Subset selection : robustness and imprecise selection

EINDHOVEN UNIVERSITY OF TECHNOLOGY
Department of Mathematics and Computing Science

Memorandum COSOR 92-10

Subset Selection: Robustness and Imprecise Selection

P. van der Laan

Eindhoven, May 1992
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

Secretariate: Dommelbuilding 0.03
Telephone: 040-47 3130

ISSN 0926 4493

Subset Selection: Robustness and Imprecise Selection

Paul van der Laan

Eindhoven University of Technology

Department of Mathematics and Computing Science

Eindhoven, The Netherlands

"Although this may seem a paradox, all exact science is dominated by the idea of approximation."

Bertrand Russell

"If you think that Subset Selection is of little importance, think again!"

Summary

Assume k (integer $k \geq 2$) independent populations are given. The associated independent random variables have distributions with an unknown location parameter. The goal is to select the best population, this is the population with largest value of the location parameter. First, this paper reviews some distributional and robustness results for Subset Selection from Normal populations. Special attention is given to the probability of correct selection. Secondly, some distributional results are given. Explicit expressions for expectation and variance of the subset size using Subset Selection are presented. Finally, some remarks are made concerning a generalized selection goal using Subset Selection. Instead of selecting precisely the best population, an imprecise selection can be applied, that is selection of a population in the neighbourhood of the best population. The generalized Subset Selection goal is to select a non-empty subset of populations that contains at least one almost best population with a certain confidence level. For a collection of populations with an unknown location parameter an almost best population, or more accurately an ε-best population, is defined as a population with location parameter on a distance less than or equal to ε (with $\varepsilon \geq 0$) from the maximal value of the location parameter for all populations. The selection of an almost best population is compared with the selection of the best one from an application point of view. Some efficiency results are presented.

1. Introduction

In practice we are often confronted with the problem of selection. For, a general goal in many experiments is to select the best treatment out of several treatments. Each treatment is described by a qualitative factor characterized by a parameter value. Often this parameter will be a location parameter. Especially in the field of testing varieties statistical selection is an essential feature. For most kinds of selection problems a quantitative methodology of selection is needed. The usual statistical approach is to test the so-called homogeneity hypothesis. This homogeneity hypothesis states that all parameter values are equal to each other. In this context two remarks can be made. First, in general the treatments will differ so that a consistent test will reject the null hypothesis if the numbers of observations are large enough. Secondly, after rejecting the homogeneity hypothesis the statistical conclusion is that the treatments differ but not which treatment is the best. Multiple comparisons and simultaneous confidence intervals can give additional information.
Let us consider the problem of selecting the best variety from a number k (integer $k \geq 2$) of varieties. The best variety is defined as the variety with the largest value of the location parameter. If there are more than one contenders for the best because there are ties, it is assumed for continuity reasons and for computational convenience that one of these is appropriately tagged. We assume that the selection is based on the average yield per plot of constant size.
Furtheron, we assume that the experiments design is a complete randomized design with n plots for each variety or as a randomized complete block design with k plots per block and the plots randomly associated to the k varieties. Statistical selection procedures can possibly help us to improve our selection. A short description of the basic approaches will be given in section 2.
There are many goals to consider and we shall do that in section 3. The advantages and disadvantages of Subset Selection will also be mentioned in section 3, together with remarks on the meaning of certain characteristics of this statistical selection procedure. Some exact results concerning the expectation and variance of the subset size are presented in section 4. In section 5 some recent results from literature concerning robustness of selection procedures are presented.
In variety testing it is important that selection of the best variety can be guaranteed in some sense or another. Otherwise, we may turn around in a circle during a large period of several years. The principle of a statistical procedure is that the probability of an error, an incorrect selection, is under control. In a selection problem the probability of correct selection is from a practical point of view as important as the power of a test. To be sure, or almost sure, that we don't miss the best variety, the probability of correct selection of the best variety has to be taken into account. However, from a practical point of view, an experimenter may be satisfied with an almost best variety instead of the best variety. But also in that situation we want to have a confidence requirement that the best or an almost best variety will be selected. We return to this point in section 6. Finally, we conclude with some comments in section 7.

