OPTIMAL SAMPLING IN SELECTION PROBLEMS*

by

Shanti S. Gupta
Purdue University

Technical Report #83-21

Department of Statistics
Purdue University

June 1983

*This research was supported by the Office of Naval Research Contract N00014-75-C-0455 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

OPTIMAL SAMPLING IN SELECTION PROBLEMS*

by

Shanti S. Gupta
Purdue University

A selection procedure typically consists of three ingredients: (1) a sampling rule, (2) a stopping rule, and (3) a decision rule, though these components are not usually explicitly so labeled. The problem of optimal sampling arises in different ways depending on the context of the problem at hand. Broadly speaking, the problem of optimal (or optimum) sampling arises because of the need for balancing between the cost of sampling and the cost of making a wrong decision. Obviously, increasing the amount of sampling increases the former cost while decreasing the latter.

1. Indifference Zone Formulation

Suppose we have k independent populations $\pi_1, \pi_2, \ldots, \pi_k$, where the CDF of π_i is $F(x; \theta_i)$, where the parameter θ_i has an unknown value belonging to an interval Θ on the real line. Our goal is to select the population associated with the largest θ_i which is called the best population. In the Indifference Zone Formulation of Bechhofer [2], it is required that the selection rule guarantees with a probability at least equal to $P^*(1/k < P^* < 1)$ that the best population will be chosen whenever the true parametric configuration $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ lies in a subset of the parametric space Ω_Δ characterizing the property that the distance between the best and the next best populations is at least Δ. The subset Ω_Δ is called the Preference Zone.

*This research was supported by the Office of Naval Research Contract N00014-75-C-0455 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

The constants P* and Δ are specified in advance by the experimenter. The probability guarantee requirement is referred to as the P*-requirement.

Now, let us consider k independent normal populations $\pi_1, \pi_2, \ldots, \pi_k$ with unknown means $\mu_1, \mu_2, \ldots, \mu_k$, respectively, and common known variance σ^2. Based on samples of size n from each population, the single-stage procedure of Bechhofer [2] for selecting the population with the largest μ_i selects the population that yields the largest sample mean. Here the preference zone is defined by the relation $\mu_{[k]} - \mu_{[k-1]} \geq \Delta$, where $\mu_{[1]} \leq \ldots \leq \mu_{[k]}$ denote the ordered μ_i. The optimum sampling problem in this case is to determine the minimum sample size n subject to the P*-requirement. The optimum value of n is given by the smallest integer n for which

$$\int_{-\infty}^{\infty} \Phi^{k-1}\left(x + \frac{\sqrt{n}\Delta}{\sigma}\right) \varphi(x)dx \geq P^*$$

where Φ and φ denote the CDF and the density function of a standard normal random variable.

Suppose that these normal distributions have unknown and possibly unequal variances. In this case, no single-stage procedure exits. Two-stage procedures have been studied in this situation by Bechhofer, Dunnett, and Sobel [4], and Dudewicz and Dalal [9]. One may take a sample of size n_0 from each population at the first stage

and on the basis of the information obtained from these samples, determine the sizes of additional samples to be taken from these populations. The selection rule is based on the total samples from all the populations. Even when the variances are known, one may use a two-stage procedure in which the first stage involves selection of a nonempty subset of random size with possible values 1,2,..., and k. If the first stage results in a subset of size larger than 1, then a second stage ensues with additional samples from those populations that still remain under consideration. Such procedures have been considered by Alam [1], Tamhane and Bechhofer [20], [21] and by Gupta and Miescke [15] with some modifications. A problem of optimum sampling in these cases is to determine the optimal combination of the sample sizes in the two stages. This can be done, for example (Tamhane and Bechhofer [20]), by minimizing the maximum of the expected total sample size for the experiment over all parametric configurations subject to the P*-requirement.

2. Minimax, Gamma Minimax and Bayes Techniques

Consider again k normal populations $\pi_1, \pi_2, \ldots, \pi_k$ with unknown means $\mu_1, \mu_2, \ldots, \mu_k$ and common known variance σ^2. If the selection procedure is to take samples of size n from these populations and choose the population that yields the largest sample mean, one can consider a loss function

$L = c_1 n + \sum_{i=1}^{k} c_2(\mu_{[k]}-\mu_i)I_i$, where c_1 is the sampling cost per observation, c_2 is a positive constant, and $I_i = 1$, if π_i is selected, and = 0 otherwise. Optimum n can be obtained by minimizing the integrated risk assuming (known) prior distributions for μ_i's; see Dunnett [10]. One may also determine the optimum n by minimizing the maximum expected loss over all parametric configurations. However, the expected loss in our case is unbounded above and we can find a minimax solution if we have prior information regarding the bounds on the differences $\mu_{[k]}-\mu_{[i]}$, $i = 1,\ldots,k-1$.

