Subset selection for an epsilon-best population : efficiency results

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics and Computing Science

COSOR-Memorandum 91-20

Subset selection for an ε-best population : Efficiency results

by

Paul van der Laan

ISSN: 0926-4493

The Netherlands

September, 1991

Subset selection for an ε-best population: Efficiency results

Paul van der Laan

Abstract

An almost best or an ε-best population is defined as a population with location parameter on a distance not larger than $\varepsilon(\geq 0)$ from the best population (with largest value of the location parameter). For the subset selection tables with the relative efficiency of selecting an ε-best population relative to selecting the best population are given.
Results are presented for confidence level $P^* = 0.50, 0.80, 0.90, 0.95$ and 0.99; the number of populations $k = 2(1)15(5)50(10)100(50)300(100)500(250)2000$, and $\varepsilon = 0.2, 0.5, 1.0, 1.5$ and 2.0, where P^* is the minimal probability of correct selection.

AMS Subject classification: Primary 62F07; Secondary 62E15.

Key Words and Phrases: subset selection, almost best population, efficiency.

1. Introduction

The tables with relative efficiencies of subset selection for an ε-best population relative to subset selection for the best population have been prepared in order to achieve that these relative efficiencies will become easily accessible to applied statisticians and users of statistical selection procedures.
A subset selection procedure selects a subset from $k(\geq 2)$ populations $\pi_1, \pi_2, ..., \pi_k$, characterized by the location parameters $\theta_1, \theta_2, ..., \theta_k$, respectively. The related independent random variables are denoted by $X_1, X_2, ..., X_k$, respectively. The population with the largest location parameter is called the best population. Tagging is used when there is more than one best population. Subset selection has as its goal to select a non-empty subset, as small as possible, from the k populations in order to include the best population with a certain confidence. The size of the subset is a random variable. The confidence requirement has to be met for all parameter configurations. The selection rule is:

$$\text{select } \pi_i \text{ in subset iff } X_i \geq \max_{1 \leq j \leq k} X_j - d$$

with selection constant $d \geq 0$. This selection constant d has to be determined such that the probability that the subset will contain the best population is at least equal to $P^*(k^{-1} < P^* < 1)$. An ε-best (almost best) population is defined as a population for which the location parameter $\theta_i \geq \theta_{[k]} - \varepsilon$, where $\varepsilon \geq 0$ is given and the ranked location parameters are denoted by $\theta_{[1]} \leq \theta_{[2]} \leq ... \leq \theta_{[k]}$. The selection goal is to select a subset that will contain at least one ε-best population with probability P^*. The selection rule is:

$$\text{select } \pi_i \text{ in subset iff } X_i \geq \max_{1 \leq j \leq k} X_j - c,$$

where the selection constant $c(\geq 0)$ has to be determined such that the probability that the subset will contain at least one ε-best population is at least equal to P^*.

2. Relative efficiency (RE)

In Van der Laan (1991) a comparison of both selection procedures has been proposed on the basis of a relative efficiency coefficient RE, defined as

$$RE = \sup_{\theta \in \Omega} ES_B / \sup_{\theta \in \Omega(\varepsilon)} ES_E \ ,$$

where

S_B: the size of the subset based on subset selection for the best,
S_E: the size of the subset based on subset selection for an ε-best,
Ω: parameter space of $\theta = (\theta_1, \theta_2, ..., \theta_k)$,
$\Omega(\varepsilon) = \{\theta : \theta \in \Omega \text{ and } \theta_{[i]} < \theta_{[k]} - \varepsilon, i = 1, 2, ..., k-1\}$.

We have for Normal populations with standard deviation 1:

$$RE = kP^*\{\int_{-\infty}^{\infty} \Phi^{k-1}(x+c+\varepsilon)\phi(x)dx +$$

$$+(k-1)\int_{-\infty}^{\infty} \Phi^{k-2}(x+c)\Phi(x+c-\varepsilon)\phi(x)dx\}^{-1} ,$$

with $\phi(\cdot)$ and $\Phi(\cdot)$ the density and distribution function, respectively, of the standard Normal distribution. For more details we refer to Van der Laan (1991).
In practical situations it is often the case that X_i is the sample mean based on $n(\geq 2)$ observations from $N(\theta, \sigma^2)$-populations. The sample mean has standard deviation $\frac{\sigma}{\sqrt{n}}$, so in this case a difference in location parameter has to be compared with $\frac{\sigma}{\sqrt{n}}$. Thus $\varepsilon = 2$ means for $\sigma = 2.5$ and $n = 100$ an amount of $\frac{2.5}{\sqrt{100}} * 2 = 0.5$.

