A FINITE URN MODEL FOR SELECTING THE POPULATION WITH THE LARGES--ETC(U)

JAN 82 K ALAM, M H RIZVI

N00014-75-C-0451
A FINITE URN MODEL FOR SELECTING
THE POPULATION WITH THE
LARGEST α-QUANTILE

*Khursheed Alam and M. Haseeb Rizvi
Clemson University and Sisorex Inc.

DEPARTMENT OF MATHEMATICAL SCIENCES
CLEMSON UNIVERSITY

Technical Report #378

January, 1982

*This work was supported in part by the U. S. Office of Naval Research under Contract N00014-75-C-0451.
A Finite Urn Model For Selecting

The Population With The Largest α-Quantile

Khursheed Alam and M. Haseeb Rizvi

Clemson University and Sisorex Inc.

ABSTRACT

Several procedures for ranking populations according to the quantile of a given order have been discussed in the literature. These procedures deal with continuous distributions. This paper deals with the problem of selecting a population with the largest α-quantile from \( k \geq 2 \) finite populations, where the size of each population is known. A selection rule is given based on the sample quantiles, where the samples are drawn without replacement. A formula for the minimum probability of a correct selection for the given rule, for a certain configuration of the population α-quantiles, is given in terms of the sample numbers.

Key Words: Ranking and Selection, Hypergeometric Distribution, Quantiles, Finite Population.

This work was supported by the U. S. Office of Naval Research under Contract N00014-75-C-0451.
1. **Introduction.** This paper deals with the problem of selecting a population from amongst several populations whose distribution functions are unknown. In many situations there are certain parameters of the distributions which are of interest, such as, the mean and variance. Then the populations are ranked according to the values of those parameters and the selection is based on their estimated values. It may be required to select, for example, the population with the smallest variance or the population with the largest mean. Whereas, the mean, variance, a scale parameter and a location parameter are parameters of general interest, it is sometimes justifiable to rank the populations according to the quantile of a given order, especially when the standard parameters do not exist. Rizvi, Sobel and Woodworth (1968) have discussed a comparison of populations in terms of $\alpha$-quantiles. Sobel (1967) and Rizvi and Sobel (1967) have considered the problem of selecting the population with the largest $\alpha$-quantile and the problem of selecting a subset of $k \geq 2$ populations which includes the population with the largest $\alpha$-quantile.

The papers mentioned above pertain to large populations. In this paper we consider the problem of selecting the population with the largest $\alpha$-quantile from $k \geq 2$ finite populations. The populations are sampled by the method of sampling without replacement. A practical situation in which the problem may arise is illustrated by the following example: Suppose that the Department of Education in a certain state is interested in selecting one of several schools in an area to implement
2.

a special training program for exceptionally bright students. As the program involves only exceptionally bright students, the school with the largest 75th quantile score on a special merit examination (SME) may be selected for the training program. A random sample of n students is taken from each school and the selected students are given the SME. The selection of the school for the special training program would depend on the SME scores.

The selection problem is formulated as follows: Let \( \pi_1, \ldots, \pi_k \) denote \( k > 2 \) finite populations and let \( N_i \) denote the size of \( \pi_i \), \( i = 1, \ldots, k \). The numbers \( N_1, \ldots, N_k \) are assumed to be known. The members of each population are ranked according to some characteristic value. Let \( X_{i,m_i} \) denote the \( m_i \)th smallest value of the elements of \( \pi_i \) for \( m_i = \alpha N_i \), where \( \alpha \) is a given positive fraction. Then \( X_{i,m_i} \) represents the \( \alpha \)-quantile of \( \pi_i \). It is assumed that \( \alpha N_i \) is integer valued for each \( i = 1, \ldots, k \). We shall call the population associated with the largest value of \( X_{i,m_i} \) as the best population. It is required to select the best population.

Suppose that a sample of \( n_i < N_i \) elements is drawn from \( \pi_i \) without replacement. Let \( S_i \) denote the sample \( \alpha \)-quantile, that is, the \( \alpha n_i \) smallest value in the sample. It is assumed for simplicity that \( \alpha n_i \) is integer values for each \( i = 1, \ldots, k \). We select the population associated with the largest value of \( S_i \) for the best population. We shall call this procedure \( S \).

