The Annals of Mathematical Statistics
1969, Vol. 40, No. 2, 492–497

AN APPROXIMATION TO THE SAMPLE SIZE IN SELECTION PROBLEMS[1]

By Edward J. Dudewicz

The University of Rochester

0. Summary. Let $f(\mathbf{x} \mid P_1)$ be the pdf of a $(k-1)$-dimensional normal distribution with zero means, unit variances, and correlation matrix P_1. Consider the integral, for $\delta > 0$,

$$\int_{-\delta}^{\infty} \cdots \int_{-\delta}^{\infty} f(\mathbf{x} \mid P_1)\, dx \cdots dx_{k-1} = \alpha(\delta), \quad \text{say.} \tag{1}$$

Assume that no element of P_1 is a function of δ. Note that $\alpha(\delta)$ is an increasing function of δ and $\alpha(\delta) \to 1$ as $\delta \to \infty$. The problem is to obtain an approximation to δ, for a large specified value, α, of $\alpha(\delta)$. This is given by the theorem of Section 1.

This result is used to obtain approximations to the sample size in a selection procedure of Bechhofer and in a problem of selection from a multivariate normal population. The closeness of the approximation is illustrated for the procedure of Bechhofer (Table 1).

1. The Approximation to δ.

Theorem. *For large α (near 1), an approximation to δ, which satisfies the equation*

$$\int_{-\delta}^{\infty} \cdots \int_{-\delta}^{\infty} f(\mathbf{x} \mid P_1)\, dx_1 \cdots dx_{k-1} = \alpha, \tag{2}$$

is $\delta^2 \sim -2 \log_e (1-\alpha)$. The ratio tends to 1 as $\alpha \to 1$. This approximation is independent of k.

Proof.

$$\begin{aligned} \alpha(\delta) &= \int_{-\delta}^{\infty} \cdots \int_{-\delta}^{\infty} f(\mathbf{x} \mid P_1)\, dx_1 \cdots dx_{k-1} \\ &= P[\textstyle\bigcap_{i=1}^{k-1} \{Z_i > -\delta\}] = P[\bigcap_i E_i], \quad \text{say.} \end{aligned} \tag{3}$$

Then,

$$1 - \alpha(\delta) = P[\textstyle\bigcup_i \bar{E}_i]. \tag{4}$$

Expressing the union as a disjoint union, we obtain

$$1 - \alpha(\delta) = \textstyle\sum_{i=1}^{k-1} P[E_1 E_2 \cdots \bar{E}_i]. \tag{5}$$

Now

$$P[E_1 E_2 \cdots \bar{E}_i] = \int_{-\delta}^{\infty} \cdots \int_{-\delta}^{\infty} \int_{-\infty}^{-\delta} f_i(\mathbf{x} \mid P_2)\, dx_1 \cdots dx_{i-1}\, dx_i\,,$$

Received 2 February 1968; revised 13 September 1968.

[1] The author wishes to acknowledge the support of an NSF Summer Fellowship and of a Research Assistantship supported by the ONR on Contract Nonr 401 (53) at Cornell University, and of ONR Contract N00014-68A-0091 at The University of Rochester. "This research is supported in whole or in part by The Center for Naval Analyses of the University of Rochester. Such support does not imply endorsement of the content by The Navy." Revision of this paper was supported by NSF Grant No. GP-8958.

where $f_i(\mathbf{x} \mid P_2)$ is the pdf of an i-dimensional normal distribution with zero means, unit variances, and correlation matrix obtained from P_1 by deleting the last $k - i$ rows and columns. The jth of the $i - 1$ integrals on $(-\delta, \infty)$ can be replaced by some multiplier $C_j(i)$ between 0 and $(\pi^{k-1}k)^{\frac{1}{2}}$ which tends to a limit as $\delta \to \infty$ $(j = 1, \cdots, i - 1)$, and using problem 1 of Feller (1957), p. 179, the remaining integral on $(-\infty, -\delta)$, which equals the integral on (δ, ∞), can be approximated as

$$1 - \Phi(\delta) \sim \delta^{-1}(2\pi)^{-\frac{1}{2}}e^{-\delta^2/2}.$$

Letting $C_i = C_1(i) \cdots C_{i-1}(i)$,

$$P[E_1E_2 \cdots \bar{E}_i] \sim C_i(\delta(2\pi)^{\frac{1}{2}})^{-1}e^{-\delta^2/2},$$

and letting $\bar{C} = (C_1 + \cdots + C_{k-1})(k - 1)^{-1}$, $\bar{C} = \bar{C}(\delta) \to$ a limit as $\delta \to \infty$ and an approximation to δ, such that $\alpha(\delta) = \alpha$, can be obtained from the equation

$$1 - \alpha \sim (k - 1)\bar{C}(\delta(2\pi)^{\frac{1}{2}})^{-1}e^{-\delta^2/2}.$$

