Some generalized subset selection procedures

EINDHOVEN UNIVERSITY OF TECHNOLOGY
Department of Mathematics and Computing Science

Memorandum COSOR 93-21

Some generalized subset selection procedures

P. van der Laan
C. van Eeden

Eindhoven, July 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

SOME GENERALIZED SUBSET SELECTION PROCEDURES

Paul van der Laan
Eindhoven University of Technology
The Netherlands

and

Constance van Eeden
University of British Columbia and
Université du Québec à Montréal
Canada

Summary

In this paper some generalizations of Gupta's subset selection procedure are discussed. Assume $k(\geq 2)$ populations are given and assume that the associated random variables have distributions with unknown location parameters θ_i, $i = 1, ..., k$. The ordered parameters are denoted by $\theta_{[1]} \leq ... \leq \theta_{[k]}$. On the basis of independent samples from these populations, Gupta (1965) selects a subset, as small as possible, which contains, with probability at least P^*, the best population, i.e. the one with the largest location parameter, $\theta_{[k]}$. The two generalizations discussed in this paper are those of van der Laan (1991, 1992a, b) and of van der Laan and van Eeden (1993). Each one of these is designed to give a smaller expected subset size, $\mathcal{E}S$, than Gupta's procedure, for which $\mathcal{E}S$ is large when $\theta_{[k]}$ is close to the other θ_i's. The procedure of van der Laan (1992a) selects, with probability at least P^*, an ε-best population whose location parameter is at least $\theta_{[k]} - \varepsilon$ (with $\varepsilon \geq 0$). Some efficiency results for normal populations, comparing van der Laan's procedure with Gupta's, are presented. The procedure of van der Laan and van Eeden (1993) uses a loss function and it upperbounds either the expected loss or the expected subset size, or both. The loss is taken as zero when the subset contains an ε-best population and as an increasing function of $\theta_{[k]} - \varepsilon - \max\{\theta_i \mid i\text{-th population in the subset}\}$ if not. Some properties of this procedure, for the case of two normal populations, are presented.

1. Introduction

In many situations an important problem with wide practical application is the selection of the "best" population from among a set of $k (k \geq 2)$ populations, where "best" is associated with the maximal (or minimal) value of an unknown parameter. In such a case, statistical selection methodology can provide a model for the problem which enables the experimenter to realistically formulate his question concerning the best population and solve it in an adequate way.
We assume that independent samples, one from each population, are given, that the parameter of interest is the population mean, that the selection is based on the sample means and that the best population is the one with the largest mean. There are two main approaches to this problem: the indifference zone approach of Bechhofer (1954) and the subset selection approach of Gupta (1965). For general theoretical considerations concerning these methodologies we refer the reader to Gupta and Panchapakesan (1979).
The indifference zone approach has as its goal to indicate the best population. The procedure selects the population which gave the largest sample mean. The probability requirement is that the probability of correct selection is at least $P^*(k^{-1} < P^* < 1)$, whenever the best population is at least δ^* away from the second best for some given positive δ^*. This minimal probability P^* can only be guaranteed if the common sample size n is large enough.
The subset selection approach has as its goal the selection of a non-empty subset, as small as possible, which contains the best population with a predetermined confidence level. A correct selection (CS) in this context is the selection of a subset which contains the best population. The size of the subset is random and the confidence requirement has to be met for all parameter configurations. When a large confidence level is required, one may expect that the size of the selected subset will be large. Also, when the parameter values are close together a large expected subset size will result.
Smaller (expected) subset sizes can be obtained by not restricting oneself to the selection of the best population, but to allow the selection of an almost-best population to be a correct selection. This aspect of subset selection is treated in Section 2. The use of loss functions is discussed in Section 3. Section 4 contains some final conclusions.

The following notation will be used: the populations are denoted by $\pi_1, ..., \pi_k$, the observed means by $X_1, ..., X_k$, the ordered sample means by $X_{[1]} \leq ... \leq X_{[k]}$, the unknown location parameters by $\theta_1, ..., \theta_k$ and the ordered parameters by $\theta_{[1]} \leq ... \leq \theta_{[k]}$. Further, the population associated with $\theta_{[i]}$ is denoted by $\pi_{[i]}$ and $\pi_{[k]}$ is the best population.