Let \( \epsilon \) be a positive fraction, such that \( \epsilon < \alpha < 1-\epsilon \). We assume for simplicity that \( \epsilon N_i \) is integer valued for each \( i=1, \ldots, k \). Let \( C_\epsilon \) denote a configuration of population quantiles given by
3.

(1.1) \[ X_j, (a+\epsilon)N_j < X_i, (a-\epsilon)N_i, j = 1, \ldots, i-1, i+1, \ldots, k \]

when \( \pi_i \) is the best population, \( i=1, \ldots, k \). This is called a preference zone. It is required that the probability of a correct selection (PCS) for the procedure \( S \) should satisfy the relation \( \text{PCS} \geq P^* \) in the preference zone, where \( P^* \) is a pre-assigned number, such that \( \frac{1}{k} < P^* < 1 \).

The value of the PCS for the procedure \( S \) depends on the sample numbers \( n_1, \ldots, n_k \). In the following section we derive a formula for the minimum value of the PCS inside \( C_\epsilon \). The sample numbers needed to satisfy the probability requirement for \( S \) are determined from the given formula.

2. Procedure \( S \).

Suppose that \( \pi_i \) is the best population. It is easy to see that the probability of a correct selection for the procedure \( S \), given the configuration \( C_\epsilon \), is minimized when

\[
(2.1) \quad X_i, (a-\epsilon)N_i-1 \leq X_j, \leq X_j, (a+\epsilon)N_j < X_i, (a-\epsilon)N_i \leq X_i, N_i \leq X_j, (a+\epsilon)N_j + 1
\]

for \( j = 1, \ldots, i-1, i+1, \ldots, k \). Therefore, given \( C_\epsilon \), the minimum probability of a correct selection is equal to

\[
(2.2) \quad \min_{i=1, \ldots, k} \min_{j=1, \ldots, i-1, i+1, \ldots, k} P \{ S_j < S_i \mid \pi_i \text{ is best population} \} = \min (R_1, \ldots, R_k)
\]
where

\[ R_i = \sum_{r=0}^{\alpha n_i - 1} \frac{(\alpha - \epsilon)N_i - 1}{n_i - r} \frac{(1 - \alpha + \epsilon)N_i + 1}{n_i} \times \]

\[ \prod_{j=1, \ldots, i-1, i+1, \ldots, k} \frac{(\alpha + \epsilon)N_j}{n_i - r} \frac{(1 - \alpha - \epsilon)N_j}{n_j} \]

\[ = Q(\alpha n_i - 1; n_i, (\alpha - \epsilon)N_i - 1, N_i) \times \]

\[ \prod_{j=1, \ldots, i-1, i+1, \ldots, k} (1 - Q(\alpha n_j - 1; n_j, (\alpha + \epsilon)N_j, N_j)) \]

and

\[ Q(x; n, M, M) = \sum_{r=0}^{x} \frac{\binom{M}{r} \binom{N-M}{n-r}}{\binom{N}{n}} \]

denotes the cumulative distribution function of the hypergeometric distribution.

The value of the PCS given by (2.2) and (2.3) can be computed from the tables of the hypergeometric distribution prepared by Lieberman and Owen (1961). If \( N_i \) is large and \( n_i \) is small compared to \( N_i \) for each \( i \), then \( R_i \) is approximately given by

\[ R_i = I_{1-\alpha+\epsilon}((1-\alpha)n_i+1, \alpha n_i) \prod_{j=1, j \neq i} I_{\alpha+\epsilon}(\alpha n_j, (1-\alpha)n_j+1). \]

where

\[ I_p(a, b) = \frac{\Gamma(a+b)}{\Gamma(a)(b)} \int_0^p x^{a-1}(1-x)^{b-1} \, dx \]

denotes the incomplete beta function. On the other hand, if \( n_i \) and \( N_i \) are large for each \( i \), such that \( n_i/N_i \rightarrow \xi_i \), say, where \( \xi_i < 2(1-\max (\frac{\epsilon}{\alpha}, \frac{\epsilon}{1-\alpha})) \), then \( R_i \) is approximately given by
5.