Hence $\delta^2 \sim -2 \log_e (1 - \alpha)$.

2. The approximation to sample size. Ranking and selection procedures have been receiving increasing interest recently, one reason being that they furnish a method of sample size selection which is appropriate in many situations. However, the tables available for this use (e.g. those of Bechhofer (1954) and Teichroew (1955)) are not extensive enough to cover all situations which arise in practice, e.g. in drug screening where the number of populations may be very large. Also, these tables do not apply if certain independence assumptions are violated. The theorem of Section 1 provides an approximation to the sample size in a selection procedure of Bechhofer and in a multivariate situation.

Consider the following *problem*: Given k populations $\pi_1, \cdots, \pi_k$ the observations from which are known to be identically distributed except for a location parameter (i.e., an observation X_i from population π_i has cdf $F(x - \nu_i)$, where $F(\cdot)$ may be either known or unknown), select any one of the (at least one) populations associated with $\nu_{[k]} = \max(\nu_1, \cdots, \nu_k)$. For the case in which $F(\cdot)$ is known to be normal with a known variance σ^2, Bechhofer (1954) has suggested use of the *means procedure*: Take N independent vectors $\underset{\sim}{X}_j = (X_{1j}, \cdots, X_{kj})$, $j = 1, \cdots, N$, where X_{ij} denotes the jth observation from the ith population π_i; choose the population associated with the largest $\sum_{j=1}^{N} X_{ij} (i = 1, \cdots, k)$ as being associated with $\nu_{[k]}$. Let $\{\lambda^*, P^*\} (0 < \lambda^* < \infty, 1/k < P^* < 1)$ be two specified constants, and denote the ranked means by $\nu_{[1]} \leqq \cdots \leqq \nu_{[k]}$. Then N is to be set as the smallest sample size which guarantees the following *probability requirement*: We are to select the population associated with $\nu_{[k]}$, i.e. we are to make a *correct selection* (CS), with probability $P(\text{CS}) \geqq P^*$ whenever $\nu_{[k]} - \nu_{[k-1]} \geqq \lambda^*\sigma$.

It is known (see Bechhofer (1954), p. 23) that, if the means procedure is used for this problem in the normal case, then subject to $\nu_{[k]} - \nu_{[k-1]} \geqq \lambda^*\sigma$ the $P(\text{CS})$

is minimized over $\nu_{[1]}, \cdots, \nu_{[k]}$ when

$$\nu_{[1]} = \cdots = \nu_{[k-1]}, \nu_{[k]} - \nu_{[k-1]} = \lambda^*\sigma, \tag{6}$$

the *least favorable configuration* (LFC) of the population parameters; call this minimal $P(\text{CS})$ by α.

THEOREM. *As* $\alpha \to 1$,

$$N \sim -4(\lambda^*)^{-2} \log_e (1 - \alpha), \tag{7}$$

the ratio tending to 1 *as* $N \to \infty$ *due to having* $\alpha \to 1$.

PROOF. From Bechhofer (1954), p. 20, eq. (13), we know that

$$\alpha = \int_{-\lambda^*(\frac{1}{2}N)^{\frac{1}{2}}}^{\infty} \cdots \int_{-\lambda^*(\frac{1}{2}N)^{\frac{1}{2}}}^{\infty} f(\mathbf{x} \mid P_1)\, dx_1 \cdots dx_{k-1}, \tag{8}$$

where P_1 is the $(k-1) \times (k-1)$ correlation matrix with $\rho_{ij} = \frac{1}{2}(i \neq j)$. The theorem follows from (2).

3. Numerical study of the approximation. We can compare the N_A calculated via (7) with N_T from the tables of Bechhofer (1954) for moderate k, obtaining Table 1. (We use $\lambda^* = 1$ for convenience.) The tables of Gupta (1963), p. 810, and of Teichroew (1955) are useful in making comparisons for large k and for high P^*, respectively.

