Statistical selection : a way of thinking!

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.

Download date: 09. Dec. 2019
Memorandum COSOR 95-25

Statistical selection: A way of thinking

P. van der Laan

Eindhoven, August 1995
The Netherlands

1Voor Vriendenboek aangeboden aan Prof.dr. F.E.J. Kruseman Aretz
Eindhoven University of Technology
Department of Mathematics and Computing Science
Probability theory, statistics, operations research and systems theory
P.O. Box 513
5600 MB Eindhoven – The Netherlands

Secretariat: Main Building 9.10
Telephone: 040-47 3130

ISSN 0926 4493
STATISTICAL SELECTION: A WAY OF THINKING!

Paul van der Laan

'An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.'
John Tukey

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

'If you think that statistical selection is of no importance, think again!'

Summary

Statistical selection of the best population is discussed in general terms and the principles of statistical selection procedures are presented. Advantages and disadvantages of Subset Selection, one of the main approaches, are indicated. The selection of an almost best population is considered and compared with the selection of the best one from an application point of view.

1. Introduction

In practice we are often confronted with the problem of selection. For instance in the field of testing varieties, testing drugs and choosing the optimal production process statistical selection is often an essential feature.

For all kinds of selection problems a quantitative methodology of selection is needed. The problem of statistical selection concerning several populations (varieties, treatments, drugs, processes, etc.) has long been a concern for statisticians. Let us consider the problem of selecting the best population from a number \( k \) (integer \( k \geq 2 \)) of populations. The best population is defined as the population with the largest expectation. If there are more than one contenders for the best because there are ties, it is assumed that one of these is appropriately tagged.

Furtheron, we assume that the experiment has designed as a complete randomized design with \( n \) observations for each population. Statistical selection procedures can possibly help us to improve our selection process. 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 a special approach of statistical selection, namely Subset Selection, will be mentioned in Section 4.
It is important that selection of the best population can be guaranteed in some sense or another. The principle of a statistical selection procedure is that the probability of an error, an incorrect selection, is under control. It is possible to compare, in a certain sense, the situation with that of statistical testing of a hypothesis H. The rejection of H, while in reality H is true, is indicated by 'an error of the first kind'. The characteristic feature of a statistical test is that the probability of an error of the first kind is under control, say smaller than or equal to 0.05. A second aspect is the probability that H will be rejected if in reality H is false. The probability that H will be rejected, if it is false, is called the power (of the test considered). A requirement may be that in a certain situation the power must be larger than 0.90, say. It is possible to compare the probability of a correct selection using a statistical selection procedure, with the power of a statistical test. One can say that this probability is as important as the power of a test. To be sure, or almost sure, that we don't miss the best population, the probability of correct selection of the best population has to be taken into account. Of course, from a practical point of view, one may be satisfied with an almost best population instead of the best population. But also in that situation we want to have a confidence requirement that the best or an almost best population will be selected. We return to this point in Section 5. Finally, we conclude with some comments in Section 6.

2. Statistical selection

The main approaches in handling with selection of the best population are the Subset Selection approach (for details we refer to Gupta and Panchapakesan; 1979) and the Indifference Zone approach (for details we refer to Gibbons, Olkin and Sobel; 1977, Gupta and Panchapakesan, 1979). These two basic approaches will be outlined in a few words.

Assume k (k ≥ 2) independent Normal random variables X1,...,Xk are given. These variables are associated with the k populations indicated by V1,...,Vk and may be sample means. The assumed Normal distributions have common known variance σ² and unknown means µ1,...,µk. The assumption of a common known variance has been made for simplicity. The goal is to select the population with mean µ[k], where µ[1] ≤ ... ≤ µ[k] denote the ordered values of µ1,...,µk. Let CS denote correct selection.

The Subset Selection procedure selects a subset, nonempty and as small as possible, with the probability requirement that the probability of a 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 Subset Selection rule is defined as follows:

Select population Vi in the subset if and only if

\[ X_i \geq \max_{1 \leq j \leq k} X_j - d \sigma / \sqrt{n} \]

(i=1,...,k), where Xi is the sample mean of the i-th sample.

The selection constant d (d > 0) must be determined in such a way that P(CS) ≥ P*, with 1/k < P* < 1, for all possible parameter configurations. Tables with values of the selection constant d can be found in, for instance, Gibbons, Olkin and Sobel (1977).
The second approach is the so-called Indifference Zone approach. The goal is to indicate or select the best population. The procedure is to select that population that resulted in the largest sample mean. The probability requirement is that the probability of a CS is at least $P^*$, with $1/k < P^* < 1$, whenever the best population is at least $\delta^* (> 0)$ away from the second best. In this context CS means that the best population with mean $\mu_{1k}$ produces the largest sample mean and consequently it is also indicated as the best population. The minimal probability $P^*$ can only be guaranteed if the common sample size $n$ is large enough. From the probability requirement the next quantity

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

can be computed and can be found in, for instance, Gibbons, Olkin and Sobel (1977). Notice that the following holds

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

With this chosen (minimal) value of $n$ it can be guaranteed with minimal probability $P^*$ that the selected population is less than $\delta^*$ away from the best one.

