Subset selection

TU/e technische universiteit eindhoven

SPOR-Report 2000-10

Subset Selection

Paul van der Laan

SPOR-Report
Reports in Statistics, Probability and Operations Research

Eindhoven, June 2000
The Netherlands

/ department of mathematics and computing science

SUBSET SELECTION[1]

Paul van der Laan
Eindhoven University of Technology

SUMMARY

Assume k ($k\geq2$) populations are given. The associated independent random variables have continuous distribution functions differing only in their unknown location parameter. The statistical selection goal of subset selection is to select a non-empty subset which contains the best population with confidence level P^*, with $k^{-1}<P^*<1$. In this context, best population means the population with the largest value of the location parameter. The subset selection approach is described and some properties are presented.

Keywords: selection, probability of correct selection, expected subset size, variance of subset size, least favourable configuration.

INTRODUCTION

In practice we often have to make decisions. Before a decision can be made, in general, a careful statistical analysis is needed. In statistical inference the basic formulation of hypothesis testing is not adequate in a number of cases. If a technician is investigating a number k of chemical processes, characterised for example by the population mean yield μ per standard time period, he usually wants to select, with a certain confidence, the best process or a subset from the set of k processes that contains the best process. In this context of population means, best is defined by having the maximal population mean. For this, in nature of things, selection problem the hypothesis testing formulation is not the most suitable one. We suppose throughout this paper that the k populations are described by qualitative variables. Statistical selection procedures are developed in order to select the best population from qualitative populations.

There are two main approaches for statistical selection, namely the indifference zone approach and the subset selection approach. The pioneer Bechhofer (1954) proposed the indifference zone approach. The selection approach of Bechhofer has as its goal to

----

[1] Submitted to the Encyclopedia of Environmetrics, to appear on John Wiley, New York, in 2001.

indicate (to select) the best population from the k populations with a certain confidence. Gupta (1956,1965) proposed the subset selection approach. His goal is to select a subset from the k populations in order to include the best population with a certain confidence.

In the field of statistical selection theory introductory papers are written by Gupta (1977), Gibbons et al. (1979), Dudewicz (1980) and Van der Laan (1987). Introductions to selection procedures can also be found in Dudewicz (1976) and Gibbons et al. (1977). Especially the last book, containing 148 pages with tables and 14 pages with references, can be considered as a bridge between theory and practice! Of interest is also the standard book of Gupta and Panchapakesan (1979), which is more theoretical of nature and contains about 750 references. General papers can also be found in the proceedings of the conference Statistical Ranking and Selection-Three Decades of Development (1984) edited by Rizvi (1985,1986). As an example we mention Gupta and Panchapakesan (1985).

In the next section we shall describe the subset selection approach proposed by Gupta (1956,1965). After that some general properties of subset selection are presented, as well as a short descriptive comparison of the two approaches of Bechhofer and Gupta. Then the distribution of the subset size is given. The expected subset size and also the variance are presented. The expected subset size can be considered as a characteristic measure for a subset selection procedure, so it is important to study it. Also the variance, as a measure of dispersion, is of interest for a subset selection procedure. Finally, in the last section some generalisations and modifications are indicated.

SUBSET SELECTION

The subset selection approach, suggested by Gupta (1956,1965), is dealing with the selection of a subset from the k populations considered, in order to include the best population with a certain confidence. The size of the subset will be stochastic of nature. In Gupta and Panchapakesan (1985) a lot of references can be found.

Given are k ($\geq$2) independent random variables $X_1,\ldots,X_k$, which may be sample means with (for simplicity) common sample size, associated with the populations $\pi_1,\ldots,\pi_k$. These k random variables characterise the populations $\pi_1,\ldots,\pi_k$. We assume that the distributions of these random variables differ only in their location parameter and have continuous cumulative distribution functions $F(x\text{-}\theta_1),\ldots,F(x\text{-}\theta_k)$, respectively, and probability densities $f(x\text{-}\theta_1),\ldots,f(x\text{-}\theta_k)$, respectively.