Note that the approximation (7) appears to overestimate N, thus being somewhat conservative, for $k \leqq 10$. If we had $\lambda^* \neq 1$, *each* entry (N_T as well as N_A) would be divided by $(\lambda^*)^2$. Thus, the ratio (which by derivation approaches 1 as $N \to \infty$ due to having $\alpha \to 1$) is independent of λ^*. Although it does not appear to be the case that $|N_A - N_T| \to 0$ (indeed the difference could conceivably be unbounded) the approach of the ratio to 1 appears fairly rapid.

4. Extension to multivariate normal selection. Use of the special nature of the correlation matrix P_1 in (8) has not been made. Thus, result (7) is valid whenever a reduction of α to form (8) is possible, provided only that P_1 does not depend on N.

Now, consider the following *problem*: Given one k-variate population Π the observations from which are k-variate normal, select any one of the (at least one) factors associated with $\nu_{[k]}$. Suppose that the *means procedure* stated in Section 2

TABLE 1

	α	0.90	0.95	0.97	0.99	0.995	0.999
	N_A	9.2104	11.9829	14.0262	18.4207	21.1933	27.6310
$k = 3$	N_T	4.9738	7.3446	9.1397	13.0849	15.6159	21.5760
	N_T/N_A	.54	.61	.65	.71	.74	.78
$k = 5$	N_T	6.7584	9.3342	11.2419	15.3633	17.9725	24.0580
	N_T/N_A	.73	.78	.80	.83	.85	.87
$k = 10$	N_T	8.8977	11.6841	13.7122	18.0251	20.7234	26.9537
	N_T/N_A	.97	.98	.98	.98	.98	.98

is used. We then observe N independent k-variate normal vectors $\underset{\sim}{X}_j = (X_{1j}, \cdots, X_{kj}), j = 1, \cdots, N$, and select the factor associated with the largest $\bar{X}_i = \sum_{j=1}^{N} X_{ij}/N$ $(i = 1, \cdots, N)$ as being associated with $\nu_{[k]}$.

Denote the k-variate normal distribution of any $\underset{\sim}{X}_j (j = 1, \cdots, N)$ by $N(\nu_1, \Sigma_1)$ where $\nu_1 = (\nu_1, \cdots, \nu_k)$ is unknown and Σ_1 is some covariance matrix, say $\Sigma_1 = (\sigma_{ij})$ for $i, j = 1, \cdots, k$. Let round brackets about a subscript denote the quantity associated with that one of the ranked means; thus, $\underset{\sim}{X}_{(j)} (j = 1, \cdots, N)$ is $N(\nu_{(1)}, \Sigma_{(1)})$ where $\nu_{(1)} = (\nu_{[1]}, \cdots, \nu_{[k]})$ and $\Sigma_{(1)} = (\sigma_{(i)(j)})$ for $i, j = 1, \cdots, k$. Now, our development will be a generalization of Bechhofer (1954), p. 20. Then (assuming $\nu_{[k]} - \nu_{[k-1]} > 0$)

$$
\begin{aligned}
P(\mathrm{CS}) &= P[\bar{X}_{(1)} < \bar{X}_{(k)}, \cdots, \bar{X}_{(k-1)} < \bar{X}_{(k)}] \\
&= P[\bar{X}_{(k)} - \bar{X}_{(1)} > 0, \cdots, \bar{X}_{(k)} - \bar{X}_{(k-1)} > 0] \qquad (9) \\
&= P[Y_1 > 0, \cdots, Y_{k-1} > 0], \qquad \text{say,}
\end{aligned}
$$

where

$$
(\bar{X}_{(1)}, \cdots, \bar{X}_{(k)}) \quad \text{is} \quad N(\nu_{(1)}, N^{-1}\Sigma_{(1)}).
$$

Now, $(Y_1, \cdots, Y_{k-1})$ is $(k-1)$-variate normal say $N(\delta, \Sigma_2)$ with $\delta_i = EY_i = \nu_{[k]} - \nu_{[i]}$ and

$$
\begin{aligned}
\sigma^2(Y_i) &= N^{-1}(\sigma^2_{(k)} + \sigma^2_{(i)} - 2\sigma_{(i)(k)}) \qquad (10) \\
\sigma(Y_iY_j) &= N^{-1}(\sigma^2_{(k)} - \sigma_{(i)(k)} - \sigma_{(j)(k)} + \sigma_{(i)(j)}).
\end{aligned}
$$