2. Selection of an almost-best population

In this section, a correct selection is the selection of a subset which contains at least one ε-best population, where an ε-best population is defined as a population π_i for which $\theta_i \geq \theta_{[k]} - \varepsilon$. Note that, for each $\varepsilon \geq 0$, there exists at least one ε-best population.
As for Gupta's selection procedure, the goal is to select a subset (as small as possible) such that, for all $\theta = (\theta_1, ..., \theta_k), P_\theta(CS) \geq P^*$ for some given $P^*(k^{-1} < P^* < 1)$. The selection rule is given by

put π_i into the subset if and only if $X_i \geq X_{[k]} - c$,

where the selection constant $c \geq 0$. This rule is of the same form as Gupta's - the difference is in the choice of c. Further, if the underlying distributions are normal with known variance σ^2, then it can easily be seen that $c = \sigma d - \varepsilon$, where d is Gupta's selection constant for standard normal populations and the same P^* is used for the two procedures. It can easily be shown that, when using Gupta's procedure, the probability of selecting an ε-best population into the subset is at least equal to the probability of selecting the best population and this for every parameter configuration. A disadvantage of the new procedure is that the least favourable configuration is more complicated than the one for Gupta's procedure.
As an illustration, consider the following two cases of normal populations:

$$\sigma = 1, k = 10, P^* = .90 \text{ and } \varepsilon = .5 \text{ or } \varepsilon = 1.0 \ .$$

The selection constants for these two cases are, resp. $c = 2.483$ and $c = 1.983$ and, using these constants, the minimal (over the parameter configurations) P_θ (subset contains the best population) = .804, resp .655. This illustrates the fact that, with this new procedure, it is possible to get a larger probability of CS than with Gupta's procedure.
For a fixed selection constant c and a fixed ε the efficiency, G, of the ε-best selection procedure is defined as the relative difference in the minimal probabilities of reaching the selection goals - the new one, resp. Gupta's - relative to Gupta's. For the two examples above, one gets $G = (.90 - .804)/.804 = .12$ for the first example and $G = (.90 - .655)/.655 = .37$ for the second.
If the density of the observations is strictly unimodel (i.e. the logarithm of the density is concave) then it can be shown (in the location parameter case) that G is a decreasing function of P^*. For the proof we refer the reader to van der Laan (1992b).
Another way to compare the ε-best selection procedure with Gupta's is to investigate the ratio

$$\sup \mathcal{E}_\theta S_b / \sup \mathcal{E}_\theta S_a \ ,$$

where S_b and S_a are the subset sizes for selecting the best and an ε-best population resp. when using Gupta's procedure and the ε-best selection procedure, respectively, with the same minimal probability of correct selection. For normal populations with standard deviation 1 we get the following results for this ratio:

		P^*		
k	ε	.80	.90	.95
10	.5	1.011	1.007	1.004
	1.0	1.051	1.033	1.021
	2.0	1.292	1.196	1.129
20	0.5	1.007	1.004	1.003
	1.0	1.033	1.021	1.013
	2.0	1.211	1.138	1.090
50	0.5	1.003	1.002	1.001
	1.0	1.018	1.011	1.007
	2.0	1.130	1.084	1.054
100	0.5	1.002	1.001	1.001
	1.0	1.011	1.007	1.004
	2.0	1.089	1.057	1.036
1000	0.5	1.001	1.000	1.000
	1.0	1.002	1.001	1.001
	2.0	1.024	1.015	1.009

3. Selection based on a loss function

Instead of asking for a probability of at least P^* of a correct selection, van der Laan and van Eeden (1993) use a loss function. They take the loss equal to zero when the subset contains an ε-best population and a nondecreasing function, h, of the difference $\theta_{[k]} - \varepsilon - \theta^{[s]}$ if the selected subset does not contain an ε-best population, where $\theta^{[s]} = \max(\theta_i | i$ such that π_i is in the subset). The subset selection goals are expressed in terms of an upper bound on the risk function and/or on the expected subset size. These upper bounds are required to hold either for all or for some θ. The selection rule is of the same form as the one used by Gupta and by van der Laan. The difference is in the value of the selection constant c.
Gupta's and van der Laan's subset selection approaches are obtained as special cases by taking $h(x) \equiv 1$ for all $x \in \mathbb{R}^+$ with $\varepsilon = 0$ for Gupta and $\varepsilon > 0$ for van der Laan.
For the case of two normal populations with equal variances $\sigma^2 = 1$ and $h(x) = x^p$ for some $p > 1$, the selection rule becomes: put $\pi_1(\pi_2)$ into the subset if $X_1 - X_2 > c(< -c)$, otherwise put both populations into the subset. The risk function, R, is given by

