Ann. Inst. Statist. Math.
Vol. 41, No. 2, 401–408 (1989)

ON COMPUTATION OF INTEGRALS FOR SELECTION FROM MULTIVARIATE NORMAL POPULATIONS ON THE BASIS OF DISTANCES

ROY C. MILTON[1] AND M. HASEEB RIZVI[2]

[1]*National Eye Institute, National Institutes of Health, Bethesda, MD 20892, U.S.A.*
[2]*Sysorex International, Inc., 10590 North Tantau Avenue, Cupertino, CA 95014, U.S.A.*

(Received August 31, 1987; revised August 31, 1988)

Abstract. A procedure for selecting the t largest of k multivariate normal populations on the basis of distance is reviewed. Computation of integrals of products of non-central Beta distribution and density functions, required for implementing the procedure, is described. A table of minimum sample sizes needed to guarantee a specified probability of correct selection is given (Table 1).

Key words and phrases: Ranking and selection procedures, multivariate normal, non-central Beta distribution, adaptive Romberg quadrature.

1. Introduction

A new statistical methodology began in the mid-1950's with the early definitive formulations by Robert Bechhofer, Shanti Gupta and Milton Sobel for problems requiring selection and ordering (or ranking) of populations.

An experimenter is often required to compare k (≥ 2) populations. For example, these populations may represent different drugs for treatment of a certain disease, or different fertilizers for increasing the yield of a certain crop. A parameter $\hat{\theta}$ characterizes each population, e.g., θ may be the success rate in treatment with a drug or the crop yield associated with a fertilizer. While a classical approach has been to test the "homogeneity" hypothesis that $\theta_1 = \cdots = \theta_k$, where $\theta_1,\ldots,\theta_k$ are the unknown values of the parameter θ for the k populations, this is often inadequate if the real goal of the experimenter is to identify the best population (i.e., the best drug or the best fertilizer).

The extensive development of methods and applications in selection and ranking procedures has generally followed either the "indifference-zone" approach of Bechhofer (1954) or the "subset-selection" approach due to Gupta (1956). In its simplest form, the indifference-zone procedure is to take a single sample of fixed size n from each of the k populations and to select the population yielding the largest value of an appropriately chosen statistic $\hat{\theta}$ as the best population, i.e., as the population with the largest (true but unknown) parameter θ. The minimum value of n is determined so that the probability of selecting the best population—the probability of correct selection—is guaranteed to be at least a specified value P^* $(1/k < P^* < 1)$ whenever the difference between the largest and second largest parameters is at least a specified value δ^*, or $\theta_{[k]} - \theta_{[k-1]} \geq \delta^*$. If the two largest (but unknown) population parameters differ by less than δ^*, we are "indifferent" to which population is selected as largest. In contrast, for fixed sample size n and P^*, a simple subset-selection procedure selects a subset of the k populations so that the best population is included in the subset (whose size may be determined by the data) with probability $\geq P^*$, whatever the configuration of the unknown parameters.

The literature on selection and ranking methods now includes many variations and generalizations of these approaches. Recent overviews are provided by Bechhofer (1985) and Gupta and Panchapakesan (1985). Other useful references with broad coverage are the books by Gibbons *et al.* (1977) and Gupta and Panchapakesan (1979), and the categorized bibliography by Dudewicz and Koo (1982).

2. Selection from multivariate normal populations

Alam and Rizvi (1966) considered two problems of selection from k multivariate normal populations. In Problem I it is required to select the t best of the k populations, $1 \leq t \leq k$. (In Problem II it is required to select a subset of the k populations which contains the t best populations. We restrict our attention to Problem I and its more difficult computational aspects for preparing tables required for its solution and application.) The probability of a correct selection is required to be at least a pre-assigned quantity P^*, $1\Big/\binom{k}{t} < P^* < 1$.