2. Basic approaches in statistical selection

The main approaches in handling with selection of the best population (variety or treatment) are the Subset Selection approach and the Indifference Zone approach (GUPTA and PANCHAPAKESAN (1979)). These two basic approaches suggested by GUPTA (1956, 1965)

and BECHHOFER (1954), respectivily, will be concisely considered.
Assume $k(k \geq 2)$ independent Normal random variables $X_1, ..., X_k$ are given. These random variables are associated with the k populations indicated by $\pi_1, ..., \pi_k$, and may be sample means. The assumed Normal distributions have common known variance σ^2 and unknown means $\mu_1, ..., \mu_k$. The collection of μ's is denoted by the vector $\mu = (\mu_1, ..., \mu_k)$.
The goal is to select the best population, that is the population with mean $\mu_{[k]}$, where $\mu_{[1]} \leq ... \leq \mu_{[k]}$ denote the ordered values of $\mu_1, ..., \mu_k$.
The Subset Selection procedure selects a subset, non-empty and as small as possible, with the probability requirement that the probability of a Correct Selection CS is at least P^* (with $1/k < P^* < 1$). A CS means in this context that the best population is an element of the selected subset. The selection procedure runs as follows. Select population π_i in the subset if and only if $X_i \geq \max_{1 \leq j \leq k} X_j - d \quad (i = 1, ..., k)$.
The selection constant d must be determined in such a way that $P(CS) \geq P^*$ for all possible parameter configurations. Tables with values of the selection constant d can be found in GIBBONS, OLKIN and SOBEL (1977) and in BUTLER and BUTLER (1987).
The second approach is the so-called Indifference Zone approach. Using the Indifference Zone approach the goal is to indicate the best population. The procedure is to select that population that resulted in the largest sample value. The probability requirement is that the probability of a CS is at least P^*, whenever the best population is at least $\delta^*(> 0)$ away from the second best. Using the Indifference Zone approach CS means that the best population produced the largest sample value and consequently it is also indicated as the best population. The minimal probability P^* can only be guaranteed if the common size n of the k samples, on which the response variables are based, is large enough. Tables for $\tau = \delta^* n^{1/2}/\sigma$ can be found in GIBBONS, OLKIN and SOBEL (1977) and BUTLER and BUTLER (1987). Notice that n can be computed as $(\tau\sigma/\delta^*)^2$. With the chosen minimal n it can be guaranteed with minimal probability P^* that the selected (indicated) variety is less than δ^* away from the best. From this it follows that the Indifference Zone approach is important for the design phase.
The Least Favourable Configuration LFC (all μ's are equal to each other) for Subset Selection is easier to handle with than the LFC for the Indifference Zone approach. It is also possible to consider two characteristics, one primary and one secondary in importance, and to choose from the subset the one which is satisfactory in terms of the secondary characteristic.
In practice the Subset Selection approach can be used as a screening procedure. Even when the ultimate goal of the experimenter is to choose the best, the approach can be applied to eliminate inferior varieties.
Some introductory remarks on selection procedures and some literature can be found in van der LAAN (1987). In this paper we shall concentrate on the Subset Selection approach.

3. Goals

In this section we shall in illustration describe a number of goals which may be relevant from a practical point of view in the field of variety testing. We mention the following goals:

i. Selection of the best variety.

ii. Selection of the t best varieties (with $2 \leq t < k$). We can do this with or without ordering of the varieties. In the first case we indicate a variety as the best one, another variety

as the second best, etc. In the second case we produce a collection of t varieties without ranking them.

iii. Selection of a subset of varieties that will contain at least the best variety.

iv. Selection of a subset of varieties that will contain at least the t (with $2 \leq t < k$) best varieties.

v. Selection of a subset that contains at least one almost best variety.

