An overview of a generalization in statistical selection

Citation for published version (APA):

Document status and date:
Published: 01/01/1995

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

Memorandum COSOR 95-20
An Overview of a Generalization in Statistical Selection
P. van der Laan
F.P.A. Coolen

Eindhoven, June 1995
The Netherlands
An Overview of a Generalization in Statistical Selection

Paul van der Laan
Eindhoven University of Technology
Department of Mathematics and Computing Science
Eindhoven, The Netherlands

and

Frank Coolen
University of Durham
Department of Mathematical Sciences
Durham, England

Summary

Some introductory remarks are made about statistical selection. Statistical selection procedures are answering questions like "Which variety can be considered to be the best?". The principles of the Indifference Zone approach of Bechhofer are summarized.
A generalization of the concept of the Indifference Zone selection is presented. The Indifference Zone approach is generalized by introducing a preference threshold. By this way there are three possibilities of decision: Correct Selection, False Selection and No Selection. A practical application in the field of Oil Palm cultivation is given.

AMS Subject Classification: 62F07.

Key Words: Indifference Zone; probability of correct selection; probability of false selection; probability of no selection; generalized selection goal; normal populations; preference threshold.

1. Introduction

In practice we are often confronted with the problem of selection of the best population or best treatment. Especially in the field of biometry (e.g. testing varieties) statistical selection is often an interesting feature. Selection problems often need a quantitative methodology of selection. The ordinary attack, using ANOVA techniques, is in some cases not completely adequate, in the sense that the formulation of the problem is not always realistic.

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 expected yield per unit plot. If there are more than one contenders for the best, because there are ties, it is assumed that one of these is appropriately tagged. We assume that the selection is based on the average yield per unit plot of constant size.

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.

In the theory of statistical selection there are two main approaches. One of the two main approaches for selection of the best variety or treatment is the so-called Indifference Zone approach of Bechhofer (see Bechhofer, 1954, and Gupta and Panchapakesan, 1979). The second one is the Subset Selection approach of Gupta (see Gupta, 1965, and Gupta and Panchapakesan, 1979).

The Subset Selection procedure selects a subset, nonempty and as small as possible, with the probability requirement that the probability of a correct selection \( CS \) is at least \( P^* \). \( CS \) means in this context that the best variety or treatment is an element of the selected subset. So

\[
P(\ CS \ ) \geq P^*,
\]

where

\[
1/k < P^* < 1.
\]

The size \( S \) of the subset, defined as the number of varieties or treatments in the subset, is a random variable. Tables with values of the selection constant \( d \) required by Gupta's selection rule, given in Gupta (1965), can be found in Gibbons, Olkin and Sobel (1977).

A short description of the basic approach of Bechhofer will be given in section 2. Some general remarks about modifications
and generalizations as well as about comparison of both main approaches are made in section 3. In section 4 a generalization of the Indifference Zone selection approach using a preference threshold is given. Section 5 gives an application in the field of Oil palm cultivation. Finally, in section 6 some concluding remarks are given.

2. Statistical selection: The main approach of Bechhofer

The basic approach of Bechhofer will be shortly reviewed. Assume \( k \) (integer \( k \geq 2 \)) independent Normal random variables \( X_1, \ldots, X_k \) are given. These variables are associated with the \( k \) varieties or treatments indicated by \( T_1, \ldots, T_k \), and are for instance sample yields. The assumed Normal distributions have common known variance \( \sigma^2 \) and unknown means \( \theta_1, \ldots, \theta_k \). The goal is to select the variety with mean \( \theta_{[k]} \), where

\[
\theta_{[1]} \leq \ldots \leq \theta_{[k]}
\]
denote the ordered values of \( \theta_1, \ldots, \theta_k \). Let CS denotes correct selection, this means that the selected variety is in fact the best variety.

The approach of Bechhofer is the so-called Indifference Zone approach (Bechhofer, 1954). The goal is to indicate or select the best variety. The selection rule is to select the variety that resulted in the largest sample mean. The confidence or probability requirement is that the probability of a CS is at least \( P^* \), whenever the best variety is at least \( \delta^* \) away from the second best. In this context CS means that the best variety with mean \( \theta_{[k]} \) produced the largest sample mean and consequently it is also selected as the best variety. The minimal probability \( P^* \) can only be guaranteed if the common sample size \( n \) is large enough.