3. Selection goals

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

i. Selecting the best population.

ii. Selecting the $t$ best populations, where integer $t$ is larger than or equal to 2. We can do this with or without ordering of the populations. In the first case we indicate a population as the best one, another population as the second best, etc. In the second case we produce a collection of $t$ populations without ranking them.

iii. Selecting a subset that contains only good populations.

iv. Elimination of inferior (or bad) populations.

v. Ordering (ranking) all populations from worst to best.

vi. Selecting a collection of populations, that will contain at least the best population.

vii. Selecting a collection of populations which will contain at least the $t$ ($t \geq 2$) best populations.

viii. Selecting a fixed number of populations that includes the $t$ best populations.

ix. Selecting a random number of populations such that all populations better than a standard population are included in the selected subset.

x. Selecting a subset whose size is smaller than or equal to $m$ ($1 \leq m < k$) and which will include at least one good population.

xi. Selecting a subset that excludes all the non-$t$-best populations, where a population is strictly non-$t$-best if the distance (in mean) for this population to the $t$-th best population is larger than or equal to a certain amount.
In the literature different generalizations and modifications have been proposed. We refer to Gupta and Panchapakesan (1979) for references.

4. Subset Selection

Subset Selection is in a certain sense 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 means of the populations are close together, or both. If a correct selection \( CS \) is defined as the event that the best population is in the subset then the probability on \( CS \) can be compared with the power of a test. Both characteristics indicate the probability on a correct decision while the populations may be (or are) different. 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 population, at least we hope so. A condition is that a minimal number of observations have been done. Moreover, when the variance is unknown then the Subset Selection can be carried out using a different value for the selection constant \( d \), but the Indifference Zone method requires a procedure which consists of two steps. The first step is necessary to estimate the unknown variance in order to make it possible to derive a value for the required common sample size for the samples of the second step.

5. An almost best population

The requirement to select the best population may be a strong one if the best population is not far away from the other populations. Let us assume that the best population and the second best are near each other. More accurately: the expectations of these two populations are close together, say on a distance less than \( \varepsilon \), where \( \varepsilon > 0 \), but relatively small. In such a situation it is often not of practical interest whether one selects the best population or the next best. In this situation the next best population can be indicated as an almost best population. The same idea can be extracted from the Indifference Zone for which the distance between the best population and the next best is smaller than \( \delta^* > 0 \). Not every difference in expectation is important. In many real world problems it is of interest to see whether the best or an almost best population can be selected. This idea leads to the consideration that one may generalize the selection goal to a selection of an almost best population. A consequence of this generalization is that the least favourable configuration becomes more difficult. Not an essential disadvantage using computers. The goal is to select a small but nonempty subset such that the selected subset will contain the best population or an almost best one with a high probability. In general, the generalization to selecting an almost best population will result in subsets of smaller expected size.
The concept of ε-best started, in fact, with the notion of delta-correct ranking introduced by Fabian (1962) and Lehmann (1963). For the location problem, Van der Laan (1992) studied the use of an ε-best population in Subset Selection with as goal the selection of a non-empty subset which includes at least one ε-best population with a minimal guaranteed probability P* (1/k < P* < 1).

6. Some concluding remarks

During decades of years we are used to apply statistical tests, like analysis of variance tests, to problems that are real selection problems. In a number of applications we want, ultimately, to find the best population, 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 the field of application. The first thing we need is to formulate adequately the problem. If a problem is a selection problem, then formulation as a selection problem has to be considered. Following this statement an exact formulation as a selection problem is worthwhile and then an analysis (exact or approximate) is required. Not for all designs of experiments this problem has been solved. For a number of designs a simulation is feasible in order to find an accurate estimate of the selection constant required for the prosecution of the selection procedure.

The change from analysis of variance type techniques to selection type techniques is presumably not simple. It is a change of 'a way of thinking'. We notice that sofar a lot of theoretical problems, not all problems, have been solved. We refer to Dudewicz (1980), Gibbons, Olkin and Sobel (1977, 1979), Gupta (1977) and Van der Laan (1987) for an introduction to statistical selection.

References

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

Fabian, V. (1962).


An introduction to ranking and selection. The American Statistician 33, 185-195.