In the literature different generalizations and modifications have been proposed. We refer to GUPTA and PANCHAPAKESAN (1979) for references.
An interesting result has been achieved by HSU (1981, 1984). Selecting a subset with confidence P^* for containing the best, he also gives simultaneously confidence intervals for the differences of the varieties compared with the best one. He integrated both approaches, the Indifference Zone approach and the Subset Selection approach. His method is known as 'Multiple Comparisons with the Best'. His method restricts the comparisons to the following differences $\mu_{[k]} - \mu_1, ..., \mu_{[k]} - \mu_k$. Let X_i be the sample mean of n observations $(i = 1, ..., k)$ and suppose that σ^2 is unknown. Let s^2 be the best estimator of σ^2 based on d degrees of freedom. Define

$$C := \{V_i : \min_{l \neq i} (X_i - X_l + hs(2/n)^{1/2}) \geq 0\}$$

and

$$D_i := \max[0, \max_{l \neq i} (X_l - X_i + hs(2/n)^{1/2})] .$$

The constant h satisfies

$$\int_{w=0}^{\infty} \int_{z=-\infty}^{\infty} \Phi^{k-1}(z + 2^{1/2}hw) d\Phi(z) dG(w) = P^* ,$$

where $G(w)$ is the distribution function of $w = s/\sigma$, $\Phi(z)$ is the standard Normal cumulative distribution function (cdf), and P^* is the desired confidence. Values of the constant h are given in table A4 of GIBBONS, OLKIN and SOBEL (1977). HSU (1981) proved that for all configurations of the μ's the following holds

$$P[V_{(k)} \in C \text{ and } \mu_{[k]} - \mu_i \leq D_i \text{ for all } i] \geq P^* .$$

Simultaneously with the selection also a confidence statement can been made, and both conclusions with simultaneous confidence level P^*. In van der LAAN and VERDOOREN (1990) an application in the field of variety testing has been given.

4. The Subset Size

Subset selection is a flexible form of selection, because the number of replications has not to be determined in advance.
Also after the experiment has been carried out, the selection can be executed. The influence of the number of replications can be conducted from the (expected) size of the subset. Hence, the expected size of the subset can be considered as a crucial quantity. A relatively large subset means, apart from random fluctuations, that the number of replications is small or the expected yields of the varieties are close together, or both.
Indicating the subset size with S the following theorems for populations with general cdf F can be proved.

Theorem 4.1. The distribution of S is given by

$$P(S=s) = \sum_{i_1<\ldots<i_s} \sum_{r=1}^{s} \int_{-\infty}^{\infty} \prod_{j=s+1}^{k} F(x-\mu_{[i_j]}+\mu_{[i_r]}-d)$$

$$\prod_{\substack{j=1 \\ j\neq r}} \{F(x-\mu_{[i_j]}+\mu_{[i_r]})-$$

$$-F(x-\mu_{[i_j]}+\mu_{[i_r]}-d)\}dF(x)\,,$$

for $s=1,\ldots,k$.

Corollary 4.1. The distribution of S under the Least Favourable Configuration LFC is

$$P(S=s|LFC) = \binom{k}{s}\binom{s}{1}\int_{-\infty}^{\infty}\{F(x)-F(x-d)\}^{s-1}F^{k-s}(x-d)\,dF(x)\,.$$

Theorem 4.2. For the expected subset size $E(S|LFC)$ one finds

$$E(S|LFC) = k\int_{-\infty}^{\infty} F^{k-1}(x+d)\,dF(x)$$

and for the variance of S under the LFC

$$\mathrm{var}\{S|LFC\} = k(2k-1)\int_{-\infty}^{\infty} F^{k-1}(x+d)\,dF(x)-$$

$$-2k(k-1)\int_{-\infty}^{\infty} F^{k-2}(x+d)F(x)\,dF(x)-$$

$$-k^2[\int_{-\infty}^{\infty} F^{k-1}(x+d)\, dF(x)]^2 .$$

In the tables 4.1 and 4.2 some results are presented for standard Normal distributions.

