ON SELECTION AND RANKING PROCEDURES*

by

Shanti S. Gupta

Department of Statistics

Division of Mathematical Sciences

Mimeograph Series No. 119

August, 1967

---

* This research was supported by the Aerospace Research Laboratories, Contract #AF 33(615)67C1244. Reproduction in whole or in part is permitted for any purpose of the United States Government.

ON SELECTION AND RANKING PROCEDURES *

Shanti S. Gupta, Department of Statistics
Purdue University, Lafayette, Indiana

1. Introduction

In many situations one encounters $k(k \geq 2)$ populations which may be categories, varieties, processes, candidates etc. Suppose for each population one can observe a random variable whose distribution depends upon an unknown parameter θ_i. This parameter may be the mean, the variance, some quantile or a function of these quantities. Usually, in the selection and ranking problems, populations with large (small) values of the parameters are considered desirable and, accordingly, we define the population with the largest (smallest) of the unknown values of the k parameters to be the best. In many situations, the experimenter is interested in selecting a subset, selecting as far as possible the best ones. In general, selection and ranking (multiple decision) rules may be defined which select a subset of a fixed or random size. One may wish to select only a single population and guarantee with probability P^* that the selected population is the best provided some other condition on the parameters is satisfied. Bechhofer (1954) considered the problem for the means of k normal populations with a common and known variance. In his formulation Bechhofer (1954) is interested in guaranteeing the probability of selecting the best one to be a specified number P^* whenever the standardized difference δ between the largest mean and the second largest mean is at least equal to a specified value δ^*. This "indifference zone" approach thus requires that the experimenter specify two quantities P^* and δ^*. The decision rule is to choose the population with the largest observed sample mean $\bar{X}_{max}$. The problem of determining the common sample size n to guarantee the probability P^* under the indifference zone $\delta \geq \delta^*$ was solved by Bechhofer (1954) in the above paper. Basically, similar problems with obvious modifications in the definition of indifference zone, etc., have been solved for other distributions by several authors.

A second and different (present) formulation requires that the probability of including the best population in the selected subset be guaranteed to be a specified number P^* regardless of the possible configurations of the parameters. Instead of two specified quantities P^* and δ^*, we now specify only one quantity P^*. In this formulation the size of the selected subset is a random variable and so are the ranks of the populations in the selected subset. Assuming $\theta_{[i]}$ to be the ith ranked parameter with rank i, where $i=k$ corresponds to the best population, we are then guaranteeing that the population with the maximal rank is included in the selected subset with probability at least equal to P^* regardless of what the unknown θ_i's may be. A procedure for achieving this objective can be carried out for all n where n denotes the common number of observations from each of the k populations. The expected size, the expected minimal rank and the expected sum of ranks of the populations selected in the subset define various (related) performance characteristics of the procedure or procedures.

It should be pointed out that there is need to consider multiple decision problems. In many situations the experimenter is really interested in finding the best treatment or population or a group containing the best treatment or population. The tests of homogeneity are inadequate for such problems. The formulation in terms of ranking and selection is more realistic and meaningful in such situations than that of tests for homogeneity or equality of parameter values.

2. Formal Statement for the Subset Selection Problem and Selection Rule for Two Cases

The selection of any subset containing the best population is called a correct selection and will be denoted by (CS). If the selection proceeds according to some rule R, then the subset selected should contain the best population with a specified probability P^* i.e.

$$P\{CS|R\} \geq P^* \tag{2.1}$$

whatever the unknown values of θ_i's may be. Moreover, the selection rule R should possess certain desirable properties.

* The writing of this paper was supported in part by Contract AF 33 (615) 67-C-1244, with the Aerospace Research Laboratory.

In the general situation we are given an observation x_i from the population π_i which has density $f_{\theta_i}(x)$. Let the ordered θ_i's be denoted by

$$\theta_{[1]} \leq \theta_{[2]} \leq \cdots \leq \theta_{[k]}. \tag{2.2}$$

The correct pairing of the ordered θ_i's and the observed x_i's is not known. Based on the observed values $x_1, x_2, \ldots, x_k$ we like to define a procedure which satisfies the basic probability requirement and is good in some sense. Now we discuss two cases separately and describe the two procedures proposed by Gupta (1965a).

Case (i) $f_\theta(x) = f(x-\theta)$ i.e. θ is a translation parameter. For this case the following rule R has been proposed.