Let π_i represent a p-variate normal population with $(p \times 1)$ mean column vector μ_i and $(p \times p)$ positive definite covariance matrix Σ_i, $i = 1,\ldots,k$. We rank the k populations according to the values of the Mahalanobis (1930) distance function $\theta_i = \mu_i'\Sigma_i^{-1}\mu_i$. Then π_i is called larger than π_j if $\theta_i > \theta_j$. A sample of size n $(n \times p)$ is taken from each population, yielding sample mean vectors $\bar{x}_i$ and covariance matrices S_i. Sample distance functions are $U_i = \bar{x}_i'\Sigma_i^{-1}\bar{x}_i$ and $V_i = (\bar{x}_i'S_i^{-1}\bar{x}_i)(n-p)/(np)$, and we use $\hat{\theta}_i = U_i$ or $\hat{\theta}_i = V_i$ according to whether the covariances are supposed to be known or unknown. Note that nU_i is distributed as non-central χ^2 with p degrees of freedom and non-centrality parameter $n\mu_i'\Sigma_i^{-1}\mu_i$, and nV_i has the non-central F distribution with p and $(n-p)$ degrees of freedom and the same non-centrality parameter.

Denoting the i-th largest sample parameter by $\hat{\theta}_{[i]}$, we select as largest the t populations corresponding to the t largest sample values $\hat{\theta}_{[k]},\ldots,\hat{\theta}_{[k-t+1]}$. We further preassign values $\delta_1 > 0$ and $\delta_2 > 1$ and then determine n to guarantee that the probability of a correct selection is at least P^* whenever

$$\theta_{[k-t+1]} - \theta_{[k-t]} \geq \delta_1 \quad \text{and}$$

$$\theta_{[k-t+1]}/\theta_{[k-t]} \geq \delta_2 ,$$

that is, whenever the t largest populations are sufficiently larger than the next largest population.

3. The solution

The ranking of k multivariate normal populations, as presented here in the formulation of Alam and Rizvi (1966), is shown by them to reduce to ranking (with respect to the non-centrality parameters) the non-central χ^2 or non-central F populations. In order to apply the procedure, one must fix P^*, δ_1 and δ_2 and solve (3.1) for n:

$$P^* = t\int_0^\infty H^{k-t}(x,\lambda_1)\{1 - H(x,\lambda_2)\}^{t-1}h(x,\lambda_2)dx , \tag{3.1}$$

where H and h are non-central distribution and density functions with non-centrality parameters

$$\lambda_1 = \frac{n\delta_1}{(\delta_2 - 1)}, \qquad \lambda_2 = \delta_2\lambda_1 .$$

For the case where covariance matrices are known, H is the non-central χ^2 distribution function with p degrees of freedom. The non-central χ^2 distribution was computed using the modified Bessel function representation and recurrence relations given by Seber (1963), and the density function was computed using recurrence relations and definitions given by Alam and Rizvi (1966). Numerical methods similar to, but somewhat simpler than, those described below were used to solve (3.1) for $n\delta_1$ for $t = 1$; $P^* = 0.90, 0.95, 0.99$; $k = 2(1)10$; $p = 1, 5, 9, 29$; $\delta_2 = 1.01, 1.05(0.05)$ $1.25(0.25)2.00(0.5)3.00$ and reported as Table S.1 in Gibbons *et al.* (1977). Solutions for $t = 2$ were computed but not reported.

For the more realistic case of unknown covariance matrices, H is $p/(n-p)$ times a non-central F distribution, i.e., the distribution function of the ratio of a non-central χ^2 variable with p degrees of freedom and non-centrality parameter λ and an independent central χ^2 variable with

$n-p$ degrees of freedom. Specifically, Alam and Rizvi (1966) write (3.1) as

$$P^* = t\int_0^\infty G_{p,q}^{k-t}(x,\lambda_1)\{1-G_{p,q}(x,\lambda_2)\}^{t-1}g_{p,q}(x,\lambda_2)dx\,, \tag{3.2}$$

where $q=n-p$, h is

$$g_{p,q}(x,\lambda)=\frac{e^{-\lambda/2}}{\Gamma(q/2)}\sum_{r=0}^{\infty}\frac{x^{(p/2)+r-1}\Gamma(p/2+q/2+r)}{(1+x)^{(p/2)+(q/2)+r}\Gamma(p/2+r)}\cdot\frac{\lambda^r}{2^r r!}\,,\qquad x>0\,,$$

and

$$G_{p,q}(x,\lambda)=\int_0^x g_{p,q}(y,\lambda)dy\,.$$

Solution of (3.2) for n involves the following features.