We are interested in choosing the best population, that is the population with the largest value of θ. The parameter space Ω is defined as $\Omega = \{\theta\text{: } \theta = (\theta_1,\ldots,\theta_k)\}$. The ranked location parameters $\theta_1,\ldots,\theta_k$ are denoted by $\theta_{[1]}\leq\ldots\leq\theta_{[k]}$. If there are more than one contenders for the highest rank, we suppose that one of these is appropriately tagged. Subset selection has as its goal to select a non-empty subset of the k populations in order to include the best population with a certain confidence. The size of the subset is a random variable and depends among other things on the sample sizes (in the case that the random variables are sample means), the distribution functions, and the distances between the location parameters. The statistical control extends over only one stage, namely selection of a subset. Hence, such a procedure can be called a one-stage subset selection procedure. The statistical control of two-stage and multi-stage procedures, or even sequential procedures, are not considered in this paper. For all these generalisations we refer to the literature.

The subset selection procedure of Gupta is defined by the following selection rule

R: Select π_i into the subset if and only if $x_i \geq x_{\max} - d,$ (1)

where x_i is the observed value of X_i $(i=1,\ldots,k)$ and $x_{\max}$ is the observed value of $X_{\max} = \max_{1 \leq i \leq k} X_i$. From the selection rule **R** it follows that the obtained subset will be non-empty. The selection constant d (≥ 0) has to be chosen such that the probability is at least P^* (with $k^{-1} < P^* < 1$) that the selected subset contains the population with the largest value $\theta_{[k]}$ of θ. A correct selection CS means selection of any subset which includes the best one. The probability of CS is equal to

$$P(CS) = P(X_{(k)} \geq X_{\max} - d), \tag{2}$$

where $X_{(k)}$ is the unknown random variable which is associated with $\theta_{[k]}$. Now we can write (see Gupta, 1965)

$$P(CS) = \int_{-\infty}^{\infty} f(t) \prod_{i=1}^{k-1} F(t + d + \theta_{[k]} - \theta_{[i]}) dt, \tag{3}$$

where $F(.)$ and $f(.)$ are the distribution function and the density, respectively, of X_i-θ_i $(i=1,\ldots,k)$ and

$$\inf_{\Omega} P(CS) = \int_{-\infty}^{\infty} f(t) F^{k-1}(t+d) dt, \tag{4}$$

which is attained, in this location parameter case, for the Least Favourable Configuration $\theta_{[1]} = \theta_{[k]}$, as a consequence of the tagging. So the smallest value of d has to be chosen for which

$$\int_{-\infty}^{\infty} f(t) F^{k-1}(t+d) dt = P^*, \tag{5}$$

to be sure that $P(CS) \geq P^*$ for all configurations $\theta_1, \ldots, \theta_k$.

The expected size E(S) of the selected subset is equal to

$$\mathrm{E}(S) = \sum_{i=1}^{k} \int_{-\infty}^{\infty} \prod_{\substack{j=1 \\ j \neq i}}^{k} F(x + d + \theta_{[i]} - \theta_{[j]}) dF(x). \tag{6}$$

If the density $f(x-\theta)$ has monotone likelihood ratio in x then

$$\max_{\Omega} \mathrm{E}(S) = k \int_{-\infty}^{\infty} F^{k-1}(x+d) dF(x) = kP^*. \tag{7}$$

For the case of *normal* populations with a common known variance σ^2 and where the k random variables $X_1,\ldots,X_k$ are the sample means of k independent samples of common n independent observations, the selection rule can be written as

$$\mathbf{R}\text{: Select } \pi_i \text{ into the subset if and only if} \quad \overline{X}_i \geq \max_{1\leq j\leq k} \overline{X}_j - d^* \frac{\sigma}{\sqrt{n}}, \qquad (8)$$

where $\overline{X}_i$ is the sample mean of sample i and the selection constant $d^* > 0$ must be determined such that the P^*-condition is satisfied. Values of d^* can be found in Gibbons et al. (1977) for different values of P^* and k.