3. Tables

The tabulated values of RE were determined using numerical integration.
These values were calculated for $\varepsilon = 0.2, 0.5, 1.0, 1.5$ and 2.0; $P^* = 0.50, 0.80, 0.90, 0.95$ and 0.99 and $k = 2(1)10(5)30(10)100(50)300(100)500(250)2000$.
Values of RE for k not provided in the tables need to be interpolated.
Simple linear interpolation with respect to k seems a reasonable procedure for these cases. Accurate interpolation for P^* is more difficult. Such an interpolation will generally be unnecessary because P^* is chosen in advance, and the tables cover a wide range of practical values.

Tables

Values of RE for k populations

		P^*				
k	ε	0.50	0.80	0.90	0.95	0.99
2	0.2	-	1.003	1.003	1.002	1.001
	0.5	-	1.018	1.016	1.011	1.004
	1.0	-	1.072	1.063	1.046	1.017
	1.5	-	1.155	1.141	1.107	1.044
	2.0	-	1.258	1.247	1.195	1.088
3	0.2	1.002	1.003	1.002	1.002	1.000
	0.5	1.015	1.019	1.014	1.010	1.003
	1.0	1.063	1.082	1.061	1.042	1.015
	1.5	1.140	1.193	1.148	1.103	1.039
	2.0	1.235	1.357	1.282	1.202	1.083

		P^*				
k	ε	0.50	0.80	0.90	0.95	0.99
4	0.2	1.003	1.003	1.002	1.001	1.000
	0.5	1.019	1.018	1.013	1.008	1.003
	1.0	1.082	1.078	1.055	1.037	1.012
	1.5	1.196	1.192	1.138	1.093	1.034
	2.0	1.355	1.372	1.274	1.190	1.074
5	0.2	1.003	1.002	1.002	1.001	1.000
	0.5	1.019	1.016	1.011	1.007	1.002
	1.0	1.087	1.072	1.050	1.032	1.011
	1.5	1.217	1.182	1.127	1.084	1.030
	2.0	1.414	1.364	1.259	1.176	1.067
6	0.2	1.003	1.002	1.001	1.001	1.000
	0.5	1.019	1.015	1.010	1.006	1.002
	1.0	1.088	1.067	1.045	1.029	1.009
	1.5	1.223	1.171	1.117	1.077	1.027
	2.0	1.441	1.350	1.243	1.164	1.062
7	0.2	1.003	1.002	1.001	1.001	1.000
	0.5	1.018	1.013	1.009	1.006	1.002
	1.0	1.086	1.062	1.041	1.026	1.008
	1.5	1.223	1.161	1.109	1.071	1.024
	2.0	1.452	1.334	1.229	1.153	1.057
8	0.2	1.002	1.002	1.001	1.001	1.000
	0.5	1.017	1.012	1.008	1.005	1.002
	1.0	1.083	1.058	1.038	1.024	1.008
	1.5	1.219	1.152	1.102	1.066	1.023
	2.0	1.453	1.319	1.217	1.144	1.053
9	0.2	1.002	1.002	1.001	1.001	1.000
	0.5	1.017	1.012	1.008	1.005	1.001
	1.0	1.080	1.054	1.036	1.022	1.007
	1.5	1.214	1.144	1.096	1.062	1.021
	2.0	1.449	1.305	1.205	1.136	1.050
10	0.2	1.002	1.002	1.001	1.001	1.000
	0.5	1.016	1.011	1.007	1.004	1.001
	1.0	1.077	1.051	1.033	1.021	1.007
	1.5	1.209	1.137	1.090	1.058	1.020
	2.0	1.442	1.292	1.196	1.129	1.047
15	0.2	1.002	1.001	1.001	1.000	1.000
	0.5	1.013	1.008	1.005	1.003	1.001
	1.0	1.065	1.040	1.026	1.016	1.005
	1.5	1.181	1.110	1.072	1.046	1.015
	2.0	1.398	1.243	1.161	1.105	1.038
20	0.2	1.001	1.001	1.001	1.000	1.000
	0.5	1.011	1.007	1.004	1.003	1.001
	1.0	1.056	1.033	1.021	1.013	1.004
	1.5	1.159	1.094	1.060	1.038	1.013
	2.0	1.358	1.211	1.138	1.090	1.032