$$(\mu - \varepsilon)^p \Phi((-c - \mu)\sqrt{(n/2)}) I(\mu > \varepsilon) ,$$

where $\mu = |\theta_1 - \theta_2|$, Φ is the standard normal distribution function and $I(A)$ is the indicator function of the set A. The expected subset size is given by

$$\mathcal{E}_\theta S_c = \Phi((c - \mu)\sqrt{(n/2)}) + \Phi((c + \mu)\sqrt{(n/2)}) .$$

It can easily be seen that the risk function is a decreasing function of c and it can be shown (see van der Laan and van Eeden (1993)) that the expected subset size is an increasing function of c. So, asking for max $R \leq R_0$ for some given positive R_0 puts a lower bound on c, while asking for max $\mathcal{E}_\theta S_c \leq 1 + \tau_0$ for some $0 < \tau_0 < 1$ puts an upperbound on c. Further, R is a decreasing function of ε for $\mu > \varepsilon$. So, if the bounds max $R \leq R_0$ and max $\mathcal{E}_\theta S_c \leq 1 + \tau_0$ can not be simultaneously satisfied for the chosen ε, then an increase in ε will

make it possible to satisfy these bounds.
Similar considerations hold for the case where the bounds on R and on $\mathcal{E}_\theta S_c$ are to hold only for some values of θ.
In order to apply in practice this new subset selection methodology, tables of R, max R, and $\mathcal{E}_\theta S_c$ are needed. Such tables can be found in van der Laan and van Eeden (1993).

4. Some final conclusions

The performance of selection procedures can be improved by either increasing the sample sizes or by weakening the confidence requirement. If an experimenter is able to specify an $\varepsilon > 0$ such that he is content to have a correct selection defined as a selection with an ε-best population in the subset, then the methodology for an almost-best population can be used. If the introduction of a loss function of the form given in this paper is realistic for a practical problem, then the methodology based on a loss function can be considered as a flexible approach. Both methods are generalizations of Gupta's subset selection procedure and are designed to give a smaller expected subset size. A comparison with Gupta's selection procedure can be found in van der Laan and van Eeden (1993).

References

Bechhofer, R.E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. Ann. Math. Statist. **25**, 16-39.

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

Gupta, S.S. and S. Panchapakesan (1979). Multiple Decision Procedures. Wiley, New York.

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. (1992a). Subset selection of an almost best treatment. Biometrical Journal **34**, 1-10.

van der Laan, P. (1992b). Subset Selection: Robustness and Imprecise Selection. Memorandum COSOR **92-10**. Department of Mathematics and Computing Science, Eindhoven University of Technology.

van der Laan, P. and C. van Eeden (1993). Subset selection with a generalized selection goal based on a loss function. Technical Report # **127**, Department of Statistics, University of British Columbia, Vancouver and Memorandum COSOR **93-15**, Department of Mathematics and Computing Science, Eindhoven of University. Submitted.

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
93-13	April	R.J.G. Wilms	On the limiting distribution of fractional parts of extreme order statistics

Number	Month	Author	Title
93-14	May	L.C.G.J.M. Habets	On the Genericity of Stabilizability for Time-Day Systems
93-15	June	P. van der Laan C. van Eeden	Subset selection with a generalized selection goal based on a loss function
93-16	June	A.A. Stoorvogel A. Saberi B.M. Chen	The Discrete-time H_∞ Control Problem with Strictly Proper Measurement Feedback
93-17	June	J. Beirlant J.H.J. Einmahl	Maximal type test statistics based on conditional processes
93-18	July	F.P.A. Coolen	Decision making with imprecise probabilities
93-19	July	J.A. Hoogeveen J.K. Lenstra B. Veltman	Three, four, five, six or the Complexity of Scheduling with Communication Delays
93-20	July	J.A. Hoogeveen J.K. Lenstra B. Veltman	Preemptive scheduling in a two-stage multiprocessor flow shop is NP-hard
93-21	July	P. van der Laan C. van Eeden	Some generalized subset selection procedures