(see Wise (1954))

\[(2.6) \quad R_i = I_{w_i} ((1-\alpha)n_i+1, an_i) \prod_{j=1, j\neq i}^k I_{w_j} (an_j, (1-\alpha)n_j+1)\]

where

\[w_i = 1 - \alpha + \frac{2\epsilon}{2-\xi_i}\]

\[w_j = 1 - \alpha + \frac{2\epsilon}{2-\xi_j}\]

Note that (2.6) reduces to (2.5) for \(\xi_1 = \ldots = \xi_k = 0\).

If the \(k\) populations are of equal size \(N\), say, we take \(n_1 = n_2 = \ldots = n_k = n\), say. Then from (2.2) and (2.3) the minimum value of the PCS is equal to

\[(2.7) \quad \frac{\sum_{r=n}^{n-1} (\alpha\epsilon)^{n-r} (1-\alpha\epsilon)^{n-1}) \times (\sum_{r=an}^{n-1} (\alpha\epsilon)^{n-r} (1-\alpha\epsilon)^{n-1})^{k-1}/(\binom{n}{k})^k .\]

Table I below gives the minimum value of \(n\) for which PCS > \(P^*\) for \(\alpha = .50, \epsilon = .05, .10, P^* = .75, .95, .99, k = 1 \ldots 5\) and \(N = 30, 40, 50, 100, 200, 400\). The minimum value of the PCS for the given \(n\) is also shown in the table. It is seen from the table that in some cases there is considerable difference between the minimum value of the PCS and the prescribed value \(P^*\), due to the discrete nature of the hypergeometric distribution. However, the discrepancy is reduced for large \(N\).
6.

In practice, the given populations would vary in size. Therefore, we consider the value of $R_i$, given by (2.6). It is easily shown that for large $m$ the beta integral $I_p(\alpha m, (1-\alpha)m+1)$ is increasing (decreasing) in $m$ for $p < (>) \alpha$. Therefore, $R_i$ is an increasing function of $n_1, \ldots, n_k$ when the sample numbers are large. Thus a smallest sample number can prescribed for each population to meet the probability requirement for the procedure $S$ in the general case when the population size varies. In this case it would be interesting to find an optimal distribution of the sample numbers, given \( \sum_{k=1}^{k} n_i = n \), say. This is an exercise in linear programming, where we maximize a function $f(\xi_1, \ldots, \xi_k)$ subject to the constraints $\sum_{i=1}^{k} \xi_i N_i = n$ and $0 < \xi_i < 1$, $i = 1, \ldots, k$. Here $f(\xi_1, \ldots, \xi_k) = \min(R_1, \ldots, R_k)$, where $R_i$ is given by (2.6) with the substitution $\xi \geq N_i$ for $n_\ell$, $\ell = 1, \ldots, k$.

Remark 1. We observe that the value of the PCS given by (2.2) where $R_i$ is given by (2.5) represents the minimum probability of a correct selection when the samples are drawn with replacement.

Remark 2. It is seen from (2.1) that the value of the PCS is equal to 1 if

\[
\frac{n_i}{N_i} > \max(1 - \frac{\xi}{\alpha}, \frac{\alpha^* \xi}{1-\alpha}) \quad , \quad i = 1, \ldots, k.
\]

Therefore, a correct selection is obtained with probability 1 with $n_i < N_i$ for each $i$.

Acknowledgement. The subject of this article was suggested to the authors by Professor J. Sedransk. Professor M. Sobel helped in the formulation of the problem.
REFERENCES


Table 1 - Minimum Value of $n$ for $PCS \geq P^*$ ($\alpha = .50$)

