

Two-stage selection procedures with attention to screening

Citation for published version (APA):

Laan, van der, P. (1991). *Two-stage selection procedures with attention to screening*. (Memorandum COSOR; Vol. 9112). Technische Universiteit Eindhoven.

Document status and date:

Published: 01/01/1991

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

EINDHOVEN UNIVERSITY OF TECHNOLOGY
Department of Mathematics and Computing Science

COSOR-Memorandum 91-12

Two-stage selection procedures with attention
to screening

by

P. van der Laan

ISSN: 0926-4493

May, 1991

Two-stage selection procedures with attention to screening

Paul van der Laan

Summary

Some literature concerning two-stage selection procedures is given. In general, the problem of selecting the Normal population with largest mean from $k(\geq 2)$ Normal populations with a common variance is considered. Special attention is drawn to two-stage procedures with screening for selecting the best. Such a procedure consists of a combination of the subset selection approach and the indifference zone approach. The first stage is used for eliminating bad populations and the second stage is used to indicate the best population from the remaining part.

1. Introduction

In a number of situations an investigator is faced with the problem of designing an experiment in order to select the best population out of k possible competing populations, where k is a fixed known integer ($k \geq 2$). We assume that the populations are described by qualitative variables and are indexed by a real-valued **location** parameter μ . We shall consider the problem how to select the $t(1 \leq t < k)$ best populations from a set of k populations. If there are more than t populations as candidates, then it is assumed that t of them are tagged as the t best populations. This assumption is made for the purpose of evaluating the infimum of the probability of correct selection. A statistical methodology using selection procedures should have many useful applications in the field of engineering, for instance quality processing.

A natural procedure would be to take n independent observations from each population and to obtain average values. Then the t populations with the largest sample means are indicated as the t best populations. However, the question remains of how large n should be in order to be sure with a certain confidence that the selected populations are indeed the t best populations. For the case of k Normal populations with a common known standard deviation the problem of selecting the best population ($t = 1$) and the determination of the common sample size was solved by Bechhofer (1954) using a fixed sample and the so-called Indifference Zone approach. The probability of correct selection has to be larger than or equal to P^* for all $\mu = (\mu_1, \dots, \mu_k)$ with the difference between the largest μ and the second largest is larger than or equal to a given positive real value δ^* (the so-called P^* -condition). Hall (1959) showed that among single-stage procedures this procedure is "most economical". Eaton (1967) proved that it has additional desirable decision properties. The case of k Normal populations with an unknown common standard deviation has been considered by Bechhofer, Dunnett and Sobel (1954). A more general formulation of the general selection problem can be found in Rasch (1984). Reviews of the literature, also with reference to the Normal means problem, appear in Wetherill and Ofosu (1974), Bechhofer (1975), Bechhofer (1985) and van der Laan and Verdooren (1989).

A different approach is that of Gupta (1965), which is indicated by "Subset Selection". The Subset Selection approach has as its goal to select a subset of the k populations considered,

in order to include the t best populations with a certain confidence. More precisely, the probability that the t best populations are in the subset is at least equal to P^* . This last requirement is indicated as the P^* -condition. The size of the subset is a random variable and depends among other things on the k sample sizes. In the book by Gupta and Panchapakesan (1979) one can find a large number of references, but also in Gupta and Panchapakesan (1985).

Bechhofer's procedure requires a common sample size n per population, which is chosen in such a way that the P^* -condition is fulfilled. This procedure is "conservative" in the sense that if μ is not an element of the so-called Least Favourable Configuration ($\mu_{[1]} = \dots = \mu_{[k-1]} = \mu_{[k]} - \delta^*$) which is mostly the case, then $P(CS) > P^*$ for the actual μ which the experimenter has encountered: in a certain sense there is an overprotection.

To overcome certain difficulties sequential procedures have been proposed. To mention a few: Bechhofer and Sobel (1954), Bechhofer, Kiefer and Sobel (1968), Paulson (1964), Fabian (1974a, 1974b), Lawing and David (1966) and Ramberg (1966). In some experimental situations, e.g. in agriculture where only one stage per growing season can be obtained, a sequential procedure may be impractical.

In this paper we will consider two-stage procedures. When the variance is unknown the first stage can be used to estimate the variance. A different idea behind a two-stage procedure is that after the first stage the experimentations can be terminated if for instance $\mu_{[k-t+1]} \gg \mu_{[k-t]}$. If not, the first stage is used to eliminate inferior populations.

2. Indifference Zone approach

The Indifference Zone approach for selecting was introduced by Bechhofer (1954).

References can be found in Gupta and Panchapakesan (1979) and Bechhofer (1985). Assume k (with $k \geq 2$) independent populations denoted by $G = (\pi_1, \dots, \pi_k)$ are given. The related random variable to π_i is X_i with cumulative distribution function $F(x; \mu_i)$ ($i = 1, \dots, k$). Let Ω denote the parameter space $\{\mu : \mu = (\mu_1, \dots, \mu_k)\}$. The ranked parameter values are indicated by $\mu_{[1]} \leq \dots \leq \mu_{[k]}$. The t ($1 \leq t < k$) best populations are $\pi_{(k-t+1)}, \dots, \pi_{(k)}$, where $\pi_{(i)}$ is the population associated with $\mu_{[i]}$ ($i = 1, \dots, k$). The goal is to select the unordered set $\{\pi_{(k-t+1)}, \dots, \pi_{(k)}\}$.

