The Annals of Mathematical Statistics
1972, Vol. 43, No. 6, 1977-1980

SELECTION SAMPLE SIZE APPROXIMATIONS[1]

BY JOHN S. RAMBERG

The University of Iowa

Two conservative sample size approximations are given for the Bechhofer formulation of the problem of selecting the population with the largest mean, when the populations have a common known variance. A table of numerical comparisons of these approximations with the exact sample size is included. In addition, both of these results are applied to the problem of selecting the factor with the largest mean from a single multivariate normal population with known covariance matrix.

1. Summary. Two sample size approximations are given for the selection procedure problem of Bechhofer (1954). One follows from the Bonferroni inequality, the other from the Slepian inequality. Both approximations are conservative in that they are larger than the sample size given by the exact theory and hence guarantee the probability requirement. Other approximations have already been given by Robbins and Siegmund (1967), Bechhofer, Kiefer and Sobel (1968) and Dudewicz (1969). However, whereas the previous approximations were for limiting cases, i.e., k(the number of populations) $\rightarrow \infty$ or P^*(the probability requirement) $\rightarrow 1$, the two approximations given in this paper hold for general k and P^*. A table of numerical comparisons of these approximations with the exact sample size is included to illustrate their practicality for certain ranges of P^*. A recent paper by Dudewicz and Zaino (1971) contains numerical comparisons of the above approximations.

Both inequalities are also used for the problem of selecting the factor with the largest mean from a single multivariate normal population with known covariance matrix. The Slepian inequality, when it can be used, is stronger than the Bonferroni inequality (and hence provides a better approximation); but the Bonferroni inequality can be used for general covariance matrices.

2. The problem. Given k populations $\pi_1, \cdots, \pi_k$ with means $\mu_1, \cdots, \mu_k$, we want to select the population associated with $\mu_{[k]}$, where $\mu_{[1]} \leqq \cdots \leqq \mu_{[k]}$ denote the ranked μ_i's. We assume that the observations from each of the populations are independent and normally distributed with common known variance σ^2.

The single-stage *means procedure* suggested by Bechhofer (1954) is: Take N (to be determined) independent observations X_{ij} on each of the k populations, where X_{ij} denotes the jth observation on the ith population, and select the population associated with $\max(\bar{X}_1, \cdots, \bar{X}_k)$, where $\bar{X}_i = \sum_{j=1}^{N} X_{ij}/N$, as the population associated with $\mu_{[k]}$. If $\{\delta^*, P^*\}$ $(0 < \delta^* < \infty, 1/k < P^* < 1)$ are

Received July 22, 1970; revised March 3, 1972.

[1] This reseach was supported in part by the U.S. Army Research Office-Durham under Contract DA-31-124-ARO-D-474 and by the ONR under Contract N-ONR-401(53) at Cornell University, and by NSF Grant No. GP-30966 at the University of Iowa.

two specified constants, then N is the smallest integer which guarantees that the probability of correct selection (PCS) is at least P^* whenever the population parameters are in the preference zone ($\mu_{[k]} \geqq \mu_{[k-1]} + \delta^*$), i.e.,

$$\text{(2.1)} \qquad \text{PCS} \geqq P^* \qquad \text{whenever} \quad \mu_{[k]} \geqq \mu_{[k-1]} + \delta^* .$$

Bechhofer ((1954) page 23) showed that for the normal means problem the infimum of the PCS subject to the constraint of (2.1) occurs when $\mu_{[i]} = \mu_{[k]} - \delta^*$ ($i = 1, \cdots, k-1$) (this parameter configuration is called the *least favorable configuration* (LFC), and we denote the PCS for this configuration by PCS_{LFC}), and (page 20) that

$$\text{(2.2)} \qquad \text{PCS}_{\text{LFC}} = \Phi_{k-1}(\delta', \cdots, \delta'; \{\tfrac{1}{2}\}) ,$$

where Φ_{k-1} is a $(k-1)$-variate normal distribution function (df) with zero means, unit variances and off-diagonal covariances of $\frac{1}{2}$ and $\delta' = (N/2)^{\frac{1}{2}}\delta^*/\sigma$.