Suppose we take a sample of size n_1 from each of k normal populations with unknown means $\mu_1,\mu_2,\ldots,\mu_k$, and common known variance σ^2. For a fixed t, $1 \leq t \leq k-1$, we discard the populations that produced the t smallest sample means and take an additional sample of size n_2 from each of the remaining k-t populations. We select as the best the population that entered the second stage and produced the largest sample mean based on all n_1+n_2 observations. Given that the total sample size $T = kn_1+(k-t)n_2$ is a constant, the problem is to determine the optimum allocation of (n_1,n_2) by minimizing the maximum expected loss, where the loss is $L = c_1 T + c_2 \sum_{i=1}^{k}(\mu_{[k]}-\mu_i)I_i$ as defined

earlier. For details see Sommerville [19], and Fairweather [11].

In these problems, we can also take the gamma-minimax approach and minimize the maximum expected risk over a specified class of prior distributions for the parameters μ_i; see Gupta and Huang [14].

3. Comparison with a Control

An optimal sampling problem can be, as we have seen, an optimal allocation problem. Such allocation problems are also meaningful when we compare several treatments with a control. Let $\pi_1,\pi_2,\ldots,\pi_k$ be k independent normal populations representing the experimental treatments and let π_0 be the control which is also a normal population. Let π_i have unknown mean μ_i and known variance σ_i^2, $i = 0,1,\ldots,k$. A multiple comparisons approach is to obtain one-and two-sided simultaneous confidence intervals for, say, $\mu_i-\mu_0$, $i = 1,2,\ldots,k$. If n_i is the size of the sample from π_i, $i = 0,1,\ldots,k$, such that $\sum_{i=0}^{k} n_i = N$, a fixed integer, then the problem is to determine the optimal allocation of the total sample size. The optimal allocation will depend, besides other known quantities, on a specified 'yardstick' associated with the width of the interval. For details of these problems see Bechhofer [3], Bechhofer and Nocturne [5], Bechhofer and Tamhane [6], and Bechhofer and Turnbull [7].

Instead of taking the above multiple comparisons approach, one can use the formulation of partitioning the set of k experimental populations into two sets one consisting of populations that are better than the control and the other consisting of the remaining (worse than the control). For a given total sample size, the problem is to determine the optimal allocation either by minimizing the expected number of populations misclassified or by maximizing the probability of a correct decision; for details see Sobel and Tong [18].

4. Subset Selection Approach

As before, consider k independent populations $\pi_1,\pi_2,\ldots,\pi_k$, where π_i is characterized by the CDF $F(x; \theta_i)$, $i = 1,\ldots,k$. In the subset selection approach, we are interested in selecting a nonempty subset of the k populations so that the selected subset will contain the population associated with the largest θ_i with a guaranteed minimum probability P*. The number of populations to be selected depends on the outcome of the experiment and is not fixed in advance as in the indifference zone approach.

Suppose we take a random sample of size n from each population. Let T_i, $i = 1,\ldots,k$, be suitably chosen statistics from these samples. In the case of location parameters, the procedure of Gupta [12], [13] selects π_i if and only if $T_i \geq T_{max}-D$, where $T_{max} = \max(T_1,\ldots,T_k)$

and $D \geq 0$ is to be chosen such that the P*-requirement is met. The constant D will depend on k, P*, and n. Unlike in the indifference zone approach, we can obtain a rule for any given n satisfying the P*-condition.

In the case of k normal populations with unknown means $\mu_1, \mu_2, \ldots, \mu_k$, and known common variance σ^2, the rule of Gupta [12] selects π_i if and only if $\bar{X}_i \geq \bar{X}_{max} - d\sigma/\sqrt{n}$, where $\bar{X}_i$ is the mean of a sample of size n from π_i, $i = 1,2,\ldots,k$. The constant d is given by the equation

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

The expected subset size, denoted by E(S), is given by

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

where $\mu_{[1]} \leq \mu_{[2]} \leq \ldots \leq \mu_{[k]}$ denote the ordered μ_i. We can define the optimum sample size as the minimum sample size for which the expected subset size or equivalently, the expected proportion of the populations selected does not exceed a specified bound when the true parametric configuration is of a specified type. Relevant tables are available in Gupta [13] for the equidistant configuration given by $\mu_{[i+1]} - \mu_{[i]} = \delta$, $i = 1,2,\ldots,k-1$, and in Deely and Gupta [8] for the slippage configuration given by $\mu_{[1]} = \ldots = \mu_{[k-1]} = \mu_{[k]} - \delta$.