Let δ_{ij} denote a measure of distance between the populations $\pi(i)$ and $\pi(j)$ with $1 \leq i < j \leq k$. In our case of location parameters, δ_{ij} is usually defined as $\mu_{[j]} - \mu_{[i]}$. We denote the probability of Correct Selection (CS) using a selection rule R , usually bases on sufficient statistics for μ , by $P(CS | R)$ or $P(CS)$. We define two subspaces of Ω , namely $\Omega(\delta^*) = \{\mu : \delta_{k-t+1, k-t} \geq \delta^* > 0\}$, the so-called preference zone, and the subset $\Omega^c(\delta^*)$, the indifference zone.

The general problem is to determine the smallest common sample size n for which

$$P(CS) \geq P^* \text{ for all } \mu \in \Omega(\delta^*) ,$$

which is called the P^* -condition or P^* -requirement. The experimenter has to specify $0 < \delta^*$

$$\text{and } \binom{k}{t}^{-1} < P^* < 1.$$

3. Subset selection approach

Gupta (1965) suggested the subset selection approach. His goal is to select a subset of the k populations to include the best population with probability at least equal to $P^*(k^{-1} < P^* < 1)$. The size of the subset is a random variable and depends among other things on the k sample sizes. We want a selection rule which makes the size of the subset as small as possible. The subset selection approach can be applied after the experiment has already been executed.

In general, the goal is to construct a selection rule R which partitions the set G into two subsets G_3 and G_4 , where

$$\begin{aligned} G &= G_3 \cup G_4 \\ G_3 \cap G_4 &= \emptyset \end{aligned}$$

and

$$P(CS) = P(\pi_{(k-t+j)} \in G_3; j = 1, \dots, t) \geq P^* \text{ for } \mu \in \Omega, (1 \leq t < k).$$

The size S of the selected subset G_3 is a random variable taking on the values $1, \dots, k$ and $P(S \geq t) = 1$.

For general t a selection rule has been presented by Carroll, Gupta and Huang (1975).

For $t = 1$ a class of selection rules has been studied by Gupta and Panchapakesan (1972). In the location parameter case $F(x; \mu) = F(x - \mu)$ the rule

$$\begin{aligned} R: \pi_i \in G_3 \text{ iff} \\ \bar{X}_i \geq \bar{X}_{[k]} - d \quad (i = 1, \dots, k) \end{aligned}$$

with $d \geq 0$, has been discussed by Gupta (1965).

The selection constant d must be solved from

$$\int_{-\infty}^{\infty} F^{k-1}(x+d) dF(x) = P^*.$$

4. The indifference zone approach:

Normal populations with common known variances

Let us consider the situation of k Normal populations with expectation μ_i ($i = 1, \dots, k$) and with common known variance σ^2 . The goal is to partition the set G into two subsets G_1 and G_2 , with

$$G_1 = \{ \pi_{(k-t+1)}, \dots, \pi_{(k)} \},$$

and

$$G_2 = \{ \pi_{(1)}, \dots, \pi_{(k-t)} \}.$$

If there are more than t contenders because there are ties, it is assumed that t of these are appropriately tagged. The selection rule is based on the sufficient statistics for μ_i , namely the sample means of samples of n independent observations from π_i . We select a subset S of G of size t as follows. Include in S the populations associated with the t largest sample means. The P^* -condition can be written as

$$P[S = G_1 | \mu \in \Omega(\delta^*)] = P[S = G_1 | \delta_{k-t+1, k-t} \geq \delta^*] \geq P^* .$$

Then

$$P(CS) \geq t \int \Phi^{k-t}(x + \tau) \{1 - \Phi(x)\}^{t-1} d\Phi(x) ,$$

where Φ is the standard Normal cumulative distribution function and

$$\tau = \frac{\sqrt{n}\delta^*}{\sigma} .$$

The minimum of $P(CS)$ for a given selection rule R is attained for the Least Favourable Configuration (LFC) given by $\mu_{[1]} = \dots = \mu_{[k-t]} = \mu_{[k-t+1]} - \delta^* = \dots = \mu_{[k]} - \delta^*$. The minimum sample size required is the smallest integer n for which the P^* -condition is fulfilled. For the LFC we have

$$P_{LFC}(CS) = t \int \Phi^{k-t}(x + \tau) \{1 - \Phi(x)\}^{t-1} d\Phi(x) .$$

The smallest value τ meeting the P^* -condition can be found by solving $P_{LFC}(CS) = P^*$.

For $t = 1$ we have $P(CS) = \int_{-\infty}^{\infty} \prod_{i=1}^{k-1} \Phi \left(x + \frac{\sqrt{n}}{\sigma} (\mu_{[k]} - \mu_{[i]}) \right) d\Phi(x)$ and one has to solve

$$\int_{-\infty}^{\infty} \Phi^{k-1}(x + \tau) d\Phi(x) = P^* ,$$

resulting into

$$n = \left(\frac{\tau\sigma}{\delta^*} \right)^2 .$$

If the computed value of n is not an integer, it can be rounded upward. Tables of τ can be found in for instance Gibbons, Olkin and Sobel (1977) and Gupta, Nagel and Panchapakesan (1973).