Now, (9) is minimized subject to $\nu_{[k]} - \nu_{[k-1]} \geqq \lambda^*\sigma$ (now $\sigma > 0$ is simply some constant) by the LFC (6). Thus,

$$
\begin{aligned}
\alpha &= P[(Y_i - \lambda^*\sigma)/\sigma(Y_i) > -\lambda^*\sigma/\sigma(Y_i), \qquad i = 1, \cdots, k-1] \qquad (11) \\
&= P[Z_i > -\lambda^*\sigma/\sigma(Y_i), \qquad i = 1, \cdots, k-1], \text{ say.}
\end{aligned}
$$

$(Z_1, \cdots, Z_{k-1})$ is $(k-1)$-variate normal $N(0, \Sigma_3)$ with

$$
\sigma^2(Z_i) = 1, \qquad \sigma(Z_iZ_j) = \sigma(Y_iY_j)/\sigma(Y_i)\sigma(Y_j). \qquad (12)
$$

Therefore,

$$
\begin{aligned}
\alpha &= |\Sigma_3|^{-\frac{1}{2}}/(2\pi)^{\frac{1}{2}(k-1)} \int_{(-\lambda\sigma^*/\sigma(Y_1))}^{\infty} \\
&\qquad \cdots \int_{(-\lambda^*\sigma/\sigma(Y_{k-1}))}^{\infty} \exp\left(-\tfrac{1}{2}x'\Sigma_3^{-1}x\right) dx_1 \cdots dx_{k-1} \\
&\geqq |\Sigma_3|^{-\frac{1}{2}}/(2\pi)^{\frac{1}{2}(k-1)} \int_{(-\lambda^*\sigma N^{\frac{1}{2}}/2^{\frac{1}{2}}\sigma_M)}^{\infty} \qquad (13) \\
&\qquad \cdots \int_{(-\lambda^*\sigma N^{\frac{1}{2}}/2^{\frac{1}{2}}\sigma_M)}^{\infty} \exp\left(-\tfrac{1}{2}x'\Sigma_3^{-1}x\right) dx_1 \cdots dx_{k-1},
\end{aligned}
$$

where

$$
\sigma_M = \max_{1 \leqq i \leqq k-1} [\tfrac{1}{2}(\sigma^2_{(k)} + \sigma^2_{(i)}) - \sigma_{(i)(k)}]^{\frac{1}{2}}. \qquad (14)
$$

Since Σ_3 is independent of N, our proof of (7) shows that the minimal N needed to make the lower bound in (13) equal to α, say $\bar{N}$, is such that we have the

THEOREM. *As* $\alpha \to 1$,

$$\bar{N} \sim -4(\lambda^*)^{-2}(\sigma/\sigma_M)^{-2} \log_e (1 - \alpha), \tag{15}$$

the ratio tending to 1 *as* $\bar{N} \to \infty$ *due to having* $\alpha \to 1$.

Due to its construction, $\bar{N}$ overprotects against an incorrect selection. Even if one has *no knowledge whatever* of σ_M, result (15) allows one to obtain an approximation to $\bar{N}$, *if* one sets $\sigma = \sigma_M$, i.e. if one makes his probability requirement in terms of multiples of the (unknown) σ_M.

If one has *special knowledge* then σ_M may sometimes be evaluated. E.g., if $\sigma_1^2 = \cdots = \sigma_k^2 = \sigma_0^2$ (say) and $\sigma_{(i)(k)} \geqq -B$ $(B \geqq 0)$ for $i = 1, \cdots, k$ with at least one equality, then $\sigma_M = (\sigma_0^2 + B)^{\frac{1}{2}}$. In this case the "best" such B (in terms of $\bar{N}$) is $B = 0$. (If one knew only $\sigma_{(i)(k)} \geqq C$ for $i = 1, \cdots, k$, then one would obtain $\sigma_M \geqq (\sigma_0^2 - C)^{\frac{1}{2}}$.)