Table 4.1. $E\{S|LFC\}$ for standard Normal distributions for some values of k and d.

k \ d	1	1.5	2	2.5	3	3.5	4	4.5	5
2	1.52	1.71	1.84	1.92	1.97	1.99	2.00	2.00	2.00
5	2.47	3.26	3.94	4.43	4.73	4.89	4.96	4.99	5.00
10	3.41	5.09	6.74	8.09	9.02	9.57	9.83	9.94	9.98
25	4.99	8.68	13.05	17.28	20.66	22.89	24.12	24.68	24.90

Table 4.2. Standard deviation of S under the LFC for standard Normal distributions for some values of k and d.

k \ d	1	1.5	2	2.5	3	3.5	4	4.5	5
2	.500	.453	.364	.267	.181	.115	.068	.038	.020
5	1.14	1.20	1.08	.848	.595	.381	.225	.125	.066
10	1.78	2.14	2.12	1.80	1.32	.863	.511	.281	.145
25	2.91	4.13	4.73	4.50	3.62	2.51	1.52	.083	.042

Theorem 4.3. If

$$C_m(a) := (a/(a-1))^{m+1}\{\ln a - \sum_{i=1}^{m} (1-1/a)^i/i\}$$

and

$$\sum_{i=1}^{m}(1-1/a)^i/i = 0 \quad \text{for} \quad m \leq 0 ,$$

then for the Logistic distribution the following holds

$$E(S|LFC) = k[1-(k-1)/aC_{k-1}(a)]$$

and

$$\text{var}(S|LFC) = k(k-1)\{k/aC_{k-1}(a) + (k-2)/a\}\{1 - (k-1)/aC_{k-1}(a)\} ,$$

where $a = \exp(\beta d)$ (> 1) with β the scale parameter of the Logistic distribution.

5. Robustness of Subset Selection

First it is possible to examine the robustness of the lower bound for the $P(CS)$ against departures from the assumption of a common known standard deviation σ_0. This, in fact, was done by DRIESSEN, van der LAAN and van PUTTEN (1990), who considered Normal populations and several values of P^*, namely .75, .90, .95 and .99, respectively. This investigation has been executed by varying the standard deviations in the interval $I = [\alpha^{-1}\sigma_0, \alpha\sigma_0]$, with $\alpha \geq 1$, and determining the minimum value of the corresponding lower bound of $P(CS)$. The vector of standard deviations $(\sigma_1, ..., \sigma_k)$, where σ_i is the standard deviation of population $\pi_i (i = 1, ..., k)$, is denoted by σ.
The following selection rule has been used:

$$\pi_i (i = 1, ..., k) \text{ in subset iff } X_i \geq \max_{1 \leq j \leq k} X_j - d\sigma_0 n^{-1/2} ,$$

where the selection constant d has to be chosen such that the P^*-requirement has been met. For the LFC the following holds

$$P_{LFC}[CS|\sigma] = \int \prod_{i=1}^{k-1} (\sigma_k \sigma_i^{-1} y + d\sigma_0 \sigma_i^{-1}) \varphi(y)\, dy .$$

DRIESSEN et al. (1990) define the loss L as a function of k, P^*, α and σ_0 by the difference between P^* and the minimum of the $P(CS)$, where the elements of the vector σ are varying in I. It follows that

$$L = P^* - \min_{\alpha^{-2} \leq s \leq 1} \int \Phi^{k-1}(\alpha^{-1} d^* + sy) \varphi(y)\, dy ,$$