SOME PROPERTIES OF SUBSET SELECTION

The subset selection approach has certain nice aspects in practice. We assume that the k random variables are based on sample means. No minimal sample size has to be determined before the experiment can be started. The method of Gupta can be executed after the experiment has been realised, because no minimal sample size is needed. The consequence is that the subset size is a stochastic variable. For large sample sizes the subset size may be small, but for small sample sizes the subset size may be large, even k is possible. The size of the subset or even the expected subset size is an interesting characteristic for the experimenter when he is trying to find the best population. A small subset would mean that either the location parameters are not close together or that the sample sizes are large, or both. A possible practical objection to subset selection is of the type ”large subsets are sometimes the result”. But that is the toll one has to pay for a well-defined strong probability requirement. The performance of a selection procedure can be improved by increasing the common sample sizes or by formulating a weaker probability requirement.

In general the experimenter will have as its ultimate goal to select the best population. But, certainly if the number of populations is large, the subset selection approach makes it possible to eliminate populations and to proceed with a smaller number of “potentially interesting” populations. In other words the subset selection procedure can be used as a screening methodology. The LFC is in general easy to handle. So to see, the subset selection is a flexible approach. In the case of normal means with σ^2 unknown a single-stage method can be used, whereas for the indifference zone approach of Bechhofer a two-stage procedure is needed. To be complete, the method of Bechhofer has the advantage to select only one population, requiring a minimal sample size. This means that in the design phase this selection methodology, based on the indifference zone approach, is important.

THE PROBABILITY DISTRIBUTION OF THE SUBSET SIZE *S*

The size of the subset S is a integer random variable with possible outcomes $1,\ldots,k$, and can be considered as a characteristic and crucial quantity for subset selection. A relative large subset size means, apart from random fluctuations and small sample sizes, that the location parameters are (too) close together in comparison with the variation of the populations. For instance, the expected subset size ES can be used as an interesting characteristic of a subset selection procedure. Therefore, it is of interest to determine ES for the selection procedure.

A general expression for the distribution of the subset size S for the location parameter case is given in Van der Laan (1996). From this general expression the next result can be derived for the Least Favourable Configuration LFC $\theta_{[1]} = \theta_{[k]}$. For s=1,...,k one has

$$P_{LFC}(S = s) = s\binom{k}{s}\int_{-\infty}^{\infty}[F(x) - F(x-d)]^{s-1} F^{k-s}(x-d)dF(x). \tag{9}$$

A proof is given in Van der Laan (1996). Some special distributions, like uniform, exponential and logistic, are considered in that paper.

The expectation and variance of the subset size S under the LFC are given in the next formulas.

For s=1,...,k one has

$$\mathrm{E}_{\mathrm{LFC}}(S) = k\int_{-\infty}^{\infty} F^{k-1}(x+d)dF(x), \tag{10}$$

this well known result has also been derived in a different way by Gupta (1965), and

$$\begin{aligned}\mathrm{var}_{\mathrm{LFC}}(S) = k\int_{-\infty}^{\infty} F^{k-1}(x+d)dF(x)\left[2k-1-\int_{-\infty}^{\infty} F^{k-1}(x+d)dF(x)\right]- \\ -2k(k-1)\int_{-\infty}^{\infty} F^{k-2}(x+d)F(x)dF(x)\end{aligned} \tag{11}$$

A proof is given in Van der Laan (1996). In this paper some numerical results concerning expectation and standard deviation of the subset size S are presented for normal populations. Also explicit expressions have been determined for uniform, exponential and logistic populations.

