The efficiency of subset selection of an epsilon-best uniform population relative to selection of the best one

EINDHOVEN UNIVERSITY OF TECHNOLOGY
Department of Mathematics and Computing Science

Memorandum COSOR 93-12

The efficiency of subset selection of an ε-best uniform population relative to selection of the best one

Paul van der Laan

Eindhoven, April 1993
The Netherlands

Eindhoven University of Technology
Department of Mathematics and Computing Science
Probability theory, statistics, operations research and systems theory
P.O. Box 513
5600 MB Eindhoven - The Netherlands

Secretariat: Dommel building 0.03
Telephone: 040-47 3130

ISSN 0926 4493

The efficiency of subset selection of an ε-best uniform population relative to selection of the best one

Paul van der Laan

Summary
Assume $k(\geq 2)$ uniform populations are given on $(\mu_i - \frac{1}{2},\ \mu_i + \frac{1}{2})$ with location parameter $\mu_i \in \mathbb{R}^1$, $i = 1, \cdots, k$. The best population is defined as the population with the largest value of the location parameter. In ε-best population (with $\varepsilon \geq 0$) is a population with location parameter on a distance not larger than ε from the largest value of μ. It is possible to consider subset selection for an ε-best population relative to subset selection for the best one. The relative efficiency is defined and computed in dependence of k and ε for some values of the confidence level P^* of selection.

AMS Subject Classification: Primary 62F07, Secondary 62E15.

Key Words: Subset selection, uniform populations, ε-best population, relative efficiency

1. Introduction

Given are $k(\geq 2)$ uniform populations on $(\mu_i - \frac{1}{2},\ \mu_j + \frac{1}{2})$ with unknown location parameter $\mu_i \in \mathbb{R}^1, i = 1, \cdots, k$. The random variables associated with these populations are denoted by $X_1, \cdots, X_k$, with cumulative distribution functions $F(x - \mu_i), i = 1, \cdots, k$, respectively. The ranked variables are denoted by $X_{[1]} \leq \cdots \leq X_{[k]}$. The ordered parameters are denoted by $\mu_{[1]} \leq \cdots \leq \mu_{[k]}$. The best population is the population corresponding with $\mu_{[k]}$. We suppose the best population is unique, otherwise a suitable flagging is used. An ε-best population (with $\varepsilon \geq 0$) is a population with parameter value larger than or equal to $\mu_{[k]} - \varepsilon$. In order to select the best population or an ε-best population Gupta's subset selection approach is used. For more details we refer to Gupta (1965), Butler and Butler (1987), and Van der Laan (1991, 1992). The following selection rules are used. In order to select the best population, rule R_B is used with

R_B: Select population i $(i = 1, \cdots, k)$ in subset if an only if $X_i \geq X_{[k]} - d$, with the selection constant $d \geq 0$.

In order to select an ε-best (almost best) population the next rule is used:

R_A: Select population i $(i = 1, \cdots, k)$ in subset if an only if $X_i \geq X_{[k]} - d + \varepsilon$, with $0 \leq \varepsilon \leq d < 1$.

2. The relative efficiency

In this section we shall define the relative efficiency of the selection rule R_A compared with the selection rule R_B, and derive a general expression for uniform populations. Some properties of this relative efficiency will be presented and proved.

Definition 2.1. The relative efficiency RE of the selection rule R_A relative with respect to the selection rule R_B is defined as

$$RE(k, P^*, \varepsilon) = \frac{\mathcal{E}_\mu\{S|LFC; d\}}{\mathcal{E}_\mu\{S|LFC; d - \varepsilon\}},$$

where $\mu = (\mu_1, \cdots, \mu_k)$, S is the size of the selected subset using selection rule R_B, and $\mathcal{E}_\mu\{S|LFC; c\}$ denotes the expected subset size under the Least Favorable Configuration (LFC) and with selection constant c.

It is well-known that under the LFC in the case considered we have $\mu_{[1]} = \mu_{[k]}$. Using this property the next theorem can easily be proved.