Select π_i iff

$$x_i \geq x_{max} - d \tag{2.3}$$

where $x_{max} = \max(x_1, x_2, \ldots, x_k)$ and the constant $d \geq 0$ is the minimal value required to satisfy the basic probability requirement. The rule R selects a non-empty subset of random size and is translation invariant. In practice, $1/k < P^* < 1$ so that $0 < d < \infty$.

Case (ii) $f_\theta(x) = 1/\theta\, f(x/\theta)$ i.e. θ is scale parameter. In this case the rule given in case (i) is modified as follows:

Select π_i iff

$$x_i \geq c\, x_{max} \tag{2.4}$$

where $0 < c < 1$ and c is again chosen to satisfy the P^* condition. It should be noted that the above rule is scale invariant. In practice, $1/k < P^* < 1$ so that $1 > c > 0$.

3. Properties of the Selection Rules

In this section, we discuss some properties and performance characteristics of the selection rules given in Section 2. These properties together with numerical evaluations of the performance of these rules provide justification for their use. The discussion here is mainly concerned with the translation case; the scale parameter case is entirely similar. The following properties are only stated here. For further discussion, reference should be made to Gupta (1965) and Gupta and Studden (1966).

(a) The probability of a Correct Selection

The probability of a correct selection using the rule R is an increasing function of each of the differences $\theta_{[k]} - \theta_{[j]}$, $j=1,2,\ldots,k-1$. The same holds for the scale parameter case with differences replaced by the ratios $\theta_{[k]}/\theta_{[j]}$, $j=1,2,\ldots,k-1$.

(b) Monotonicity and Unbiasedness

For the rule R, the probability of selecting the population corresponding to $\theta_{[i]}$ is greater than or equal to the probability of selecting the population corresponding to $\theta_{[j]}$ provided $\theta_{[i]} \geq \theta_{[j]}$. Thus, the rule R is unbiased in the sense that the probability of rejecting any population not having the largest parameter θ is not less than the probability of rejecting the best population.

(c) Expected Subset Size

Subject to the basic P^* requirement the procedure R satisfies the condition that the expected size of the selected subset is $\leq kP^*$ for all choices of $\theta_1, \theta_2, \ldots, \theta_k$.

(d) Minimax Property

If we impose the invariance or symmetry condition which says that if the ith and jth observations are interchanged, then we select the jth population with the same probability $\varphi_i(x_1, \ldots, x_k)$ where φ_i represents the probability of selecting the ith population. More specifically, we shall require that

$$(3.1) \qquad \varphi_i(x_1,\ldots,x_i,\ldots,x_j,\ldots,x_k) = \varphi_j(x_1,\ldots,x_j,\ldots,x_i,\ldots,x_k)$$

for all i and j.

Then it can be shown that the rule R is minimax in the sense that it minimizes $\sup_{\Omega} E_{\underline{\theta}}(S|R')$ over the class of rules satisfying the basic P^* condition and the above invariance condition. Here Ω is the whole parameter space.

4. Selection of Normal Populations

First we discuss the selection problem for the means of normal populations. It is assumed that the k normal populations $\pi_1,\pi_2,\ldots,\pi_k$ have unknown means $\mu_1,\mu_2,\ldots,\mu_k$ and a common variance σ^2. Suppose we are given n observations $x_{ij}(j=1,2,\ldots,n)$ from the population $\pi_i(i=1,2,\ldots,k)$. Let $\bar{x}_i$ be the sample mean from π_i and let s_ν^2 be the usual pooled estimate of σ^2 based on $\nu = k(n-1)$ degrees of freedom. Then the selection procedure is:

Select the population π_i iff

$$(4.1) \qquad \bar{x}_i \geq \bar{x}_{max} - D\, s_\nu/\sqrt{n}$$

where the constant $D = D(k,\nu,P^*)$ chosen to satisfy the P^* condition is positive if $1/k < P^*$. This constant D is determined by

$$(4.2) \qquad \int_0^\infty \int_{-\infty}^\infty \Phi^{k-1}(u+Dy)\, \varphi(u)\, q_\nu(y)\, du\, dy = P^*$$

where $q_\nu(y)$ is the density function of $\chi_\nu/\sqrt{\nu}$ and Φ and φ refer to the cumulative distribution function and the density of the standard normal random variable. The solutions of the above equation for $\nu = \infty$ (σ known) in the form of values of $D\sqrt{2}$ is in Gupta (1963a) as Table I for $k = 2(1)51$ and $P^* = .75, .90, .95, .975$ and $.99$. For selected values of P^*, k and ν the values of D are tabulated by Gupta and Sobel (1957). A brief table of the values of D is excerpted from Gupta (1963a) for the σ-known case and follows.