The parameter space is defined as

\[
\Omega = \{ \theta \in \mathbb{R}^k : \theta = (\theta_1, \theta_2, \ldots, \theta_k) \}.
\]

Bechhofer (1954) introduced the next measure of distance

\[
\delta = \theta_{[k]} - \theta_{[k-1]}.
\]

So the probability requirement
with $1/k < P^* < 1$, has to hold only for all $\theta \in \Omega(\delta')$, where the subspace $\Omega(\delta')$ is defined as

$$\Omega(\delta') = \{ \theta: \delta \geq \delta' > 0 \},$$

the so-called Preference Zone. P* and $\delta'$ have to be specified by the experimenter. The problem is to determine the common sample size $n (= n_i$ for $i = 1, 2, ..., k)$ for which

$$\inf P( CS ) \geq P^*,$$

where the infimum is taken over the Preference Zone. For this location parameter case the Least Favourable Configuration (LFC), where the $P( CS )$ for Bechhofer's selection procedure is minimal, is given by

$$\theta_{[1]} = \theta_{[k]} = \theta_{[k]} - \delta'.$$

and is part of the preference zone.

In the case considered, where the observations $X_{ij}$ ($j = 1, 2, ..., n$) on $X_i$ are Normally distributed with mean $\theta_i$ and common known variance $\sigma^2$ ($i = 1, 2, ..., k$) and all $X_{ij}$ are independent of each other, one can find (Bechhofer, 1954) that

$$P_{LFC} = \int_{-\infty}^{+\infty} \phi^{k-1}(x+\tau) \, d\Phi(x),$$

with $\phi(.)$ is the Normal cumulative distribution function and

$$\tau := \delta' \sqrt{n} / \sigma.$$

If one requires that

$$P( CS \mid LFC ) \geq P^*,$$

then the value of $\tau$ follows, and thus

$$n = ( \tau \sigma / \delta' )^2.$$

Tables for $\tau = \delta' \sqrt{n} / \sigma$ can be found in for instance Gibbons, Olkin and Sobel (1977). With the chosen minimal $n$ it can be guaranteed with minimal probability $P^*$ that the selected variety is less than $\delta^*$ away from the best.
3. Modifications and some general remarks

In the literature different goals are considered. We mention a few:

a. Selecting the t best varieties, where integer t is larger than or equal to 2. A possibility is to produce a collection of t varieties without ranking them.

b. Selecting a subset that contains only good varieties.

c. Selecting a collection of varieties which will contain at least the t ( t ≥ 2 ) best varieties.

d. Selecting a random number of varieties such that all varieties better than a standard variety are included in the selected subset.

e. Selecting a subset whose size is smaller than or equal to m ( 1 ≤ m < k ) and which will include at least one good variety.

In the literature different generalizations and modifications have been proposed. We refer to Gupta and Panchapakesan ( 1979 ) for references, see also Gupta and Panchapakesan ( 1985 ) and Rizvi ( 1985, 1986 ).

Subset Selection is a flexible form of selection, because the number of replications has not to be determined in advance. After the experiment has been carried out, the selection can be prosecuted. The influence of the number of replications can be conducted from the ( expected ) size of the subset. A relatively large subset means, apart from random fluctuations, that the number of replications is small or the variety means are close together, or both. If a correct selection CS is defined as the event that the best variety is in the subset then the probability on CS can be compared with the power of a test. Both characteristics indicate the probability of a correct decision while the variety may be ( or are ) different. The probability distribution of the size S, the number of treatments in the selected subset, can be found in van der Laan ( 1995 ). Also the distribution, expectation and variance of S for the Least Favourable Configuration are given.
Whereas Subset Selection can be used as a screening procedure, the Indifference Zone approach produces, in a certain sense, a more precise result. For the last method indicates the best variety, with a certain confidence. A condition is that a minimal number of observations have been done.

4. A generalization of the Indifference Zone approach using a preference threshold

It is possible to generalize, in a certain sense, the concept of Indifference Zone selection by introducing a preference threshold (Coolen and van der Laan, 1995 a,b). Again we consider k independent samples of common size from normal populations with equal known variance. Starting with the preference zone given by the parameter subspace

$$\Omega(\delta^*) = \{ \Theta = (\Theta_1, \Theta_2, \ldots, \Theta_k) \in \mathbb{R}^k : \theta_{[k]} - \theta_{[k-1]} \geq \delta^* > 0 \}.$$ 

we accept three kinds of decisions, namely CS, FS ( = False Selection) and NS ( = No Selection). We apply the selection rule $R_c$: Select population i if and only if

$$\Sigma_{j=1}^{n} (Y_{ij} - Y_{ij}) > c,$$

for $l = 1, 2, \ldots, k$; $l+i$, with the so-called threshold $c \geq 0$. It is possible to require that for the parameter configuration $\Omega(\delta^*)$ the following holds