Theorem 2.1. $RE(k, P^*, \varepsilon) = \frac{kP^*}{1-(d-\varepsilon)^k + k(d-\varepsilon)}$,

with $0 \leq \varepsilon \leq d = d(P^*) < 1$.

Proof: We have for the numerator of RE and for general d:

$$\begin{aligned}
\mathcal{E}_\mu\{S|LFC;d\} &= \sum_{i=1}^{k} P(X_i \geq X_{[k]} - d|\mu_{[1]} = \mu_{[k]}) \\
&= \sum_{i=1}^{k} P(X_j \leq X_i + d;\ j = 1,\cdots,k;\ j \neq i|\mu_{[1]} = \mu_{[k]}) \\
&= \sum_{i=1}^{k} \int \prod_{\substack{j=1 \\ j\neq 1}}^{k} F(x + d - \mu_j + \mu_i|\mu_{[1]} = \mu_{[k]})dF(x) \\
&= \sum_{i=1}^{k} \int \prod_{\substack{j=1 \\ j\neq 1}}^{k} F(x+d)dF(x) \\
&= k \textstyle\int F^{k-1}(x+d)dF(x)\ ,
\end{aligned}$$

and after some elementary computations we get

$$\mathcal{E}_\mu\{S|LFC;\ d\} = \begin{cases} 1 - d^k + kd & \text{for} \quad 0 \leq d < 1 \\ k & \text{for} \quad 1 \leq d. \end{cases}$$

When we take $d = d(P^*)$ such that the probability requirement for the subset selection is fulfilled, then $\mathcal{E}_\mu\{S|LFC;\ d = d(P^*)\}$ is equal to the maximum value of $\mathcal{E}_\mu\{S\}$, which is equal to kP^* (see Gupta (1965)), from which the result follows immediately. □

Lemma 2.1. The relative efficiency $RE(k, P^*, \varepsilon)$ satisfies

i) for each $k^{-1} < P^* < 1$ and $0 \leq \varepsilon < P^*$, $RE(k, P^*, \varepsilon)$ is strictly increasing in k,

ii) for each $2 \leq k < \infty$ and $k^{-1} < P^* < 1$, $RE(k, P^*, \varepsilon)$ is strictly increasing in ε (with $0 \leq \varepsilon < P^*$),

iii) for each $2 \leq k < \infty$ and $0 \leq \varepsilon < P^*$, $RE(k, P^*, \varepsilon)$ is strictly decreasing in P^*.

Proof: For i) it has to be proved that ($c := d - \varepsilon$, with $0 \leq c < 1$)

$$\frac{(k+1)P^*}{1 - c^{k+1} + (k+1)c} > \frac{kP^*}{1 - c^k + kc}$$

or that

$$g(c) := kc^{k+1} - (k+1)c^k + 1 > 0 \ \text{ for integer } k \geq 2$$

It is easy to see that $g(0) = 1$, $g(1) = 0$ and $\frac{dg(c)}{dc} = k(k+1)c^{k-1}(c-1) < 0$, and i) follows immediately.
Part ii) follows from the fact that for $m(c) := 1 - c^k + kc$ the following holds: $m(0) = 1$, $m(1) = k$ and $\frac{dm(c)}{dc} = -kc^{k-1} + k > 0$, and that c is strictly decreasing in ε.

In order to prove iii), it is sufficient to prove that RE is decreasing function of d, because $d = d(P^*)$ is an increasing function of P^*. In order to prove that

$$\frac{1 - d^k + kd}{1 - (d - \varepsilon)^k + k(d - \varepsilon)}$$

is a decreasing function of d, notice that $g(c) := 1 - c^k + kc$ is an increasing function of c with $\frac{d^2 g(c)}{dc^2} < 0$, and the result follows immediately. □

Lemma 2.2. $\lim_{k \to \infty} RE(k, P^*, \varepsilon) = \frac{P^*}{P^* - \varepsilon}$, with $0 \leq \varepsilon < P^* < 1$.