5. Subset selection: Normal populations with common known variance

The goal is to select from the set G of k Normal populations with common known variance a non-empty subset containing the population associated with $\mu_{[k]}$. We determine the k sample means \bar{X}_i ($i = 1, \dots, k$), each based on n independent observations.

The procedure R given by Gupta (1965) is as follows. Select the population corresponding to \bar{X}_i iff $\bar{X}_i \in \left[\bar{X}_{[k]} - \frac{d\sigma}{\sqrt{n}}, \bar{X}_{[k]} \right]$, where d can be found in Gibbons, Olkin and Sobel (1977). The least favourable configuration LFC is given by $\mu_{[1]} = \mu_{[k]}$ and the d values were determined from

$$\int_{-\infty}^{\infty} \Phi^{k-1}(x+d)\Phi(x)dx = P^* .$$

The expected size of the selected subset is equal to

$$\begin{aligned} E(S) &= \sum_{i=1}^k P \left(\text{selecting } \pi_{(i)} \right) \\ &= \sum_{i=1}^k \int_{-\infty}^{\infty} \prod_{j=1, j \neq i}^k F \left(x+d+\mu_{[i]}-\mu_{[j]} \right) dF(x) . \end{aligned}$$

In our case of Normal populations we have

$$\max_{\Omega} E(S) = k \int_{-\infty}^{\infty} F^{k-1}(x+d)dF(x) = kP^* .$$

In the Normal case of a common known variance σ^2 and k independent samples of n independent observations, the rule R can be written as

$$R : \pi_i \in G_3 \text{ iff } \bar{X}_i \geq \bar{X}_{[k]} - d\frac{\sigma}{\sqrt{n}} \quad (i = 1, \dots, k)$$

with $\int_{-\infty}^{\infty} \Phi^{k-1}(x+d)d\Phi(x) = P^*$, and

$$E(S) = \sum_{i=1}^k \int_{-\infty}^{\infty} \prod_{\substack{j=1 \\ j \neq i}}^k \Phi \left(x+d\frac{\sqrt{n}}{\sigma} (\mu_{[i]}-\mu_{[j]}) \right) d\Phi(x) .$$

Rizvi (1963, 1971) considered the goal of selecting a non-empty subset so as include the population with the largest $\eta_i = |\mu_i|$. He uses a rule of a similar type bases on $W_i = |\bar{X}_i|$. For his procedure

$$\sup_{\Omega} E(S) = 2k \int_{-\infty}^{\infty} [2\Phi(x+d) - 1]^{k-1} d\Phi(x) ,$$

where d is as before. This bound for $E(S)$, however, exceeds kP^* .

In Gupta (1965), Deely and Gupta (1968) and Gupta (1980) some properties and measures of the performance of a subset selection rule can be found.

Among all rules satisfying the P^* -condition, the rule R presented here has the smallest value (kP^*) for $\max_{\mu} E_{\mu}(S | R)$, where $E_{\mu}(S | R)$ denotes the expected subset size S for R when μ is the true parameter configuration. It is possible that samples of unequal sizes are given. Then the rule R_4 given by Gupta and D.Y. Huang (1976) is given by ($i = 1, \dots, k$):

$$\text{select } \pi_i \text{ iff } \bar{X}_i \geq \max_{1 \leq j \leq k} \left(\bar{X}_j - d_4 \sigma \left(\frac{1}{n_i} + \frac{1}{n_j} \right)^{\frac{1}{2}} \right).$$

The constant $d_4 > 0$ depends on k, n_1, \dots, n_k (and P^*) and must be chosen to satisfy the P^* -condition. For any given association between (n_1, \dots, n_k) and $(n_{(1)}, \dots, n_{(k)})$ (with $n_{(i)}$ the sample size associated with the population with mean $\mu_{[i]}$), the infimum of $P(CS)$ is attained when $\mu_{[1]} = \mu_{[k]}$. This infimum is given by

$$\int_{-\infty}^{\infty} \prod_{j=1}^{k-1} \Phi \left(\frac{d_4 - \alpha_j x}{(1 - \alpha_j^2)^{\frac{1}{2}}} \right) d\Phi(x),$$

where

$$\alpha_j = n_{[j]}^{\frac{1}{2}} (n_{[j]} + n_{[k]})^{-\frac{1}{2}} \quad (j = 1, 2, \dots, k-1)$$

with $n_{[1]} \leq \dots \leq n_{[k]}$ the ordered n_i . An upper bound for the expected subset size given by

$$\sup_{\Omega} E(S | R_4) \leq k\Phi(d_4).$$

For $k = 2$ and equal sample sizes, the equality sign holds. Tables for d_4 -values can be found in Gupta and D.Y. Huang (1976). A different selection rule is given by Gupta and W.T. Huang (1974):

$$\text{select } \pi_i \text{ iff } \bar{X}_i \geq \bar{X}_{[k]} - d_3 \sigma n_i^{-\frac{1}{2}},$$

where the selection constant d_3 has to satisfy the next equation

$$\int_{-\infty}^{\infty} \prod_{j=1}^{k-1} \Phi \left(\frac{n_{[j]} (x + d_3)}{n_{[k]}} \right) d\Phi(x) = P^*.$$

with $n_{[1]} \leq \dots \leq n_{[k]}$ the ordered sample sizes.