3. The approximations. Denoting the standard univariate normal df by Φ and its inverse by Φ^{-1}, we have:

THEOREM 1.

$$\text{PCS}_{\text{LFC}} \geqq 1 - (k-1)\Phi(-\delta')$$

and hence $N_B = (2\sigma^2/\delta^{*2})[\Phi^{-1}((1-P^*)/(k-1))]^2$, *the sample size approximation obtained by a Bonferroni inequality, satisfies the probability requirement* (2.1).

PROOF. Let $(Z_1, \cdots, Z_{k-1})$ have a $(k-1)$-variate normal distribution with zero means, unit variances, and off-diagonal covariances $\frac{1}{2}$. Then, from (2.2),

$$\begin{aligned} \text{PCS}_{\text{LFC}} &= P[\textstyle\bigcap_i \{Z_i \geqq -\delta'\}] \\ &= 1 - P[\textstyle\bigcup_i \{Z_i < -\delta'\}] \\ &\geqq 1 - \textstyle\sum_i P[Z_i < -\delta'] = 1 - (k-1)\Phi(-\delta') . \end{aligned}$$

(This may be regarded as use of a Bonferroni inequality.)

THEOREM 2.

$$\text{PCS}_{\text{LFC}} \geqq [\Phi(\delta')]^{k-1}$$

and hence $N_S = (2\sigma^2/\delta^{*2})[\Phi^{-1}(P^{*1/(k-1)})]^2$, *the sample size obtained by the Slepian inequality, satisfies the probability requirement* (2.1).

PROOF. From (2.1) and Slepian's ((1962) page 468) inequality (presented in a slightly more general form by Gupta ((1963) page 805)),

$$\text{PCS}_{\text{LFC}} \geqq \Phi_{k-1}(\delta', \cdots, \delta'; I) = [\Phi(\delta')]^{k-1} ,$$

where I is the $k-1$ by $k-1$ identity matrix.

We note that $[\Phi(\delta')]^{k-1} \geqq 1 - (k-1)\Phi(-\delta')$ follows from the Bonferroni inequality and hence N_S is a better approximation (i.e., less conservative) than N_B.

4. Numerical study of the approximation. A comparison of these two approximations with the exact sample size calculated from Bechhofer's (1954) tables

TABLE 1

P^*	$k=3$		$k=5$		$k=10$	
	N/N_B	N/N_S	N/N_B	N/N_S	N/N_B	N/N_S
.5	.3407	.5215	.3925	.5215	.4282	.5201
.6	.5533	.6895	.5373	.6386	.5410	.6142
.7	.7134	.7966	.6657	.7348	.6474	.6997
.8	.8310	.8729	.7784	.8177	.7492	.7808
.9	.9188	.9331	.8793	.8945	.8506	.8639
.95	.9555	.9609	.9286	.9348	.9058	.9115
.99	.9858	.9864	.9747	.9755	.9631	.9639
.999	.9962	.9963	.9928	.9928	.9884	.9885

is given in Table 1. Note that both of the approximations are quite close to N for high P^* and that while neither does particularly well for low P^*, N_S is considerably better than N_B.

The values of N_S and N_B for the ratios in Table 1 were computed using a Chebyshev approximation for Φ^{-1} given by Hastings (1955).

5. One multivariate normal population. The lower bounds given in Section 3 can also be applied to a selection problem involving a single multivariate normal population. Suppose that the goal is to select the factor associated with the largest population mean, that the covariance matrix is known, and that the *means procedure* of Section 2 will be used. (If we specify our preference zone in terms of the variance, e.g., see Dudewicz (1969), knowledge of the covariance matrix is not required.)