Proof: The selection constant $d = d(P^*)$ (with $0 < d < 1$) follows from the equation

$$1 - d^k + kd = kP^* \ ,$$

so d can be written as

$$d = P^* - \frac{1}{k} + \frac{d^k}{k} \ .$$

For large values for k we have $d \approx P^*$, and the result follows immediately. □

Lemma 2.3. $\lim_{k \to \infty} RE(k, P^*, \varepsilon)$ for fixed P^* is a strictly increasing function of ε, and for fixed $\varepsilon < P^*$ a strictly decreasing function of P^*.

Proof: The result follows immediately. □

3. Numerical results

The relative efficiency $RE(k, P^*, \varepsilon)$ has been computed for $P^* = .80, .90, .95, .99$ and $k = 2(1)5, 10, 25, 100, 1000, 2000$. The values considered for ε are $\sqrt{12}\varepsilon = .1, .2, .5(.5)2$. In this way ε has been expressed as a factor times the standard deviation of the distribution. Numerical results can be found in the tables 1, 2, 3 and 4 for $P^* = .80, .90, .95$, and $.99$, respectively.

Table 1. $RE(k, P^*, \varepsilon)$ for $P^* = .80$.

k	d	$\sqrt{12}\ \varepsilon$ .1	.2	.5	1	1.5	2
2	.367544	1.024	1.050	1.146	1.389		
3	.511195	1.028	1.059	1.169	1.449	1.945	
4	.577880	1.031	1.064	1.186	1.488	1.027	3.193
5	.618034	1.032	1.068	1.196	1.513	1.078	3.324
10	.702946	1.036	1.075	1.215	1.556	1.163	3.546
25	.760042	1.037	1.078	1.220	1.564	1.180	3.592
100	.790000	1.037	1.078	1.220	1.565	1.180	3.593
1000	.799000	1.037	1.078	1.220	1.565	1.180	3.593
2000	.799500	1.037	1.078	1.220	1.565	1.180	3.593
$\lim_{k\to\infty} RE = P^*/(P^*-\varepsilon)$	.80	1.0374	1.0778	1.2201	1.5646	2.1799	3.5931

Table 2. $RE(k, P^*, \varepsilon)$ for $P^* = .90$.

k	d	$\sqrt{12}\ \varepsilon$ .1	.2	.5	1	1.5	2
2	.552786	1.015	1.032	1.091	1.234	1.469	
3	.664450	1.019	1.040	1.116	1.302	1.605	2.142
4	.715533	1.021	1.045	1.133	1.346	1.695	2.319
5	.746303	1.023	1.049	1.145	1.377	1.755	2.440
10	.812545	1.029	1.060	1.174	1.443	1.877	2.685
25	.860947	1.032	1.068	1.190	1.470	1.923	2.781
100	.890000	1.033	1.069	1.191	1.472	1.927	2.789
1000	.899000	1.033	1.069	1.191	1.472	1.927	2.789
2000	.899500	1.033	1.069	1.191	1.472	1.927	2.789
$\lim_{k\to\infty} RE = P^*/(P^*-\varepsilon)$	.90	1.0331	1.0685	1.1910	1.4722	1.9272	2.7894

Table 3. $RE(k, P^*, \varepsilon)$ for $P^* = .95$.

k	d	$\sqrt{12}\ \varepsilon$ .1	.2	.5	1	1.5	2
2	.683772	1.010	1.021	1.063	1.163	1.321	1.581
3	.767176	1.013	1.028	1.085	1.225	1.450	1.824
4	.804967	1.016	1.033	1.101	1.269	1.539	1.992
5	.827691	1.017	1.037	1.113	1.302	1.603	2.110
10	.876878	1.023	1.049	1.147	1.381	1.747	2.378
25	.914253	1.029	1.061	1.173	1.427	1.823	2.521
100	.940021	1.031	1.065	1.179	1.436	1.837	2.549
1000	.949000	1.031	1.065	1.179	1.437	1.838	2.549
2000	.949500	1.031	1.065	1.179	1.437	1.838	2.549
$\lim_{k\to\infty} RE = P^*/(P^* - \varepsilon)$	.95	1.0313	1.0647	1.1792	1.4365	1.8376	2.5493