**6. Normal populations with common unknown variance:
The Indifference Zone approach**

Consider the case of k Normal Populations with unknown means and common unknown variance σ^2 , which must be regarded as a nuisance parameter. The goal is to select the best population following the Indifference Zone approach. It can be seen by increasing σ^2 that there is no common fixed sample size large enough such that the usual confidence statement about the selection procedure will hold for all possible values of σ^2 . If the true value of σ^2 is sufficiently large, then the probability of a correct selection will be arbitrarily close to k^{-1} , which is smaller than any reasonable value of $P^*(k^{-1} < P^* < 1)$. So there exists no fixed sample size solution for selecting the Normal population with the largest mean satisfying the P^* -requirement for all possible σ^2 . However, for a large total sample size $\sum_{j=1}^k n_j$, we can obtain a good estimate of σ^2 by pooling the sample variances:

$$s^2 = \left(\left(\sum_{j=1}^k n_j \right) - k \right)^{-1} \sum_{j=1}^k (n_j - 1) s_j^2$$

with

$$s_j^2 = (n_j - 1)^{-1} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2 \quad \text{and} \quad \bar{X}_j = n_j^{-1} \sum_{i=1}^{n_j} X_{ij} .$$

Then we use s^2 as an approximation to σ^2 and apply a procedure for known σ^2 . If we use a two-stage procedure (cf. Bechhofer, Dunnett and Sobel (1954) and Dunnett and Sobel (1954)) instead of a one-stage procedure, the problem can be solved. In the first stage a sample of $n (> 1)$ observations is taken from each of the k populations. The pooled sample variance s^2 is equal to

$$s^2 = \nu^{-1} \sum_{i=1}^k (n - 1) s_i^2 = k^{-1} \sum_{i=1}^k s_i^2$$

with $\nu = k(n - 1)$, the number of degrees of freedom of the unbiased estimator s^2 . Then a second sample of size $N - n (\geq 0)$ is taken from each of the k populations. $N (\geq n)$ is a random variable, the value of which depends on s^2 , and equals

$$N = \max \left(n, \left[2 \left(\frac{hs}{\delta^*} \right)^2 \right] \right) ,$$

where $[y]$ denotes the smallest integer $\geq y$, and h can be obtained from Gibbons, Olkin and Sobel (1977). Calculate k over-all sample means $\bar{X}_i = N^{-1} \sum_{j=1}^N X_{ij}$ ($i = 1, \dots, k$) and select the population that produced $\max_{1 \leq i \leq k} \bar{X}_i$.

The probability of making a correct selection is at least P^* whenever $\delta = \mu_{[k]} - \mu_{[k-1]} \geq \delta^*$. It

is desirable to keep the sample size small at the second stage without making n too large since the additional information about s^2 that is available in the second stage is never used. Any prior information about σ^2 could easily be used to provide some idea of how large n should be.

7. Normal populations with Common Unknown Variance: Subset Selection

This section presents procedures following the Subset Selection approach for Normal populations with common unknown variance. First we consider the case of a common sample size n . The procedure (Gupta (1965)) is to select the population that yielded the sample mean \bar{X}_i iff

$$\bar{X}_i \geq \bar{X}_{[k]} - d_2 s n^{-\frac{1}{2}},$$

where s^2 is given in section 6.

The d_2 -values are tabulated by Gupta and Sobel (1957). They can also be obtained from table A4 with $d_2 = h\sqrt{2}$ in Gibbons, Olkin and Sobel (1977).

An advantage of the Subset Selection approach is that a single-stage procedure guarantees the subset selection probability requirement for all $\mu = (\mu_1, \dots, \mu_k)$, while for this same case a two-stage procedure is necessary to guarantee the indifference-zone probability requirement. Secondly we consider the case of unequal sample sizes. The procedure R_6 is to select population π_i that yielded \bar{X}_i iff

$$\bar{X}_i \geq \max_{1 \leq j \leq k} \left\{ \bar{X}_j - d_6 s \left(n_i^{-1} + n_j^{-1} \right)^{\frac{1}{2}} \right\}$$

(cf. Gupta and D.Y. Huang (1976)). In this case the infimum of $P(CS)$ is given by

$$\int_0^\infty \int_{-\infty}^\infty \prod_{i=1}^{k-1} \Phi \left(\frac{d_6 s - \alpha_i u}{\sqrt{1 - \alpha_i^2}} \right) d\Phi(u) dQ_\nu(s),$$

where

$$\alpha_i = \sqrt{\frac{n_{[i]}}{n_{[i]} + n_{[k]}}, \quad \nu = n_1 + \dots + n_k - k.$$

Q_ν denotes the cumulative distribution function of $\sqrt{(\mathcal{X}_\nu^2/\nu)}$, where \mathcal{X}_ν^2 is a chi-square random variable with ν degrees of freedom. We also refer to Gupta and Panchapakesan (1979; section 12.2.4).