The upper bound $\bar{N}$ will be "exact" as N is in (7) in that the minimal $N = \bar{N}$ iff (13) is an equality; this is so iff

$$\tfrac{1}{2}(\sigma_{(1)}^2) - \sigma_{(1)(k)} = \cdots = \tfrac{1}{2}(\sigma_{(k-1)}^2) - \sigma_{(k-1)(k)}, \tag{16}$$

which occurs if, e.g., $\sigma_1^2 = \cdots = \sigma_k^2$ and $\sigma_{(1)(k)} = \cdots = \sigma_{(k-1)(k)}$. For the latter to be satisfied it is sufficient (but not necessary) that *all* covariances be equal. (Note that the case when all covariances are equal and they and σ are known has been mentioned by Milton (1963), pp. 5–6, whose tables may then be used.)

$\bar{N}$ will be "exact" *and* approximation (15) identical to approximation (7) in the case $\sigma_1^2 = \cdots = \sigma_k^2 = \sigma^2$ and all covariances are equal to 0 (the case of independence). *Now*, if one knew that $\sigma_1^2 = \cdots = \sigma_k^2 = \sigma^2$ and that components of any k-vector tended to vary together if at all (i.e. $\sigma_{ij} \geqq 0$ $(i, j = 1, \cdots, k)$), then one might extend the LFC and choose the "worst" such covariances (those which would maximize the approximation to the "exact" $\bar{N}$). One sees (from criterion (16) and the role of σ_M in (13)) that all covariances equal to 0 accomplish this.

One may consider such extended LFC's in other cases specified by one's knowledge. If one knows $\sigma_1^2 = \cdots = \sigma_k^2 = \sigma^2$ and $\sigma_{ij} = C$ $(i \neq j; i, j = 1, \cdots, k)$ then $\bar{N}$ will be "exact" *and* the worst value of C is the smallest possible (smallest such that the variance-covariance matrix is positive definite). Since for this case the matrix is positive definite iff $\sigma^2 > C > -\sigma^2/(k - 1)$, as $C \downarrow -\sigma^2/(k-1)$, $\sigma_M \uparrow [k/(k-1)]^{\frac{1}{2}}\sigma$. Thus, in this case approximation (15) to $\bar{N}$ will differ from approximation (7) by a factor $k/(k - 1) = 1 + (k - 1)^{-1}$.

Finally, since $\sigma_M \leqq 2^{\frac{1}{2}}\sigma$ (in the case $\sigma_1^2 = \cdots = \sigma_k^2 = \sigma^2$), the approximation to $\bar{N}$ is at most twice that of (7). Thus, for large α at *most* twice as many observations per population will be required as in the case of independence.

5. Remarks. We make the following remarks.

1. For the case $k = 2$, result (7) may be derived in a manner similar to that used by Dudewicz (1966), Section 2-B (i), in another context.

2. As the problem statement of Section 2 should intimate, result (7) is con-

jectured to hold for a general location parameter family whose observations obey the (classical) Central Limit Theorem. A similar result can be conjectured for Section 4, and is not intimated by our problem statement there only so as to avoid a cumbersome notation.

3. In their monograph, Bechhofer, Kiefer, and Sobel (1968) have independently previously obtained (7) via a different proof. Although they obtain higher order corrections, their proof does not appear to generalize to multivariate problems. (See their Theorem 6.2.1, eq. 6.4.1, eq. 6.4.3, Lemma 6.5.1, and eq. 14.2.10.)

4. Thanks are especially due to Professor Robert E. Bechhofer for his suggestions for and guidance of Dudewicz (1966), which contains result (7), and to a referee for substantial simplification of the presentation and for the reference to Gupta (1963).

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. (1966). The efficiency of a nonparametric selection procedure: largest location parameter case. Technical Report No. 14, Department of Operations Research, Cornell Univ.

[4] FELLER, W. (1957). *An Introduction to Probability Theory and Its Applications, Volume I* (Second Edition). Wiley, New York.

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

[6] MILTON, R. C. (1963). Tables of the equally correlated multivariate normal probability integral. Technical Report No. 27, Department of Statistics, Univ. of Minnesota.

[7] TEICHROEW, D. (1955). Probabilities Associated with Order Statistics in Samples from Two Normal Populations with Equal Variance. Chemical Corps Engineering Agency, Engineering Statistics Unit, Army Chemical Center, Maryland, December **7**, ENASR No. ES-3.