<table>
<thead>
<tr>
<th>K</th>
<th>N</th>
<th>P*</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>30</td>
<td>.75</td>
<td>22</td>
<td>28</td>
<td>28</td>
<td>24</td>
</tr>
<tr>
<td>PCS</td>
<td>.750</td>
<td>1.000</td>
<td>1.000</td>
<td>.809</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>40</td>
<td>.788</td>
<td>.955</td>
<td>1.000</td>
<td>.783</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>PCS</td>
<td>.784</td>
<td>.950</td>
<td>.990</td>
<td>.755</td>
<td>.953</td>
<td>.990</td>
</tr>
<tr>
<td>50</td>
<td>.778</td>
<td>.976</td>
<td>1.000</td>
<td>.781</td>
<td>.976</td>
<td>1.000</td>
</tr>
<tr>
<td>PCS</td>
<td>.764</td>
<td>.954</td>
<td>.992</td>
<td>.769</td>
<td>.952</td>
<td>.995</td>
</tr>
<tr>
<td>100</td>
<td>.752</td>
<td>.955</td>
<td>.991</td>
<td>.756</td>
<td>.954</td>
<td>.992</td>
</tr>
<tr>
<td>PCS</td>
<td>.755</td>
<td>.950</td>
<td>.990</td>
<td>.753</td>
<td>.952</td>
<td>.991</td>
</tr>
</tbody>
</table>

$\epsilon = .10$

<table>
<thead>
<tr>
<th>K</th>
<th>N</th>
<th>P*</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>30</td>
<td>.784</td>
<td>16</td>
<td>22</td>
<td>24</td>
<td>18</td>
</tr>
<tr>
<td>PCS</td>
<td>.784</td>
<td>.980</td>
<td>1.000</td>
<td>.794</td>
<td>.972</td>
<td>1.000</td>
</tr>
<tr>
<td>40</td>
<td>.771</td>
<td>.958</td>
<td>.994</td>
<td>.760</td>
<td>.972</td>
<td>.992</td>
</tr>
<tr>
<td>PCS</td>
<td>.773</td>
<td>.951</td>
<td>.993</td>
<td>.752</td>
<td>.958</td>
<td>.990</td>
</tr>
<tr>
<td>50</td>
<td>.753</td>
<td>.952</td>
<td>.992</td>
<td>.766</td>
<td>.957</td>
<td>.993</td>
</tr>
<tr>
<td>PCS</td>
<td>.755</td>
<td>.951</td>
<td>.991</td>
<td>.755</td>
<td>.953</td>
<td>.991</td>
</tr>
<tr>
<td>100</td>
<td>.755</td>
<td>.951</td>
<td>.991</td>
<td>.755</td>
<td>.953</td>
<td>.991</td>
</tr>
<tr>
<td>PCS</td>
<td>.751</td>
<td>.953</td>
<td>.990</td>
<td>.765</td>
<td>.953</td>
<td>.990</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>K</th>
<th>N</th>
<th>P*</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>200</td>
<td>.755</td>
<td>26</td>
<td>62</td>
<td>88</td>
<td>34</td>
</tr>
<tr>
<td>PCS</td>
<td>.755</td>
<td>.951</td>
<td>.991</td>
<td>.755</td>
<td>.953</td>
<td>.991</td>
</tr>
<tr>
<td>400</td>
<td>.751</td>
<td>.953</td>
<td>.990</td>
<td>.765</td>
<td>.953</td>
<td>.990</td>
</tr>
</tbody>
</table>
**A Finite Urn Model for Selecting the population with the largest $\alpha$-quantile**

**Authors:** Khursheed Alam and M. Haseeb Rizvi

**Performing Organization Name and Address:**
Clemson University
Dept. of Mathematical Sciences
Clemson, South Carolina 29631

**Contract or Grant Number:** N00014-75-C-0451

**Program Element, Project, Task Area & Work Unit Numbers:** NR 365-049

**Report Date:** January, 1982

**Number of Pages:** 11

**Security Classification:** Unclassified

**Distribution Statement:** Approved for public release; distribution unlimited.

**Keywords:** Ranking and Selection, Hypergeometric Distribution, Quantiles, Finite Population.

**Abstract:** Several procedures for ranking populations according to the quantile of a given order have been discussed in the literature. These procedures deal with continuous distributions. This paper deals with the problem of selecting a population with the largest $\alpha$-quantile from $k \geq 2$ finite populations, where the size of each population is known. A selection rule is given based on the sample quantiles, where the samples are drawn without replacement. A formula for the minimum probability of a correct (over)
selection for the given rule, for a certain configuration of the population $\alpha$-quantiles, is given in terms of the sample numbers.