8. Two-stage procedures

In the literature a large number of papers on two-stage selection procedures has been written. The first stage can be used to get some information about an unknown nuisance parameter. For instance, assume we have the following situation. Given are k Normal populations with a common unknown variance σ^2 . Our goal is to select the population with the largest mean. Then it is possible to use the first stage to estimate the unknown σ^2 . The second stage can

be used to indicate the best population.

In the list of publications a large number of papers can be found dealing with two-stage selection procedures. A small number of papers concern three-stage and multi-stage selection procedures.

9. Screening experiments

Two-stage procedures have been introduced also in order to use the first stage as a screening procedure. After the first stage the inferior populations can be eliminated. The second stage can be used to indicate the best population in the collection of the remaining populations. Such a procedure can be indicated as two-stage procedures with screening for selecting the best.

Assume $k(k \geq 2)$ Normal populations are given with unknown location parameters (means) $\mu_1, \mu_2, \dots, \mu_k$ and a common variance σ^2 . The t best populations are associated with $\mu_{[k-t+1]}, \dots, \mu_{[k]}$. For selecting the best population the Indifference Zone approach can be used. If σ^2 is unknown a two-stage procedure is necessary. The main purpose of the first stage is to obtain an estimate of σ^2 so that the total sample size necessary to meet the P^* -requirement can be determined.

Another possibility is to use a two-stage procedure in the known σ^2 case, where the first stage has been introduced in order to screen out inferior populations. This can be done by using the subset selection approach to select superior populations in a subset. Early investigations were confined to the special case $k = 2$, see Cohen (1959) and Alam (1970). An interesting combination of the Indifference Zone approach and the Subset Selection approach has been given by Alam (1970). He considered the problem of selecting the Normal populations with the known mean from k Normal Populations with a common known variance. In Alam's procedure selection is carried out in two stages. The first stage is used for eliminating populations with sample means much smaller than the maximum observed.

Tamhane and Bechhofer (1977) have introduced for $k \geq 2$ the following procedure, which may be especially important for screening experiments. One can see this procedure as a kind of combination of the Subset Selection approach and the Indifference Zone approach. Given k and P^* compute

$$n_1 = \left[\left(\frac{\hat{c}_1 \sigma}{\delta^*} \right)^2 \right], \quad n_2 = \left[\left(\frac{\hat{c}_2 \sigma}{\delta^*} \right)^2 \right], \quad h = \frac{\hat{d} \delta^*}{\hat{c}_1},$$

where $[a]$ denotes the smallest integer $\geq a$. The constants \hat{c}_1, \hat{c}_2 and \hat{d} can be found in Tamhane and Bechhofer (1979). Then the following stages must be carried out:

Stage 1.

Take a sample of size n_1 from each population. Let $\bar{X}_{[1]}^{(1)} < \dots < \bar{X}_{[k]}^{(1)}$, denote the ranked first-stage sample means.

- If $\bar{X}_{[k-1]}^{(1)} < \bar{X}_{[k]}^{(1)} - h$, then stop. Select the population that produced $\bar{X}_{[k]}^{(1)}$.
- If $\bar{X}_{[k]}^{(1)} - h \leq \bar{X}_{[k-s]}^{(1)} < \dots < \bar{X}_{[k-1]}^{(1)}$ for exactly s populations ($1 \leq s \leq k-1$), then proceed to the second stage.

Stage 2.

Take a sample of size n_2 from each of the $s+1$ populations that produced $\bar{X}_{[k-s]}^{(1)} < \dots < \bar{X}_{[k]}^{(1)}$.

Compute the cumulative sample mean $(n_1+n_2)^{-1} \sum_{j=1}^{n_1+n_2} X_{ij}$ for each of these $s+1$ populations.

Let $\bar{X}_{[1]}^{(1+2)} < \dots < \bar{X}_{[s+1]}^{(1+2)}$ denote the ranked cumulative sample means. Select the population that produced $\bar{X}_{[s+1]}^{(1+2)}$.

The constants \hat{c}_1, \hat{c}_2 and \hat{d} were chosen in such a way that the Indifference Zone probability requirements is guaranteed and the average number of observations required per population is minimized when $\mu_{[i]} = \mu_{[k]}$.

The procedure of Bechhofer (1954) is a special case of the two-stage procedure by letting $d = 0$.

The P^* -requirement is

$$P(CS) \geq P^* \text{ for } \mu = (\mu_1, \mu_2, \dots, \mu_k) \in \Omega(\delta^*) = \{\mu : \delta_{k,k-1} \geq \delta^*\} .$$

There are many solutions (n_1, n_2, d) . Tamhane and Bechhofer (1977) used a minimax criterion: minimize $\sup E(kn_1 + \underline{r}n_2)$, where $kn_1 + \underline{r}n_2$ is the total sample size required. However, the LFC is shown to be $\mu_{[1]} = \mu_{[k-1]} = \mu_{[k]} - \delta^*$ only in the case of $k = 2$. Tamhane and Bechhofer (1977) obtained a conservative solution by taking the infimum over $\Omega(\delta^*)$ of a lower bound of the PCS. Tamhane and Bechhofer (1979) obtained an improvement by using a sharper lower bound. This paper contains efficiency results. Their numerical study shows that the procedure is very effective as a screening procedure, especially as k increases. One can try to generalize this two-stage procedure to a selection procedure with three or more stages. One may expect that the largest increase in efficiency results from going from one to two stages.