with $\varphi(\cdot)$ the standard Normal density. From this it follows that L is independent of σ_0. As a matter of fact a discretization is necessary. It appears that there is a serious lack of robustness in the sense that the actual lower bound of the $P(CS)$ can be considerably smaller than the pretended lower bound P^* of the probability of correct selection based on the assumption of a common known variance for the chosen values of P^*. This lack of robustness is substantial for large values of k and for large values of α. But also for small values of k and α the experimenter must not neglect in general the loss due to the departure from the assumption of equal variances. For $P^* = 0.90$, $k = 100$ and $\alpha = 1.5$ the loss is 0.362. For the same value of k and $\alpha = 2.0$ the loss is very large, namely 0.776. But also for small values of k the loss is essential, for instance for $k = 5$ the loss for $\alpha = 1.5$ is already 0.180.
For $P^* = 0.95$ and $k = 10$ the loss runs from 0.0 for $\alpha = 1$, via (all the time approximately) 0.38 for $\alpha = 2$ and 0.66 for $\alpha = 3$, to 0.8 for $\alpha = 4$. For $k = 3$ the numbers are 0.0, 0.22, 0.35 and 0.42, respectivily. For $k = 100$ these numbers are 0.0, 0.7, 1.0 and 1.0, respectivily.
For $P^* = 0.90$ and $\alpha = 2$ table 5.1 gives an illustration of the loss for different values of k.

Table 5.1. The loss L for different values of k ($P^* = 0.90$ and $\alpha = 2$).

k	2	3	5	10	25	100
L	0.16	0.22	0.31	0.40	0.56	0.78

Secondly, it is also possible to investigate the robustness of the Subset Selection procedure against deviations from the assumption of Normality. One can study the concequences of deviations in that sense that instead of Normality the observations are Logistically distributed, all the time standardized distributions. The Logistic distribution is symmetric and has a shape similar to that of the Normal distribution. The tails of the Logistic density are heavier than those of the Normal density.
In order to study robustness of the $P(CS)$ the value of P^* is compared with the real lower bound, thus under the LFC, based under the assumption of Logistically distributed observations with known scale parameter β, thus with density

$$f(x) = \beta \ \exp(-\beta(x - \mu_i))\{1 + \exp(-\beta(x - \mu_i))\}^{-2} \ .$$

For standard Logistic observations the $P(CS|LFC)$ can be determined using results in van der LAAN (1989). One finds

$$P(CS|LFC) = 1 - (k-1)a^{-1}C_{k-1}(a) \ ,$$

where $a = \exp(\pi 3^{-\frac{1}{2}} d)$ with d the selection constant of the selection rule defined in section 2. Exact values of the actual minimal probability for some values of P^*-Normal, when in reality the observations are standard Logistically distributed are given in table 5.2.

Table 5.2. Minimal probability of CS when in reality the Logistic distribution is valid, for some values of P^*.

P^*-Normal = 0.90

k	2	3	4	5	6	7	10
min. P	.906	.905	.902	.900	.898	.896	.890

P^*-Normal = 0.95

k	2	3	4	5	6	7	10
min P	.951	.949	.946	.944	.943	.941	.937

For the situation that the random variables are the means of $n > 1$ independent and standard Logistically distributed observations simulations have been carried out (van der LAAN and van PUTTEN (1990)). Having the Central Limit Theorem in mind it is not surprising that

for large values of n the results are generally in still better agreement with Normal results. Some results are presented in table 5.3.

Table 5.3. Some simulation results for the Logistic distribution.

P^*-Normal = 0.90 and $n = 10$

k	2	3	4	5	6	7	10
min P	.899	.901	.893	.892	.901	.899	.889

P^*-Normal = 0.95 and $n = 10$

k	2	3	4	5	6	7	10
min P	.952	.949	.946	.952	.949	.950	.943

Finally, it is possible to make a comparison between the selection constants for standard Normal populations and those for standard Logistic populations. In table 5.4 some results are presented.

Table 5.4. A comparison between Normal and Logistic results for $P^* = 0.90$ and some values of k.

k	Normal	Logistic	Rel. error
2	1.8124	1.7581	-0.030
3	2.2302	2.1906	-0.018
4	2.4516	2.4319	-0.008
5	2.5997	2.5995	-0.000
6	2.7100	2.7280	0.007
7	2.7972	2.8322	0.013
8	2.8691	2.9198	0.018
9	2.9301	2.9954	0.022
10	2.9829	3.0619	0.026
25	3.3911	3.6104	0.065
50	3.6584	4.0063	0.095

The differences between the Normal and Logistical results are for $k \leq 10$ rather small. Only for the cases $k = 25$ and $k = 50$ the difference is not so small. The difference becomes larger with increasing k.

