial cells. Technical Report No. S-42, Department of Mathematics, Syracuse University, Syracuse, NY.

Bechhofer, R.E. and Goldsman, D.M. (1985). Truncation of the Bechhofer-Diefer-Sobel sequential procedure for selecting the

multinomial event which has the largest probability. *Communications in Statistics--Simulation and Computation.* B14 (2), 283-315.

Bechhofer, R.E. and Goldsman, D.M. (1986). Truncation of the Bechhofer-Kiefer-Sobel sequential procedure for selecting the multinomial event which has the largest probability (II): Extended tables and an improved procedure. *Communications in Statistics--Simulation and Computation.* B15 (3), 829-851.

Chen, P. (1986). Inverse sampling subset selection for multinomial distribution. *American Journal of Mathematical and Management Sciences,* Vol. 6, Nos. 1 and 2, 41-64.

Chen, P. (1987). Truncated inverse sampling procedure for multinomial subset selection. Tech. Report No. S-41, Department of Mathematics, Syracuse University, Syracuse, N.Y.

Gibbons, J.D., Olkin, I., and Sobel, M. (1977). *Selecting and ordering populations: a new statistical methodology.* Wiley, New York.

Goldsman, D.M. (1984a). On selecting the best of K systems: An expository survey of indifference-zone multinomial procedures. *Proceedings of the 1984 Winter Simulation Conference,* 107-112.

Goldsman, D.M. (1984b). A multinomial ranking and selection procedure: Simulation and Applications. *Proceedings of the 1984 Winter Simulation Conference,* 259-264.

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

Gupta, S.S. and Panchapakesan, S. (1979). *Multiple decision procedures.* Wiley, New York.

Panchapakesan, S. (1971). On a subset selection procedure 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.

AUTHOR'S BIOGRAPHY

PINYUEN CHEN is an Associate Professor in the Department of Mathematics at Syracuse University. He received his M.S. degree from the University of Miami in 1978, a Ph.D. degree in statistics from the University of California at Santa Barbara in 1982. His research interests include ranking and selection, group testing, and paired comparison.

Pinyuen Chen
Department of Mathematics
Syracuse University
Syracuse, NY 13244-1150
(315) 443-1573