Table 1. Values of D for the Rule (4.1) when σ is known

P^* \ k	2	5	10	15	20
.75	.95	1.85	2.26	2.47	2.60
.90	1.81	2.60	2.98	3.17	3.30
.95	2.33	3.06	3.42	3.60	3.72

Evaluation of the efficiency of the above selection procedure have been made by Gupta (1965a) and Deely and Gupta (1966) for the case where σ is known. In this case we can assume that the common value of σ is unity without any loss of generality. Since the expected size or proportion in the selected subset, the probability of a correct selection depend on the unknown means, we assume the simple configuration known as the slippage configuration in which the means are $\mu,\mu,\ldots,\mu,\mu+\delta$. Then for selected values of $\delta\sqrt{n}$, k and P^*, we give below a brief excerpt from Deely and Gupta (1966) of the values of the actual probability of a correct selection and the expected proportion i.e. $1/k$ times the expected size.

Table 2. This table gives the probability of a correct selection (top) and the expected proportion (bottom) in the selected subset for the slippage configuration.

P* \ k	2			5			10		
	$\delta\sqrt{n}$ = 1.00	2.00	5.00	1.00	2.00	5.00	1.00	2.00	5.00
.75	.917	.982	1.000	.930	.988	1.000	.935	.990	1.000
	.702	.606	.501	.695	.537	.210	.709	.564	.124
.90	.977	.996	1.000	.981	.998	1.000	.983	.998	1.000
	.847	.722	.506	.857	.715	.236	.871	.753	.169
.95	.991	.999	1.000	.993	.999	1.000	.993	.999	1.000
	.908	.795	.515	.923	.816	.247	.930	.841	.218

Selection for Small Variance--The selection of normal populations into a subset to contain the population with the smallest variance is given in Gupta and Sobel (1962). The rule based on the observed sample variances $s_1^2, s_2^2, \ldots, s_k^2$ is as follows:

Select the population π_i iff

$$s_i^2 \leq s_{min}^2/c \tag{4.3}$$

where $c = c(k,\nu,P^*)$ depends on k, P^* and ν = the common degrees of freedom for each s_i^2 and lies between 0 and 1. A brief table of c-values as excerpted from Gupta and Sobel (1962) is given below.

Table 3. This table gives the necessary c-values required for the variance selection procedure (4.3).

	P* \ k	2	3	5	10
ν = 2	.75	.33	.17	.08	.03
4		.48	.32	.21	.13
10		.64	.51	.41	.32
20		.74	.62	.54	.46
40		.81	.72	.65	.59
ν = 2	.90	.11	.06	.03	.01
4		.24	.16	.11	.07
10		.43	.35	.28	.23
20		.56	.48	.42	.36
40		.66	.60	.55	.50
ν = 2	.95	.05	.03	.01	.01
4		.16	.11	.07	.05
10		.34	.27	.22	.18
20		.47	.41	.36	.32
40		.59	.54	.49	.45

5. Selection Procedures for other Distributions

Using the same formulation selection procedures for other distributions have been computed and their efficiency has also been evaluated. A procedure for subset selection containing the largest of the scale parameters of the gamma population is proposed and evaluated in Gupta (1963b). This is similar to the procedure for ranking the variances of the normal populations. For the binomial distributions, the problem is solved by Gupta and Sobel (1960) and Gupta (1963c). More recently, a paper by Gupta and Nagel (1966) solves the problem for the multinomial distribution. It should be pointed out

that for the multinomial distribution, the subset selection problems for the cell with the largest probability and the cell with the smallest probability are not equivalent so that two different procedures have been worked out in the paper mentioned above. Selection and ranking problems for multivariate populations in terms of $\mu_i' \Sigma_i^{-1} \mu_i$ (μ_i is the mean vector and Σ_i is the covariance matrix of the ith population) has been done in two papers by Gupta (1965b) and Gupta and Studden (1965).