6. Imprecise Subset Selection

The requirement to select precisely the best variety may be a strong one if the best variety is not far away from the other varieties. In practice, objections against selection procedures are more or less concentrated on large subsets using the subset selection approach. In the situation sketched before it is clear that with a large P^* one has to pay automatically with large expected subsets. Another possibility is to increase the common number of observations on which X is based. However, using the subset selection approach the common sample size is generally assumed fixed in practice. A possible way out is a so-called impricise subset selection.
Let us assume that the best variety and the second best are near each other. More accurately: the average yields are close together, say on a distance less than ε, where $\varepsilon > 0$, but relatively small. In such a situation it is often not of pactical interest whether one selects the best one or the next best. Not every difference in yield is important. In other words we are content with a more or less imprecise selection. In many real world problems it is of interest to select the best or an almost best variety. This approach leads to the consideration that one may generalize the selection goal to a selection of an almost best variety. A consequence of this generalization is that the least favourable configuration becomes more difficult. This is not an essential disadvantage. The goal is to select a small but nonempty subset such that the selected subset will contain the best variety or an almost best one with confidence P^*. In general, the generalization of selecting the best to selecting an almost best variety will result in subsets of smaller expected size.
Some aspects of selecting an almost best population have been considered in van der LAAN (1991a, 1991b, 1992a).

Definition 6.1. A treatment T_i is called an ε-best treatment if and only if $\mu_i \geq \mu_{[k]} - \varepsilon$, with $\varepsilon \geq 0$.

From Def. 6.1 it follows that the best treatment is also ε-best. Thus for each $\varepsilon \geq 0$ there always exists at least one ε-best treatment. If $\varepsilon = 0$, then there exists only one ε-best treatment, namely the best treatment, assuming there are no ties.

Definition 6.2. A correct selection CS is a selection of a subset C which contains at least one ε-best treatment.

We use the following selection rule R for determining a subset C. A treatment $T_i (i = 1, ..., k)$ is in the subset C if and only if $X_i \geq \max_{1 \leq j \leq k} X_j - c$, were the selection constant $c \geq 0$ has to be determined such that $P(CS) \geq P^*$. The best treatment is denoted by $T_{(k)}$.
It can be proved that

$$\inf \mu P(CS) = P(T_{(k)} \in C | \mu_{[1]} = \mu_{[k-1]} = \mu_{[k]} - \varepsilon) .$$

From this it follows that for Normal distributions with common scale parameter σ the selection constant c is equal to

$$c = d\sigma - \varepsilon ,$$

where d can be found in BUTLER and BUTLER (1987). Without loss of generality we

assume that $\sigma = 1$. In table 6.1 a comparison is made between the P^* value for subset selection of an ε-best treatment and the probability of correct selection of the best treatment for different values of k, ε and P^*. The computations are based on interpolations so are of limited accuracy. The efficiency of both procedures is measured by the relative gain G_r in minimal probability of correct selection. G_r is defined as the gain in minimal probability of correct selection for an ε-best treatment relative to that of a best treatment.

Table 6.1. The probability of correct selection of the best treatment using the selection constant for selection of an ε-best treatment, and the relative gain G_r (with $P^* = 0.90$).

$\varepsilon = 0.2$

k	Min. Pr. for best	G_r in %
2	0.868	3.7
3	0.865	4.0
4	0.864	4.2
5	0.863	4.3
6	0.863	4.3
7	0.862	4.4
8	0.862	4.4
9	0.862	4.4
10	0.862	4.4
25	0.860	4.7
50	0.859	4.8
100	0.859	4.8
500	0.858	4.9
1000	0.858	4.9
2000	0.857	5.0

$\varepsilon = 0.5$

k	min. Pr. for best	G_r in %
2	0.820	9.8
3	0.813	10.7
4	0.810	11.1
5	0.809	11.2
6	0.807	11.5
7	0.806	11.7
8	0.805	11.8
9	0.805	11.8
10	0.804	11.9
25	0.801	12.4
50	0.799	12.6
100	0.797	12.9
500	0.795	13.2
1000	0.794	13.4
2000	0.793	13.5

For $\varepsilon = 1$ one finds for $k = 10$ a minimal probability of 0.655 so $G_r = 37\%$.
Some results for G_r are summarized in table 6.2 for different values of P^*.