First, transformation of variables from non-central F to non-central Beta yields

$$\begin{aligned} P^* = t\int_0^1 B^{k-t}&\left(y;\ \frac{p}{2},\frac{q}{2},\lambda_1\right)\\ &\cdot\left\{1-B\left(y;\frac{p}{2},\frac{q}{2},\lambda_2\right)\right\}^{t-1}b\left(y;\frac{p}{2},\frac{q}{2},\lambda_2\right)dy\,, \end{aligned} \tag{3.3}$$

where the region of integration is now finite. The non-central Beta density function is

$$b(y;c,d,\lambda)=\sum_{r=0}^{\infty}\frac{e^{-\lambda/2}\lambda^r}{2^r r!}\frac{y^{c+r-1}(1-y)^{d-1}}{B(c+r,d)}dy\,,$$

with $B(c+r,d)=\Gamma(c+r)\Gamma(d)/\Gamma(c+r+d)$, and the non-central Beta distribution function is

$$B(x;c,d,\lambda)=\sum_{j=0}^{\infty}\frac{e^{-\lambda/2}(\lambda/2)^j}{j!}I(x;c+j,d)\,,$$

with

$$I(x;c+j,d)=\int_0^x\frac{w^{c+j-1}(1-w)^{d-1}dw}{B(c+j,d)}\,,$$

the complete Beta function.

Second, evaluation of $B(x;c,d,\lambda)$ is done using the representation by recurrence relations among generalized Laguerre polynomials given by

Seber (1963). Following a suggestion by Donald E. Amos (personal communication), evaluation of $b(x; c, d, \lambda)$ is done by differentiating the recurrence relation representation for $B(x; c, d, \lambda)$.

Third, numerical integration is done using a subroutine from the University of Wisconsin Computer Center for adaptive Romberg integration of single integrals, with requested accuracy of $5\cdot 10^{-5}$.

Fourth, efficiency in computing and avoidance of underflow and overflow is achieved by rewriting the integral in (3.3) as $t\int_0^1 = t\int_0^L + t\int_L^U + t\int_U^1$, applying inequalities to the outer integrals, choosing (L, U) so that the value of each outer integral is less than $5\cdot 10^{-6}/t$, and then ignoring the outer integrals.

Fifth, fixing δ_1, δ_2, k, t and p, the right-hand side of (3.3) is computed for an array of judiciously spaced values of n which yield P^*-values from below 0.90 to just above 0.99.

Finally, values of n corresponding to $P^* = 0.90$, 0.95, 0.99 are computed using IMSL subroutine IQHSCU for quasi-Hermite piecewise cubic polynomial interpolation, as described by Akima (1970).

4. The table

Table 1 gives the minimum sample size n needed in each of k populations to satisfy the specified δ_1, δ_2 and P^* requirements in selecting the t $(< k)$ p-variate normal populations with the largest Mahalanobis

Table 1. Minimum sample size n needed in each of k populations to satisfy the $(\delta_1, \delta_2, P^*)$ requirement in selecting the t $(< k)$ p-variate normal populations with the largest Mahalanobis distance when covariance matrices are unknown.

t	k	p	δ_1	$P^* = .90$			$P^* = .95$			$P^* = .99$		
				$\delta_2 = 1.5$	2.0	3.0	1.5	2.0	3.0	1.5	2.0	3.0
1	2	2	1	74.7	35.0	20.0	121.9	56.5	31.7	241.9	111.2	61.5
			2	58.6	25.5	13.8	95.2	40.8	21.6	188.4	79.7	41.2
			3	53.3	22.4	11.9	86.4	35.6	18.2	170.6	69.3	34.5
			4	50.6	20.9	10.9	82.0	33.1	16.6	161.7	64.1	31.2
			10	45.9	18.2	9.2	74.0	28.5	13.7	145.7	54.7	25.3
1	2	6	1	79.1	39.8	25.3	126.2	61.3	37.1	246.2	116.0	66.9
			2	62.7	29.9	18.4	99.4	45.1	26.2	192.6	84.1	45.9
			3	57.4	26.6	16.2	90.5	39.8	22.6	174.7	73.4	38.9
			4	54.7	25.0	15.2	86.0	37.2	20.9	165.8	68.2	35.5
			10	49.9	22.2	13.3	78.1	32.5	17.8	149.7	58.7	29.3
1	2	10	1	83.4	44.4	30.2	130.6	66.0	42.1	250.6	120.8	72.2
			2	66.9	34.1	22.8	103.5	49.4	30.6	196.7	88.3	50.4
			3	61.4	30.8	20.4	94.5	44.0	26.9	178.8	77.6	43.2
			4	58.7	29.1	19.3	90.1	41.3	25.1	169.8	72.3	39.7
			10	53.9	26.3	17.4	82.1	36.6	21.9	153.7	62.7	33.4