SOME CONCLUDING REMARKS

In the literature many generalisations and modifications of the subset selection method can be found. We mention, e.g., Mahamumulu (1967). In this paper a generalised goal is defined as follows: to select a subset of size s from k populations so that the selected subset contains at least c of the t best populations with a certain confidence. As an example, for the case of normal means the confidence requirement is to determine the minimal sample size such that the $P(CS)$ is larger than or equal to a certain bound P^*, whenever $\theta_{[k-t+1]} - \theta_{[k-t]} \geq \delta^*$. Santner (1975) and Gupta and Santner (1973) proposed so called restricted subset size procedures. The restriction is to select a subset of random size, subject to the condition that this size does not exceed a maximum. Desu (1970) and Carroll, Gupta, and Huang (1975) considered the goal to exclude any population that is sufficiently inferior to the best population. Verheijen et al. (1997) introduced a preference threshold procedure. This procedure is a combination of indifference zone selection and subset selection, the two standard approaches of statistical selection. In Verheijen et al. (1999) the robustness of the preference threshold procedure is investigated with respect to deviations from the normality assumption.

Multivariate and nonparametric methods have also been considered in the literature. In short, too many different cases (e.g., scale parameters, various distributions), modifications, generalisations are considered in the literature to describe them here. We refer to the literature.

REFERENCES

Bechhofer, R.E. (1954). A single–sample multiple decision procedure for ranking means of normal populations with known variances. *Annals of Mathematical Statistics* **25**, 16-39.

Desu, M. M. (1970). A selection problem. *Annals of Mathematical Statistics* **41**, 1596-1603.

Dudewicz, E. J. (1976). *Introduction to Statistics and Probability.* Holt, Rinehart and Winston, New York.

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

Carroll, R. J., Gupta, S. S., and Huang, D. Y. (1975). On selection procedures for the *t* best populations and some related problems. Communications of Statistics **4**, 987-1008.

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

Gibbons, J. D., Olkin, I., and Sobel, M. (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. Dissertation (and Mimeograph Series No.150), Institute of Statistics, University of North Carolina, Chapel Hill.

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

Gupta, S. S. (1977). Ranking and Selection. *Special Issue, Communications of Statistics-Theory and Methods* **A6**, 993-1001.

Gupta, S.S. and Panchapakesan, S. (1979). *Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations.* John Wiley & Sons, Inc., New York.

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 Santner, T. J. (1973). On selection and ranking procedures- a restricted subset selection rule. *Proceedings of the 39th Session of the International Statistical Institute*, Vol. **45**, book **1**, 409-417.

Mahamunulu, D. M. (1967). Some fixed–sample ranking and selection problems. *Annals of Mathematical Statistics* **38**, 1079-1091.

Rizvi, M. H. (Ed.) (1985). Modern Statistical Selection, Part I - Robert E. Bechhofer, Shanti S. Gupta, and Milton Sobel, *Proceedings of the Conference "Statistical ranking and Selection - Three Decades of Development"*, Univ. of California at Santa Barbara, December 1984. *American Journal of Mathematical and Management Sciences*, Vol. **5**, NOS. **3&4**.

Rizvi, M. H. (Ed.) (1986). Modern Statistical Selection, Part II - Robert E. Bechhofer, Shanti S. Gupta, and Milton Sobel, *Proceedings of the Conference "Statistical ranking and Selection - Three Decades of Development"*, Univ. of California at Santa Barbara, December 1984. *American Journal of Mathematical and Management Sciences*, Vol. **6**, NOS. **1&2**.

Santner, T. J. (1975). A restricted subset selection approach to ranking and selection problems. *Annals of Statistics* **3**, 334-349.

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

Van der Laan, P. and Verdooren, L. R. (1989). Selection of Populations: An Overview and Some Recent Results. *Biometrical Journal* **31**, 383-420.

Van der Laan, P. (1996). Distributional and efficiency results for subset selection. *Journal of Statistical Planning and Inference* **54**, 159-174.

Verheijen, J. H. M., Coolen, F. P. A., and Van der Laan, P. (1997). Combining two classical approaches for statistical selection. Communications of Statistics-Theory and Methods **26**, 1291-1312.

Verheijen, J. H. M., Coolen, F. P. A., and Van der Laan, P. (1999). Preference threshold procedure for selection: Robustness against deviations from normality. *Biuletyn Oceny Odmian* (*Cultivar Testing Bulletin*) **30**, 151-168.