Table 6.2. The relative gain G_r for different values of P^*.

$k = 10, \varepsilon = 0.2$

P^*	0.80	0.90	0.95	0.975	0.99	0.995	0.999
min P	0.739	0.862	0.927	0.962	0.983	0.9917	0.9977
G_r in %	8.3	4.4	2.5	1.4	0.7	0.33	0.13

$k = 10, \varepsilon = 0.5$

P^*	0.80	0.90	0.95	0.975	0.99	0.995	0.999
min P	0.648	0.804	0.893	0.942	0.973	0.9868	0.9959
G_r in %	23.5	11.9	6.4	3.5	1.7	0.83	0.31

The relative gain G_r becomes smaller with increasing P^* for the cases considered. For fixed $\varepsilon (\varepsilon \geq 0)$ we can write $G_r = G_r(c)$ as

$$G_{r(c)} = \frac{\int_{-\infty}^{+\infty} F^{k-1}(x+c+\varepsilon)\, f(x)\, dx}{\int_{-\infty}^{+\infty} F^{k-1}(x+c)\, f(x)\, dx} - 1 ,$$

where c is the selection constant for subset selection of an ε-best treatment meeting the P^*-requirement. One has $c = c(P^*)$, an increasing function of P^*, and

$$\lim_{c \uparrow \infty} G_r(c) = 0 .$$

Definition 6.3. A density is strongly unimodal if and only if $f(x)$ is log-concave (that is log $f(x)$ is a concave function).
(DHARMADHKARI and JOAG-DEV (1988)).
The following theorem holds.

Theorem 6.1. If $f(x)$ is strongly unimodal, then $G_r(c)$ is a decreasing function of P^*.

Proof:
We notice that the quantity

$$G(c) := \int_{-\infty}^{+\infty} F^{k-1}(x+c)\, f(x)\, dx$$

can be written as

$$G(c) = P[\max(X_2, X_3, ..., X_k) - X_1 \leq c]$$
$$= F^{k-1} * F(c) ,$$

where $*$ means the convolution. F and hence (cf. van der LAAN (1970)) F^{k-1} are strongly unimodal. Since strong unimodality is preserved under convolution (DHARMADHKARI and JOAG-DEV (1988)), $G(c)$ has a log-concave density. It follows (KARLIN 1968)) that G itself is log-concave. From this it follows that $\log G(c+\varepsilon) - G(c)$ is a decreasing function of c, so $G_r(c)$ is a decreasing function of c.

7. Concluding remarks

During decades of years we are used to apply statistical tests, like analysis of varince tests, to problems that are real selection problems. Especially in the field of variety testing many problems are in fact selection problems. We think it is important to investigate the possibilities to use statistical selection procedures for certain problems in variety testing. The statement of JOHN TUKEY 'An approximate answer to the right problem is worth a good deal more

than an exact answer to an approximate problem' illustrates that an exact formulation as a selection problem is worthwile. Not for all designs of experiments this problem has been solved.
In this paper we have tried to discuss some aspects of selection which are of interest from a practical point of view.
From the presented results it can be concluded that subset selection for Normal populations is not very robust against deviations of a common variance. It seems that small deviations from the Normal shape are not of great influence on the selection characteristics. The results from Logistic populations can be considered as an indication. Selection of an almost best population is worthwile to be considered in practice. In any case, subset selection remains an important tool in the statistician's tool bag.

Acknowledgement

I would like to thank Professor Fred Steutel for his helpful remarks with respect to Theorem 6.1.

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.

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.

Dharmadhikari, S. and K. Joag Dev (1988). Unimodality, Convexity, and Applications. Academic Press, New York - London.