Using the notation and equations (9)—(12), (14) of Section 4 of Dudewicz (1969), we have

$$\text{PCS}_{\text{LFC}} \geqq \Phi_{k-1}((N/2)^{\frac{1}{2}}\delta^*/\sigma_{\text{MAX}}, \cdots, (N/2)^{\frac{1}{2}}\delta^*/\sigma_{\text{MAX}}; P) , \tag{5.1}$$

where $\sigma_{(i)(j)}$ is the covariance between the factors associated with $\mu_{[i]}$ and $\mu_{[j]}$,

$$\sigma^2_{\text{MAX}} = \tfrac{1}{2} \max_{1 \leqq i < j \leqq k} (\sigma^2_{(i)} + \sigma^2_{(j)} - 2\sigma_{(i)(j)}) ,$$

and P is a $k-1$ by $k-1$ correlation matrix with off-diagonal elements given by

$$\rho_{ij} = \frac{(\sigma^2_{(k)} - \sigma_{(i)(k)} - \sigma_{(j)(k)} + \sigma_{(i)(j)})}{((\sigma^2_{(i)} + \sigma^2_{(k)} - 2\sigma_{(i)(k)})(\sigma^2_{(j)} + \sigma^2_{(k)} - 2\sigma_{(j)(k)}))^{\frac{1}{2}}} .$$

Then the sample size approximation given in Theorem 1 is valid if σ_{MAX} is inserted for σ.

If in addition the off-diagonal elements of P are nonnegative, which follows if for example $\sigma_r{}^2 - \sigma_{ir} - \sigma_{jr} + \sigma_{ij} \geqq 0$ $(i, j, r = 1, \cdots, k; i \neq j \neq r)$, then the sample size approximation given in Theorem 2, inserting σ_{MAX} for σ, is valid.

Note that

$$\sigma^2_{\text{MAX}} \geqq \text{Dudewicz's } \sigma_M{}^2 = \tfrac{1}{2} \max_{1 \leqq i \leqq k-1} (\sigma^2_{(k)} + \sigma^2_{(i)} - 2\sigma_{(i)(k)}) .$$

However, since we do not know which factor is associated with $\sigma^2_{(k)}$, even when

all of the σ_{ij} are known, σ_M is not always computable; whereas σ_{MAX}, which is larger, is. (See Dudewicz ((1969) page 496) for some cases where σ_M is computable.)

6. Acknowledgment. The author is indebted to the referee for numerous suggestions and also wishes to thank Professor Robert Bechhofer for acquainting him with multiple decision procedures.

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** 16–39.

[2] BECHHOFER, R. E., KIEFER, J. and SOBEL, M. (1968). *Sequential Identification and Ranking Procedures (with special reference to Koopman–Darmois populations)*. Univ. of Chicago Press.

[3] DUDEWICZ, E. J. (1969). An approximation to the sample size in selection problems. *Ann. Math. Statist.* **40** 492–497.

[4] DUDEWICZ, E. J. and ZAINO, N. A. (1971). Sample size for selection. *Statistical Decision Theory and Related Topics* (eds. S. S. Gupta and J. Yackel). Academic Press, New York, 347–362.

[5] GUPTA, S. S. (1963). Probability integrals of the multivariate normal and multivariate t. *Ann. Math. Statist.* **34** 792–828.

[6] HASTINGS, CECIL (1955). *Approximations for Digital Computers*. Princeton Univ. Press.

[7] ROBBINS, H. and SIEGMUND, D. (1967). Mathematics of the decision sciences, Part 2. *Lectures in Applied Mathematics* **12** 267–279.

[8] SLEPIAN, D. (1962). The one-sided barrier problem for Gaussian noise. *The Bell System Tech. J.* **41** 463–502.

DEPARTMENT OF STATISTICS
UNIVERSITY OF IOWA
IOWA CITY, IOWA 52240