6. Examples

Example 1. Suppose we are given a one-way classification of $k = 5$ normal populations. Let the observed sample means based on 16 observations each be 7.50, 5.21, 10.80, 13.25, 15.00. Let the common variance $\sigma^2 = 25$ and let us assume that the experimenter would like a correct selection probability of $P^* = .75$. Then the rule (4.1) is to be used with s_ν replaced by σ and the constant D from Table I is 1.85. Since $\bar{x}_{max} = 15.00$, all the populations with observed sample means in the interval $[15-(1.85)5/4, 15]$ will be selected i.e. the populations with means ≥ 12.69 are to be selected. Thus the selected subset consists of the two populations with observed sample means 13.25 and 15.00. After the selection has been made it can be stated with confidence P^* that the best population is one of the two that have been selected.

Example 2. Based on $\nu = 10$ degrees of freedom the observed sample variances s_i^2 $(i=1,2,3)$ of 3 normal populations are 2.50, 4.58 and 5.75. Corresponding to $P^*=.75$, the selected subset contains of all populations with $s_i^2 \leq (2.5)/(.51) = 4.9$ (from Table 3 $c = .51$). Thus, in this example, the population with the two smallest observed variances 2.50 and 4.58 are selected.

REFERENCES

1. Bechhofer, R.E. 1954. A single-sample multiple decision procedure for ranking means of normal populations with known variances. Ann. Math. Statist. 25, 16-29.

2. Deely, J.J. 1965. Multiple decision procedures from an empirical Bayes approach. Ph.D. Thesis, Dept. of Statistics, Mimeograph Series No. 45, Purdue University.

3. Deely, J.J. and Gupta, S.S. 1966. On the properties of subset selection procedures. Dept. of Statistics, Mimeograph Series No. 49, Purdue University. Submitted.

4. Gupta, S.S. and Sobel, M. 1957. On a statistic which arises in selection and ranking problems. Ann. Math. Statist. 28, 957-967.

5. Gupta, S.S. and Sobel, M. 1960. Selecting a subset containing the best of several binomial populations. Contributions to Probability and Statistics. Ch. XX. Stanford University Press.

6. Gupta, S.S. and Sobel, M. 1962a. On selecting a subset containing the population with the smallest variance. Biometrika 49, 495-507.

7. Gupta, S.S. and Sobel, M. 1962b. On the smallest of several correlated F statistics. Biometrika 49, 509-523.

8. Gupta, S.S. 1963a. Probability integrals of the multivariate normal and multivariate t. Ann. Math. Statist. 34, 792-828.

9. Gupta, S.S. 1963b. On a selection and ranking procedure for gamma populations. Ann. Inst. Statist. Math. Tokyo 14, 199-216.

10. Gupta, S.S. 1963c. Selection and ranking procedures and order statistics for the binomial distribution. Classical and Contagious Discrete Distributions, Proceedings of the International Symposium, Montreal, 1963.

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

12. Gupta, S.S. and Nagel, K. 1966. On selection and ranking procedures and order statistics from the multinomial distribution. Department of Statistics, Mimeograph Series No. 77, Purdue University.

13. Gupta, S.S. and Studden, W.J. 1966. Some aspects of selection and ranking procedures with applications. Dept. of Statistics, Mimeograph Series No. 81, Purdue University.

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*

Purdue University

2a. REPORT SECURITY CLASSIFICATION

Unclassified

2b. GROUP

3. REPORT TITLE

On Selection and Ranking Procedures

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Technical Note

5. AUTHOR(S) *(Last name, first name, initial)*

Shanti S. Gupta

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
	5	13

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
AF 33(615)67C1244	Mimeo #119
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
d.	

10. AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Air Force Office of Scientific Research Aerospace Research Laboratories Wright Patterson AFB, Ohio

13. ABSTRACT

This paper describes some parametric procedures for selecting a subset containing the best population. Properties of selection rules for the location and scale parameter cases are mentioned. Brief tables giving the constants that define the selection procedures for the means and variances of normal populations are given. Also a brief table relevant to efficiency of the means procedure is given.

DD FORM 1 JAN 64 1473

Unclassified

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Selection and Ranking Subset Selection Expected Proportion Correct Selection						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through

_____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through

_____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through

_____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS), (S), (C),* or *(U)*

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, roles, and weights is optional.