If σ^2 is unknown a three-stage procedure can be used, where the first stage is used to determine the additional sample sizes necessary in the second and third stage. The second stage is used to eliminate inferior populations by a subset rule. The third stage is necessary (unless the size of the subset is equal to one) to make the final decision. Such procedures have been studied by Tamhane (1976) and Hochberg and Marcus (1981). It is perhaps possible to use two (instead of three) stages, where the first stage is used both to screen out inferior populations, as well as to obtain an estimate for σ^2 . Following Bechhofer (1985) an (unsolved) problem is to determine the LFC for the basic Tamhane-Bechhofer (1977) two-stage procedure for four or more populations. Such a finding would make it possible (at least theoretically) to calculate constants to implement the exact version of their procedure.

A research project (van der Laan and Rasch (1991)) has been started in order to answer the following questions:

- What are the sample sizes in the two stages in dependence of P^* (size n_1 of stage 1 fixed, size n_2 of stage 2 depends also on \underline{r})?
- What is the optimal n_1 , so that $E(kn_1 + \underline{r}n_2)$ is minimal. A good estimate of σ^2 is required.
- How robust is the two-stage selection procedure against deviations from the assumptions of common variance and normality?

Bibliography

- ALAM, K. (1970). A two-sample procedure for selecting the population with the largest mean for k Normal populations. *Ann. Inst. of Statist. Math.* **22**, 127-136.
- BECHHOFFER, R.E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Ann. Math. Statist.* **25**, 16-39.
- BECHHOFFER, R.E. (1967). A two-stage subsampling procedure for ranking means of finite populations with an application to bulk-sampling problems. *Technometrics* **9**, 355-366.
- BECHHOFFER, R.E. (1975). Ranking and selection procedures. *Proc. 20th Conf. Design Exp. Army Res. Develop. Testing, Part. 2. Durham, N.C.: U.S. Army Research Office, 929-949.*
- BECHHOFFER, R.E. (1985). Selection and Ranking Procedures - Some personal reminiscences, and thoughts about its past, present and future. *American Journal of Mathematical and Management Sciences. Vol. 5, NOS, 3 & 4, 201-234.*
- BECHHOFFER, R.E. ET AL (1986). Two-stage selection of the best factor-level combination in multifactor experiments: Common unknown variance. *Statis. Des. Ph.D.*, 3-15.
- BECHHOFFER, R.E., DUNNET, C.W. AND SOBEL, M. (1954). A two-sample multiple decision procedure for ranking means of normal populations with a common unknown variance. *Biometrika* **41**, 170-176.
- BECHHOFFER, R.E., KIEFER, J. AND SOBEL, M. (1968). *Sequential Identification and Ranking Procedures.* The University of Chicago Press, Chicago.
- BECHHOFFER, R.E. AND SOBEL, M. (1954). On a sequential ranking procedure (preliminary report). *Abstract, Bull. Amer. Math. Soc.* **60**, 34-35.
- BOFINGER, E. (1979). Two-stage selection problem for normal populations with unequal variances. *Austr. J. of Statist.* **21**, 149-156.
- CARROLL, R.J., GUPTA, S.S. AND HUANG, D.Y. (1975). On selection procedures for the t best populations and some related problems. *Commun. Statist. - Theor. Meth.* **4**, 987-1008.
- COHEN, D.S. (1959). A two sample decision procedure for ranking means of normal populations with a common known variance. Unpublished M.S. Thesis, Dept of Ind. Eng., Cornell Univ., Ithaca, N.Y.
- DESU, M.M., NARULLA S.C. AND VILLARREAL, B. (1977). A two-stage procedure for selecting the best of k exponential distributions. *Commun. Statist. - Theor. Meth. A* **6**, 1223-1230.
- EATON, M.L. (1967). Some optimum properties of ranking procedures. *Ann. Math. Statist.* **38**, 124-137.