Table 1. (continued).

t	k	p	δ_1	$P^* = .90$			$P^* = .95$			$P^* = .99$		
				$\delta_2 = 1.5$	2.0	3.0	1.5	2.0	3.0	1.5	2.0	3.0
1	3	2	1	113.3	52.8	29.7	165.8	76.7	42.7	292.8	134.5	74.2
			2	88.7	38.3	20.4	129.5	55.3	29.0	228.1	96.4	49.7
			3	80.6	33.5	17.3	117.4	48.2	24.4	206.6	83.7	41.6
			4	76.5	31.1	15.8	111.4	44.7	22.2	195.7	77.4	37.6
			10	69.2	26.9	13.1	100.5	38.3	18.1	176.3	66.0	30.3
1	3	6	1	117.7	57.6	35.1	170.2	81.5	48.2	297.2	139.3	79.7
			2	92.9	42.6	25.0	133.6	59.6	33.6	232.2	100.7	54.3
			3	84.6	37.7	21.7	121.5	52.4	28.8	210.6	87.9	46.0
			4	80.5	35.3	20.1	115.4	48.8	26.4	199.7	81.5	41.8
			10	73.2	30.9	17.2	104.5	42.4	22.2	180.3	70.0	34.4
1	3	10	1	122.0	62.2	40.1	174.5	86.2	53.3	301.5	144.1	85.0
			2	97.0	46.9	29.4	137.7	63.9	38.0	236.4	105.0	58.9
			3	88.7	41.9	25.9	125.5	56.5	33.1	214.7	92.1	50.3
			4	84.6	39.4	24.2	119.5	52.9	30.6	203.8	85.6	46.0
			10	77.2	35.0	21.3	108.5	46.4	26.3	184.3	74.0	38.4
1	4	2	1	137.1	63.7	35.8	192.3	88.8	49.4	#	148.3	81.7
			2	107.3	46.2	24.4	150.2	64.0	33.4	251.5	106.3	54.8
			3	97.4	40.4	20.7	136.1	55.8	28.2	227.7	92.3	45.8
			4	92.5	37.5	18.9	129.1	51.7	25.5	215.8	85.3	41.3
			10	83.6	32.3	15.6	116.5	44.3	20.8	194.4	72.7	33.3
1	4	6	1	141.9	68.5	41.1	196.6	93.7	54.8	#	153.1	87.2
			2	111.5	50.5	29.0	154.3	68.4	38.0	255.7	110.6	59.4
			3	101.5	44.6	25.1	140.2	60.0	32.5	231.8	96.5	50.2
			4	96.5	41.6	23.1	133.2	55.8	29.8	219.8	89.4	45.6
			10	87.6	36.3	19.7	120.5	48.4	24.9	198.4	76.7	37.3
1	4	10	1	145.8	73.2	46.2	201.0	98.4	60.0	#	157.9	92.6
			2	115.6	54.8	33.5	158.4	72.6	42.5	259.8	114.9	63.9
			3	105.6	48.7	29.3	144.3	64.1	36.8	235.9	100.6	54.5
			4	100.6	45.7	27.3	137.2	59.9	34.0	223.9	93.5	49.8
			10	91.6	40.4	23.7	124.5	52.5	29.0	202.4	80.7	41.4
1	5	2	1	154.3	71.7	40.1	211.2	97.6	54.2	#	158.1	87.1
			2	120.8	51.9	27.4	165.0	70.3	36.7	268.3	113.3	58.3
			3	109.6	45.4	23.2	149.6	61.3	30.8	242.9	98.4	48.8
			4	104.1	42.1	21.1	141.8	56.7	28.0	230.2	90.9	44.0
			10	94.0	36.2	17.4	128.0	48.6	22.7	207.3	77.5	35.4
1	5	6	1	158.7	76.5	45.5	215.6	102.4	59.6	#	162.9	92.6
			2	124.9	56.2	32.0	169.1	74.6	41.3	272.4	117.6	63.0
			3	113.7	49.6	27.5	153.6	65.4	35.2	246.9	102.6	53.2
			4	108.1	46.2	25.3	145.9	60.8	32.2	234.2	95.0	48.3
			10	98.0	40.3	21.5	132.0	52.6	26.8	211.3	81.5	39.5
1	5	10	1	163.0	81.2	50.7	220.0	107.1	64.9	#	167.7	98.0
			2	129.1	60.5	36.4	173.2	78.9	45.8	276.5	121.9	67.5
			3	117.8	53.7	31.8	157.7	69.6	39.5	251.0	106.7	57.5
			4	112.1	50.3	29.5	149.9	64.9	36.4	238.3	99.1	52.5
			10	102.0	44.3	25.5	136.0	56.6	30.9	215.4	85.5	43.5