Table 4. $RE(k, P^*, \varepsilon)$ for $P^* = .99$.

k	d	$\sqrt{12}\ \varepsilon$ .1	.2	.5	1	1.5	2
2	.858579	1.005	1.010	1.032	1.091	1.186	1.335
3	.898260	1.006	1.014	1.048	1.141	1.294	1.539
4	.916016	1.008	1.018	1.061	1.181	1.376	1.691
5	.926649	1.009	1.021	1.072	1.212	1.439	1.806
10	.949659	1.013	1.031	1.108	1.304	1.606	2.096
25	.967522	1.020	1.046	1.147	1.377	1.723	2.301
100	.981554	1.028	1.060	1.169	1.408	1.772	2.390
1000	.989000	1.030	1.062	1.171	1.412	1.777	2.399
2000	.989500	1.030	1.062	1.171	1.412	1.777	2.399
$\lim_{k\to\infty} RE = P^*/(P^* - \varepsilon)$	.99	1.0300	1.0619	1.1707	1.4116	1.7774	2.3991

4. Concluding remarks

From the tables it follows that for the smaller values of P^* the relative efficiency tends rather rapidly to the limit value. Even for $P^* = .95$ the relative difference between the actual value and the limit is for $k \geq 25$ smaller than 1.1%.

The relative efficiency increases rather rapidly from a value a little bit larger than 1 to, for $k = 10$, a value of approximately 3.5, 2.7, 2.4, and 2.1 for $P^* = .80$, .90, .95, and .99, respectively.

References

Butler, K.L. and D.G. Butler (1987). Tables for selecting the best population. Queensland Biometrical Bulletin 2, Queensland Department of Primary Industries, Brisbane, QLD 4001, Australia.

Gupta, S.S. (1965). On some multiply decision (selection and ranking) rules. Technometrics **7**, 225-245.

van der Laan, P. (1991). The efficiency of subset selection of an almost best treatment. Memorandum COSOR 91-19, Department of Mathematics and Computing Science, Eindhoven University of Technology.

van der Laan, P. (1992). Subset selection of an almost best treatment. Biometrical Journal **34**, 647-656.

List of COSOR-memoranda - 1993

Number	Month	Author	Title
93-01	January	P. v.d. Laan C. v. Eeden	Subset selection for the best of two populations: Tables of the expected subset size
93-02	January	R.J.G. Wilms J.G.F. Thiemann	Characterizations of shift-invariant distributions based on summation modulo one.
93-03	February	Jan Beirlant John H.J. Einmahl	Asymptotic confidence intervals for the length of the shortt under random censoring.
93-04	February	E. Balas J. K. Lenstra A. Vazacopoulos	One machine scheduling with delayed precedence constraints
93-05	March	A.A. Stoorvogel J.H.A. Ludlage	The discrete time minimum entropy H_∞ control problem
93-06	March	H.J.C. Huijberts C.H. Moog	Controlled invariance of nonlinear systems: nonexact forms speak louder than exact forms
93-07	March	Marinus Veldhorst	A linear time algorithm to schedule trees with communication delays optimally on two machines
93-08	March	Stan van Hoesel Antoon Kolen	A class of strong valid inequalities for the discrete lot-sizing and scheduling problem
93-09	March	F.P.A. Coolen	Bayesian decision theory with imprecise prior probabilities applied to replacement problems
93-10	March	A.W.J. Kolen A.H.G. Rinnooy Kan C.P.M. van Hoesel A.P.M. Wagelmans	Sensitivity analysis of list scheduling heuristics
93-11	March	A.A. Stoorvogel J.H.A. Ludlage	Squaring-down and the problems of almost-zeros for continuous-time systems
93-12	April	Paul van der Laan	The efficiency of subset selection of an ε-best uniform population relative to selection of the best one