- EHRMAN, C.M., KRIEGER, A., MIESCKE, K.J. (1987) Subset selection toward optimizing the best performance at a second stage. *J. of Business and Economic Statistics* **5**, 295-303.
- FABIAN, V. (1974a). Note on Anderson's sequential procedures with triangular boundary. *Ann. Statis.* **2**, 170-176.
- FABIAN, V. (1974b). Acknowledgement of priority to "Note on Anderson's sequential procedures with triangular boundary". *Ann. Statis.* **2**, 1063.
- FINNEY, D.J. (1984). Improvement by planned multistage selection. *J. Am. Statist. Assoc.* **79**, 501-509.
- GIBBONS, J.D., OLKIN, I. AND SOBEL, M. (1977). *Selection and Ordering Populations: A New Statistical Methodology*. John Wiley & Sons, Inc., New York.
- GUPTA, S.S. (1965). On some multiple decision (selection and ranking) rules. *Technometrics* **7**, 225-245.
- GUPTA, S.S. ET AL (1984). A two-stage elimination type procedure for selecting the largest of several normal means with a common unknown variance. *Design of Experiments. Ranking and Selection* (Santner and Tamhane, eds.). Dekker, New York.
- GUPTA, S.S. AND HUANG, D.Y. (1976). Selection procedures for the means and variances of normal populations: unequal sample sizes case. *Sankhyā Ser. A*, **38**, 112-128.
- GUPTA, S.S. AND HUANG, W.T. (1974). A note on selecting a subset of normal populations with unequal sample sizes. *Sankhyā Ser. B*, **36**, 389-396.
- GUPTA, S.S. AND LIANG, T. (1989). On Bayes and empirical Bayes two-stage allocation procedures for selection problems. *Statist. Dat. An. Inf.* 61-70.
- GUPTA, S.S. AND MIESCKE, K.J. (1982). On the least favourable configurations in certain two-stage selection procedures. *Stat. Prb. CRR*, 295-305.
- GUPTA, S.S. AND MIESCKE, K.J. (1984). On two-stage Bayes selection procedures. *Sankhyā B* **46**, 123-134.
- GUPTA, S.S. AND MIESCKE, K.J. (1987). Optimum two-stage selection procedures for Weibull populations. *J. Statist. Planning and Inference* **15**, 147-156.
- GUPTA, S.S., NAGEL, K., AND PANCHAPAKESAN, S. (1973). On the order statistics from equally correlated normal random variables. *Biometrika* **60**, 403-413.
- GUPTA, S.S. AND PANCHAPAKESAN, S. (1972). On a class of subset selection procedures. *Ann. Math. Statis.* **43**, 814-822.
- GUPTA, S.S. AND PANCHAPAKESAN, S. (1979). *Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations*. John Wiley and Sons, New York-London.
- GUPTA, S.S. AND PANCHAPAKESAN, S. (1985). Subset Selection procedures: Review and assessment. *American Journal of Mathematical and Management Sciences*. Vol. **5**, NOS. 3 & 4, 235-312.

- GUPTA, S.S AND SOBEL, M. (1957). On a statistic which arises in selection and ranking problems. *Ann. Math. Statist.* **28**, 957-967.
- HAN, K.S., KIM, W.C. (1987). A two-stage selection procedure for exponential populations. *J. Korea Stat.* **16**, 37-44.
- HALL, W.J. (1959). The most-economical character of some Bechhofer and Sobel decision rules. *Ann. Math. Statist.* **30**, 964-969.
- HOCHBERG, Y., AND MARCUS, R. (1981). Three-stage elimination type procedures for selecting the best normal population when variances are unknown. *Commun. Statist. - Theor. Meth. A* **10**, 597-612,
- HOEL, D.G., SOBEL, M. AND WEISS, G.W. (1972). A two-stage procedure for choosing the better of two binomial populations. *Biometrika* **59**, 317-322.
- LAAN, P. VAN DER, AND RASCH, D. (1991). A two-stage procedure with screening for selecting the best. Unpublished report.
- LAAN, P. VAN DER, AND VERDOOREN, L.R. (1989). Selection of populations: An overview and some recent results. *Biometrical Journal* **31**, 383-420.
- LAM, K. (1988). An improved two-stage selection procedure. *Commun. Statist. B* **17**, 995-996.
- LAWING, W.D. AND DAVID, H.T. (1966). Likelihood ratio computations of operating characteristics. *Ann. Math. Statist.* **37**, 1704-1716.
- LEE, S. AND KIM, W. (1985). An elimination type two-stage selection procedure for exponential distributions. *Commun. Statist. A* **14**, 2563-2571.
- MCMAHAN, C.A. (1982). Regression toward the mean in a two-stage selection program. *American J. of Epidemiology*, **116**, 394-401.
- MCMAHAN, C.A. (1987). Regression toward the mean in a two-stage selection program. II. Correlated within-subject observations. *American J. of Epidemiology*, **125**, 912-916.
- MIESCKE, K.J. (1984). Two-stage selection procedures based on tests. *Design of Experiments. Ranking and Selection* (Santner and Tamhane, eds.) Dekker, New York, 165-178.
- MUKHOPADHYAY, N. (1979). Some comments on two-stages selection procedures. *Commun. Statist. - Theor. Meth. A* **8**, 671-684.
- MUKHOPADHYAY, N. (1983). Theoretical investigations of some sequential and two-stage procedures to select the larger mean. *Sankhyā A* **45**, 346-356.
- MUKHOPADHYAY, N. (1984). Sequential and two-stage procedures for selecting the better exponential population covering the case of unknown and unequal scale parameters. *J. Statist. Planning and Inference* **9**, 33-43.
- MUKHOPADHYAY, N. AND HAMDY, H.I. (1983). Two-stage procedure for selecting the best exponential populations when the scale parameters are unknown and unequal. *Sequential Analysis* **3**, 51-74.