Driessen, S., P. van der Laan and B. van Putten (1990). Robustness of the probability of correct selection against deviations from the assumption of a common known variance. Biometrical Journal **32**, 131-142.

Dudewicz, E.J. (1980). Ranking (ordering) and selection: An overview of how to select the best. Technometrics **22**, 113-119.

Gibbons, J.D., I. Olkin, M. Sobel (1977). Selecting and Ordering Populations: A New Statistical Methodology. Wiley, New York.

Gibbons, J.D., I. Olkin and M.Sobel (1979). An introduction to ranking and selection. The American Statistician **33**, 185-195.

Gupta, S.S. (1956). On a decision rule for a problem in ranking means. Ph.D. thesis, Department of Statistics, University of North Carolina, Chapel Hill, N.C.

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

Gupta, S.S. (1977). Selection and ranking procedures: a brief introduction. Commun. Statist. Theor. Meth. **A6**, 993-1001.

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

Hsu, J.C. (1981). Simultaneous confidence intervals for all distances from the 'best'. Annals of Statistics **9**, 1026-1034.

Hsu, J.C. (1984). Constrained simultaneous intervals for multiple comparisons with the best. Annals of Statistics **12**, 1136-1144.

Karlin, S. (1968). Total Positivity. Stanford University Press, Vol. **1**.

Keilson, J. (1979). Markov Chain Models - Rarity and Exponentiality. Applied Math. Sciences **28**, Springer-Verlag, New York - Berlin.

Laan, P. van der (1970). Simple distribution-free confidence intervals for a difference in location. Eindhoven University of Technology.

Laan, P. van der (1987). Some remarks on ranking and selection of treatments. Cultivar Testing Bulletin **12**, 203-218.

Laan, P. van der (1989). Selection from Logistic populations. Statistica Neerlandica **43**, 169-174.

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

Laan, P. van der (1991b). Subset selection for an ε-best population: efficiency results. COSOR-Memorandum **91-20**, Eindhoven University of Technology, Department of Mathematics and Computing Science.

Laan, P. van der (1992a). Subset Selection of an almost best treatment. Biometrical Journal **34**, no. 5 or 6.

Laan, P. van der (1992b). On subset selection from logistic populations. Statistica Neerlandica **46**, no. 2.

Laan, P. van der, and B. van Putten (1990). Robustness of the Normal means selection procedure with common known variance against Logistic deviations and the use of the Logistic approximation for the Normal distribution. Pub. Inst. Stat. Univ. Paris **35**, 79-92.

Laan, P van der, and L.R. Verdooren (1990). A review with some applications of statistical selection procedures for selecting the best variety. Euphitica **51**, 67-75.

List of COSOR-memoranda - 1992

Number	Month	Author	Title
92-01	January	F.W. Steutel	On the addition of log-convex functions and sequences
92-02	January	P. v.d. Laan	Selection constants for Uniform populations
92-03	February	E.E.M. v. Berkum H.N. Linssen D.A. Overdijk	Data reduction in statistical inference
92-04	February	H.J.C. Huijberts H. Nijmeijer	Strong dynamic input-output decoupling: from linearity to nonlinearity
92-05	March	S.J.L. v. Eijndhoven J.M. Soethoudt	Introduction to a behavioral approach of continuous-time systems
92-06	April	P.J. Zwietering E.H.L. Aarts J. Wessels	The minimal number of layers of a perceptron that sorts
92-07	April	F.P.A. Coolen	Maximum Imprecision Related to Intervals of Measures and Bayesian Inference with Conjugate Imprecise Prior Densities
92-08	May	I.J.B.F. Adan J. Wessels W.H.M. Zijm	A Note on "The effect of varying routing probability in two parallel queues with dynamic routing under a threshold-type scheduling"
92-09	May	I.J.B.F. Adan G.J.J.A.N. v. Houtum J. v.d. Wal	Upper and lower bounds for the waiting time in the symmetric shortest queue system
92-10	May	P. v.d. Laan	Subset Selection: Robustness and Imprecise Selection