		P^*				
k	ε	0.50	0.80	0.90	0.95	0.99
25	0.2	1.001	1.001	1.001	1.000	1.000
	0.5	1.009	1.006	1.004	1.002	1.001
	1.0	1.049	1.029	1.018	1.011	1.003
	1.5	1.143	1.082	1.053	1.033	1.011
	2.0	1.326	1.189	1.123	1.080	1.028
30	0.2	1.001	1.001	1.000	1.000	1.000
	0.5	1.008	1.005	1.003	1.002	1.001
	1.0	1.044	1.025	1.016	1.010	1.003
	1.5	1.130	1.074	1.047	1.030	1.010
	2.0	1.300	1.171	1.111	1.072	1.025
40	0.2	1.001	1.001	1.000	1.000	1.000
	0.5	1.007	1.004	1.002	1.001	1.000
	1.0	1.037	1.021	1.013	1.008	1.002
	1.5	1.111	1.062	1.039	1.025	1.008
	2.0	1.261	1.147	1.095	1.061	1.021
50	0.2	1.001	1.000	1.000	1.000	1.000
	0.5	1.006	1.003	1.002	1.001	1.000
	1.0	1.032	1.018	1.011	1.007	1.002
	1.5	1.098	1.054	1.034	1.021	1.007
	2.0	1.234	1.130	1.084	1.054	1.019
60	0.2	1.001	1.000	1.000	1.000	1.000
	0.5	1.005	1.003	1.002	1.001	1.000
	1.0	1.028	1.016	1.010	1.006	1.002
	1.5	1.088	1.048	1.030	1.019	1.006
	2.0	1.213	1.118	1.076	1.049	1.017
70	0.2	1.001	1.000	1.000	1.000	1.000
	0.5	1.005	1.003	1.002	1.001	1.000
	1.0	1.026	1.014	1.009	1.005	1.002
	1.5	1.081	1.044	1.027	1.017	1.005
	2.0	1.197	1.108	1.070	1.044	1.015
80	0.2	1.001	1.000	1.000	1.000	1.000
	0.5	1.004	1.002	1.001	1.001	1.000
	1.0	1.023	1.013	1.008	1.005	1.001
	1.5	1.074	1.040	1.025	1.016	1.005
	2.0	1.183	1.101	1.065	1.041	1.014
90	0.2	1.000	1.000	1.000	1.000	1.000
	0.5	1.004	1.002	1.001	1.001	1.000
	1.0	1.022	1.012	1.007	1.004	1.001
	1.5	1.069	1.037	1.023	1.014	1.005
	2.0	1.172	1.094	1.060	1.038	1.013
100	0.2	1.000	1.000	1.000	1.000	1.000
	0.5	1.003	1.002	1.001	1.001	1.000
	1.0	1.020	1.011	1.007	1.004	1.001
	1.5	1.065	1.035	1.022	1.013	1.004
	2.0	1.163	1.089	1.057	1.036	1.012

		P^*				
k	ε	0.50	0.80	0.90	0.95	0.99
150	0.2	1.000	1.000	1.000	1.000	1.000
	0.5	1.003	1.001	1.001	1.001	1.000
	1.0	1.015	1.008	1.005	1.003	1.001
	1.5	1.051	1.027	1.017	1.010	1.003
	2.0	1.131	1.071	1.045	1.029	1.010
200	0.2	1.000	1.000	1.000	1.000	1.000
	0.5	1.002	1.001	1.001	1.000	1.000
	1.0	1.012	1.006	1.004	1.002	1.001
	1.5	1.042	1.022	1.014	1.008	1.003
	2.0	1.112	1.060	1.038	1.024	1.008
250	0.2	1.000	1.000	1.000	1.000	1.000
	0.5	1.002	1.001	1.001	1.000	1.000
	1.0	1.011	1.006	1.003	1.002	1.001
	1.5	1.037	1.019	1.012	1.007	1.002
	2.0	1.099	1.053	1.034	1.021	1.007
300	0.2	1.000	1.000	1.000	1.000	1.000
	0.5	1.002	1.001	1.001	1.000	1.000
	1.0	1.009	1.005	1.003	1.002	1.001
	1.5	1.033	1.017	1.011	1.006	1.002
	2.0	1.089	1.048	1.030	1.019	1.006
400	0.2	1.000	1.000	1.000	1.000	1.000
	0.5	1.001	1.001	1.000	1.000	1.000
	1.0	1.008	1.004	1.002	1.001	1.000
	1.5	1.027	1.014	1.009	1.005	1.002
	2.0	1.076	1.040	1.026	1.016	1.005
500	0.2	1.000	1.000	1.000	1.000	1.000
	0.5	1.001	1.001	1.000	1.000	1.000
	1.0	1.006	1.003	1.002	1.001	1.000
	1.5	1.024	1.012	1.008	1.005	1.001
	2.0	1.067	1.036	1.022	1.014	1.005
750	0.2	1.000	1.000	1.000	1.000	1.000
	0.5	1.001	1.001	1.000	1.000	1.000
	1.0	1.005	1.003	1.002	1.001	1.000
	1.5	1.018	1.009	1.006	1.004	1.001
	2.0	1.053	1.028	1.018	1.011	1.004
1000	0.2	1.000	1.000	1.000	1.000	1.000
	0.5	1.001	1.001	1.000	1.000	1.000
	1.0	1.004	1.002	1.001	1.001	1.000
	1.5	1.015	1.008	1.005	1.003	1.001
	2.0	1.045	1.024	1.015	1.009	1.003
1250	0.2	1.000	1.000	1.000	1.000	1.000
	0.5	1.001	1.001	1.000	1.000	1.000
	1.0	1.004	1.002	1.001	1.001	1.001
	1.5	1.013	1.007	1.004	1.003	1.001
	2.0	1.040	1.021	1.013	1.008	1.003