- MUKHOPADHYAY, N. (1987). Three-stage procedures for selecting the best exponential population. *J. Statist. Planning and Inference* **16**, 345-352.
- OFOSU, J.B. (1975). A two-stage minimax procedure for selecting the normal population with the smallest variance. *J. Amer. Statist. Assoc.* **70**, 171-174.
- PAULSON, E. (1964). A sequential procedure for selecting the population with the largest mean from k normal populations. *Ann. Math. Statist.* **35**, 174-180.
- RAMBERG, J. (1966). A comparison of the performance characteristics of two sequential procedures for ranking the means of normal populations. Tech. Report 4, Dept. of Ind. Eng. and Operations Res. Cornell Univ., Ithaca, N.Y.
- RASCH, D. (1984). Einführung in die mathematische Statistik II. Varianzanalyse, Regressionsanalyse und weitere Anwendungen. VEB Deutscher Verlag der Wissenschaften, Berlin.
- RINOTT, Y. (1978). On two-stage procedures and related probability-inequalities. *Commun. Statist. - Theor. Meth. A* **7**, 799-811.
- RÜCKER, G. (1989). A two-stage trial design for testing treatment, selfselection and treatment preference effects. *Statist. Med.* **8**, 477-485.
- SANTNER, T.J. (1976). A two-stage procedure for selection of δ^* -optimal means in the normal case. *Commun. Statist. - Ther. Meth. A* **5**, 283-292.
- SEHR, J. (1988). On a conjecture concerning the least favorable configuration of a two-stage selection procedure. *Commun. Statist. A* **17**, 3221-3233.
- SIMMONDS, N.W. (1985). Two-stage selection strategy in plant breeding. *Heredity* **55**, 393-399.
- SOMERVILLE, P.N. (1971). A technique for obtaining probabilities of correct selection in a two-stage selection problem. *Biometrika* **58**, 615-623.
- SOMERVILLE, P.N. (1974). On allocation of resources in a two-stage selection procedure. *Sankhyā Ser. B* **36**, 194-203.
- SRIVENKATARAMANA, T AND TRACY, D.S. (1989). Two-phase sampling for selection with probability proportional to size in sample surveys. *Biometrika* **76**, 818-821.
- SWANEPOEL J.W.H. AND GEERTSEMA, J.C. (1976). Sequential procedures with elimination for selecting the best of k normal populations. *S. Afr. Statist. J.* **10**, 9-36.
- SWANEPOEL J.W.H. AND VENTER, J.H. (1975). A class of elimination selection procedures based on ranks. *S. Afr. Statist. J.* **9**, 119-128.
- TAMHANE, A.C. (1975a). A minimax multistage elimination type rule for selecting the largest normal mean. Ph.D. Dissertation. Techn. Rep. No. 259. Dept. of Operations Research, Cornell Univ. Ithaca, New York.
- TAMHANE, A.C. (1975b). A minimax two-stage permanent elimination type procedure for selecting the smallest normal variance. Techn. Rep. No. 260. Dept. of Operations Research, Cornell Univ., Ithaca, New York.

- TAMHANE, A.C. (1976). A three-stage elimination type procedure for selecting the largest normal mean (common unknown variance case). *Sankhyā B* **38**, 339-340.
- TAMHANE, A.C. (1977). On a class of multistage selection procedures with screening for the normal means problem. Discussion Paper No. 27, Center for Statistics and Probability, Northwestern Univ., Evanston, Illinois.
- TAMHANE, A.C. AND BECHHOFFER, R.E. (1977). A two-stage minimax procedure with screening for selecting the largest normal mean. *Commun. Statis. - Theor. Meth. A* **6**, 1003-1033.
- TAMHANE, A.C. AND BECHHOFFER, R.E. (1979). A two-stage minimax procedure with screening for selecting the largest normal mean (II): an improved PCS lower bound and associated tables. *Commun. Statist. - Theor. Meth. A* **8**, 337-358.
- TAYLOR, R.J. AND DAVID, H.A. (1962). A multistage procedure for the selection of the best of several populations. *J. Amer. Statis. Assoc.* **57**, 785-795.
- THALL, P.F., SIMON, R. AND ELLENBERG, S.S. (1988). Two-stage selection and testing designs for comparative clinical trials. *Biometrika* **75**, 303-310.
- THIBODEAU, L.A. (1980). Evaluating selection criteria. *Am. Statist. Assoc., Proceed. of Social Statist. Section* 283-284.
- TONG, Y.L. (1970). Multi-stage interval estimation of the largest mean of k normal populations. *J. Roy. Statis. Soc. Ser. B* **32**, 272-277.
- WETHERILL, G.B. AND OFOSU, J.B. (1974). Selection of the best of k Normal populations. *J.R. Statist. Sco. C, Appl. Statis.* **23**, 253-277.
- WILCOX, R.R. (1979). A two-stage procedure for selecting the best of several binomial populations. *Educational and Psychological Measurement* **39**, 715-724.
- WILCOX, R.R. (1983). Approximating the probability of selecting the best treatment with a heteroscedastic procedure when the first stage has unequal sample sizes. *J. of Educational Statistics* **8**, 45-58.

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

<u>Number</u>	<u>Month</u>	<u>Author</u>	<u>Title</u>
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.