If we use the restricted subset selection approach in which the size of the selected subset is random subject to a specified upper bound, then the P*-condition is met whenever the parametric configuration belongs to a preference zone as in the case of Bechhofer's formulation. In this case, the minimum sample size (assuming common sample size) can be determined in a similar way (Gupta and Santner [17]).

In our discussion so far, the optimal sampling related to optimal sample sizes or optimal allocation under a given sampling scheme such as single-stage, two-stage, etc. One can also seek the optimal sampling scheme by comparing single-stage, multi-stage and sequential procedures. Comparisons of different sampling schemes for several selection goals have been made and are available in the literature. In addition to the usual sampling schemes, inverse sampling rules with different stopping rules and comparisons involving vector-at-a-time sampling and Play-the-Winner sampling scheme have been studied in the case of clinical trials involving dichotomous data. References to these and other problems discussed can easily be obtained from Gupta and Panchapakesan [16].

REFERENCES

[1] Alam, K. (1970). Ann. Inst. Statist. Math., 22, 127-136.

[2] Bechhofer, R.E. (1954). Ann. Math. Statist., 25, 16-39. A pioneering paper introducing the indifference zone formulation.

[3] Bechhofer, R.E. (1969). In Multivariate Analysis-II, P. R. Krishnaiah, ed. Academic Press, New York, pp. 463-473.

[4] Bechhofer, R.E., Dunnett, C.W., and Sobel, M. (1954). Biometrika, 41, 170-176.

[5] Bechhofer, R.E. and Nocturne, D.J. (1972). Technometrics, 14, 423-436.

[6] Bechhofer, R. E. and Tamhane, A. C. (1983). Technometrics, 25, 87-95.

[7] Bechhofer, R.E. and Turnbull, B.W. (1971). In Statistical Decision Theory and Related Topics, S. S. Gupta and J. Yackel, eds. Academic Press, New York, pp. 41-78.

[8] Deely, J.J. and Gupta, S.S. (1968). Sankhyā Ser. A, 30, 37-50. The first paper to consider a Bayesian approach to subset selection.

[9] Dudewicz, E.J. and Dalal, S.R. (1975). Sankhyā Ser. B, 37, 28-78.

[10] Dunnett, C.W. J. Roy. Statist. Soc. Ser. B, 22, 1-40. This is followed by a discussion by several statisticians.

[11] Fairweather, W.R. (1968). Biometrika, 55, 411-418.

[12] Gupta, S.S. (1956). On a decision rule for a problem in ranking means. Mimeo. Ser. No. 150, Inst. of Statistics, University of North Carolina, Chapel Hill, North Carolina.

[13] Gupta, S.S. (1965). Technometrics, 7, 225-245. The first paper to present a general theory of subset selection.

[14] Gupta, S.S. and Huang, D.-Y. (1977). In The Theory and Applications of Reliability, C. P. Tsokos and I. N. Shimi, eds. Academic Press, New York, pp. 495-505.

[15] Gupta, S.S. and Miescke, K.-J. (1983). In Statistical Decision Theory and Related Topics-III, Vol. 1, S. S. Gupta and J. O. Berger, eds. Academic Press, New York, pp. 473-496.

[16] Gupta, S.S. and Panchapakesan, S. (1979). Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations. Wiley, New York. A comprehensive survey of all aspects of selection and ranking problems with an extensive bibliography.

[17] Gupta, S.S. and Santner, T.J. (1973). Proc. 39th Session Int. Statist. Inst., Vol. 45, Book 1, pp. 409-417.

[18] Sobel, M. and Tong, Y.L. (1971). Biometrika, 58, 171-181.

[19] Sommerville, P.N. (1954). Biometrika, 41, 420-429.

[20] Tamhane, A.C. and Bechhofer, R.E. (1977). Commun. Statist. Theor. Meth., A6, 1003-1033.

[21] Tamhane, A.C. and Bechhofer, R.E. (1979). Commun. Statist. Theor. Meth., A8, 337-358.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Technical Report #83-21	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): OPTIMAL SAMPLING IN SELECTION PROBLEMS	5. TYPE OF REPORT & PERIOD COVERED: Technical
	6. PERFORMING ORG. REPORT NUMBER: Technical Report #83-21
7. AUTHOR(s): Shanti S. Gupta	8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0455
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Purdue University, Department of Statistics, West Lafayette, IN 47907	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Washington, DC	12. REPORT DATE: June 1983
	13. NUMBER OF PAGES: 11
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release, distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Optimal sampling, indifference zone formulation, subset selection approach, single-stage, two-stage, comparison with control, restricted subset selection, minimax, gamma-minimax.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper describes several optimal sampling problems that arise in connection with selection and ranking procedures.

DD FORM 1 JAN 73 1473

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)