		P^*				
k	ε	0.50	0.80	0.90	0.95	0.99
1500	0.2	1.000	1.000	1.000	1.000	1.000
	0.5	1.001	1.001	1.001	1.000	1.000
	1.0	1.003	1.002	1.001	1.001	1.001
	1.5	1.012	1.006	1.004	1.002	1.001
	2.0	1.036	1.019	1.012	1.008	1.003
1750	0.2	1.001	1.001	1.001	1.001	1.001
	0.5	1.001	1.001	1.001	1.001	1.001
	1.0	1.003	1.002	1.001	1.001	1.001
	1.5	1.011	1.006	1.004	1.002	1.001
	2.0	1.033	1.017	1.011	1.007	1.003
2000	0.2	1.001	1.001	1.001	1.001	1.001
	0.5	1.001	1.001	1.001	1.001	1.001
	1.0	1.003	1.002	1.001	1.001	1.001
	1.5	1.010	1.005	1.003	1.002	1.001
	2.0	1.031	1.016	1.010	1.007	1.002

References

BUTLER, K.L. and BUTLER, D.G. (1987). *Tables for selecting the best population.* Queensland Biometrical Bulletin 2, Department of Primary Industries, Queensland Government, Brisbane.

GUPTA, S.S. and PANCHAPAKESAN, S. (1979). *Multiple decision procedures.* Wiley, New York.

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

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

Secretariate: Dommelbuilding 0.03
Telephone : 040-473130

-List of COSOR-memoranda - 1991

Number	Month	Author	Title
91-01	January	M.W.I. van Kraaij W.Z. Venema J. Wessels	The construction of a strategy for manpower planning problems.
91-02	January	M.W.I. van Kraaij W.Z. Venema J. Wessels	Support for problem formulation and evaluation in manpower planning problems.
91-03	January	M.W.P. Savelsbergh	The vehicle routing problem with time windows: minimizing route duration.
91-04	January	M.W.I. van Kraaij	Some considerations concerning the problem interpreter of the new manpower planning system formasy.
91-05	February	G.L. Nemhauser M.W.P. Savelsbergh	A cutting plane algorithm for the single machine scheduling problem with release times.
91-06	March	R.J.G. Wilms	Properties of Fourier-Stieltjes sequences of distribution with support in [0,1).
91-07	March	F. Coolen R. Dekker A. Smit	Analysis of a two-phase inspection model with competing risks.
91-08	April	P.J. Zwietering E.H.L. Aarts J. Wessels	The Design and Complexity of Exact Multi-Layered Perceptrons.
91-09	May	P.J. Zwietering E.H.L. Aarts J. Wessels	The Classification Capabilities of Exact Two-Layered Peceptrons.
91-10	May	P.J. Zwietering E.H.L. Aarts J. Wessels	Sorting With A Neural Net.
91-11	May	F. Coolen	On some misconceptions about subjective probability and Bayesian inference.

COSOR-MEMORANDA (2)

91-12	May	P. van der Laan	Two-stage selection procedures with attention to screening.
91-13	May	I.J.B.F. Adan G.J. van Houtum J. Wessels W.H.M. Zijm	A compensation procedure for multiprogramming queues.
91-14	June	J. Korst E. Aarts J.K. Lenstra J. Wessels	Periodic assignment and graph colouring.
91-15	July	P.J. Zwietering M.J.A.L. van Kraaij E.H.L. Aarts J. Wessels	Neural Networks and Production Planning.
91-16	July	P. Deheuvels J.H.J. Einmahl	Approximations and Two-Sample Tests Based on P - P and Q - Q Plots of the Kaplan-Meier Estimators of Lifetime Distributions.
91-17	August	M.W.P. Savelsbergh G.C. Sigismondi G.L. Nemhauser	Functional description of MINTO, a Mixed INTeger Optimizer.
91-18	August	M.W.P. Savelsbergh G.C. Sigismondi G.L. Nemhauser	MINTO, a Mixed INTeger Optimizer.
91-19	August	P. van der Laan	The efficiency of subset selection of an almost best treatment.
91-20	September	P. van der Laan	Subset selection for an ε-best population: efficiency results.