$$P(CS) \geq p^* \quad \text{and} \quad P(FS) \leq q^*.$$

These two conditions turn out to be

$$\int \Phi^{k-1}(z+\tau_{c,k}) \, d\Phi(z) = p^*$$

and

$$(k-1) \int \Phi^{k-2}(z-\frac{1}{2}(\tau_{f,k}-\tau_{c,k})) \, \Phi(z-\tau_{f,k}) \, d\Phi(z) = q^*$$

with

$$\tau_{c,k} = (n\delta^*-c)/(\sigma/\sqrt{n})$$
and

$$\tau_{f,k} = (n\delta' + c) / (\sigma\sqrt{n}) .$$

For details we refer to Coolen and van der Laan (1995a). The two conditions determine these two constants, and lead to

$$n = \left\{ \sigma \left( \tau_{c,k} + \tau_{f,k} \right) / (2\delta) \right\}^2$$

$$c = \sigma^2 \left( \tau^2_{f,k} - \tau^2_{c,k} \right) / (4\delta).$$

These $n$ and $c$ are such that in $\Omega(\delta')$ both probability requirements are satisfied for given $\delta'$, $P'$ and $Q'$, when using selection rule $R_c$. If $c=0$ we get the standard Indifference Zone selection procedure of Bechhofer.

5. An application

The generalized selection procedure will be illustrated using an application in the field of Oil Palm cultivation (see van der Laan and Verdooren, 1990). Given is an experiment with 10 families of Oil Palm in four complete blocks. The response variable $x$ is the percentage Magnesium content. The reason for this response variable is the good correlation between the percentage Magnesium content and the yield of oil for the first five years of production. The experimental results are summarized in the following table.

<table>
<thead>
<tr>
<th>Family $V_i$</th>
<th>$x_i$ Mg</th>
<th>Rank number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.212</td>
<td>[5]</td>
</tr>
<tr>
<td>2</td>
<td>0.222</td>
<td>[7]</td>
</tr>
<tr>
<td>3</td>
<td>0.242</td>
<td>[8]</td>
</tr>
<tr>
<td>4</td>
<td>0.204</td>
<td>[3]</td>
</tr>
<tr>
<td>5</td>
<td>0.210</td>
<td>[4]</td>
</tr>
<tr>
<td>6</td>
<td>0.186</td>
<td>[2]</td>
</tr>
<tr>
<td>7</td>
<td>0.218</td>
<td>[6]</td>
</tr>
<tr>
<td>8</td>
<td>0.244</td>
<td>[9]</td>
</tr>
<tr>
<td>9</td>
<td>0.162</td>
<td>[1]</td>
</tr>
<tr>
<td>10</td>
<td>0.248</td>
<td>[10]</td>
</tr>
</tbody>
</table>
The standard deviation of the % Mg determination is known to be 0.0186. The selection procedure of Bechhofer with $P' = 0.90$, $k = 10$ and $n = 4$ gives $\tau_{.90, 10} = 2.98293$, thus

$$\delta' = 0.0186 \times 2.98293 / \sqrt{4}$$

$$= 0.028.$$ 

Otherwise we can also determine, before the experiment, the number $n$ of complete blocks to determine a $\delta' = 0.01$. This leads to

$$n = (0.0186 \times 2.98293 / 0.01)^2$$

$$= 31.$$ 

For $\delta' = 0.02$ we find $n = 8$.

If we require

$$P(\text{FS} | R_c) \leq 0.05$$

for $\theta_{[1]} = \theta_{[k-1]} = \theta_{[k]} - 0.01$, then

$$n = \left(\frac{0.0186 \times (2.98293 + 3.35582)}{2 \times 0.01}\right)^2$$

$$= 35$$

and

$$c = 0.0186 \times \frac{(3.35582^2 - 2.98293^2)}{(4 \times 0.01)}$$

$$= 0.020.$$ 

6. Some concluding remarks

Especially in the field of variety testing many problems are in fact selection problems. Ultimately, we want often to find the best variety, where best is defined in a less or more complicated manner. We think it is important to investigate the possibilities to use statistical selection procedures for certain problems in variety testing. The first thing we need is to formulate adequately the problem. If the problem we
consider is a selection problem, then it must be considered to formulate the problem as a selection problem. An exact formulation as a selection problem is worthwhile and then an analysis (exact or approximate) is required. The use of selection procedures with a certain confidence requirement can help us to study practical problems in a realistic way. Not for all designs of experiments this problem has been solved. Finally, we refer to Gupta (1977), Gibbons, Olkin and Sobel (1979) and Dudewicz (1980) for an introduction to statistical selection procedures.

References