Table 1. (continued).

t	k	p	δ_1	$P^* = .90$			$P^* = .95$			$P^* = .99$		
				$\delta_2 = 1.5$	2.0	3.0	1.5	2.0	3.0	1.5	2.0	3.0
2	3	2	1	110.9	51.3	28.7	163.6	75.3	41.8	291.0	133.4	73.4
			2	86.5	36.9	19.4	127.4	54.0	28.1	226.5	95.4	49.0
			3	78.4	32.2	16.4	115.4	47.0	23.6	205.0	82.7	40.9
			4	74.4	29.9	14.9	109.4	43.5	21.3	194.2	76.4	36.9
			10	67.2	25.7	12.4	98.6	37.2	17.4	174.9	65.1	29.7
2	3	6	1	115.3	56.1	34.1	167.9	80.1	47.2	295.5	138.2	79.0
			2	90.7	41.2	24.0	131.5	58.3	32.7	230.6	99.7	53.7
			3	82.5	36.4	20.8	119.5	51.1	27.9	209.1	86.9	45.3
			4	78.4	34.0	19.2	113.4	47.6	25.6	198.3	80.5	41.2
			10	71.2	29.7	16.4	102.6	41.2	21.5	178.9	69.1	33.8
2	3	10	1	119.6	60.8	39.1	172.3	84.8	52.4	299.9	143.0	84.3
			2	94.8	45.5	28.4	135.7	62.6	37.1	234.7	104.0	58.2
			3	86.6	40.5	25.0	123.5	55.3	32.2	213.1	91.1	49.6
			4	82.5	38.1	23.4	117.5	51.7	29.8	202.3	84.6	45.4
			10	75.2	33.8	20.5	106.6	45.3	25.6	182.9	73.1	37.8
2	4	2	1	155.7	71.9	40.0	212.2	97.6	54.0	#	157.9	86.8
			2	121.5	51.7	27.0	165.3	70.0	36.2	268.2	112.9	57.9
			3	110.1	45.0	22.7	149.7	60.8	30.4	242.7	97.9	48.4
			4	104.4	41.7	20.6	141.9	56.3	27.5	230.0	90.4	43.6
			10	94.2	35.8	16.8	127.9	48.1	22.3	207.1	77.0	35.0
2	4	6	1	160.1	76.7	45.4	216.6	102.4	59.5	#	162.8	92.4
			2	125.6	56.0	31.6	169.5	74.3	40.9	272.2	117.2	62.6
			3	114.2	49.2	27.1	153.8	65.0	34.7	246.7	102.1	52.7
			4	108.5	45.8	24.9	146.0	60.4	31.7	234.2	94.6	47.8
			10	98.2	39.8	20.9	131.9	52.1	26.3	211.1	81.0	39.0
2	4	10	1	164.5	81.4	50.6	220.9	107.2	64.7	#	167.5	97.7
			2	129.7	60.3	36.1	173.6	78.6	45.4	276.4	121.5	67.1
			3	118.2	53.4	31.4	157.9	69.2	39.0	250.8	106.3	57.0
			4	112.5	49.9	29.1	150.0	64.5	35.9	238.0	98.7	52.0
			10	102.2	43.8	25.0	135.9	56.1	30.4	215.1	85.0	43.1
2	5	2	1	182.7	84.3	46.8	241.1	110.9	61.3	#	172.3	94.7
			2	142.6	60.6	31.6	187.9	79.5	41.1	292.6	123.2	63.2
			3	129.2	52.8	26.5	170.1	69.1	34.4	264.9	106.9	52.7
			4	122.5	48.9	24.0	161.3	63.9	31.1	250.9	98.7	47.5
			10	110.5	41.8	19.6	145.3	54.5	25.2	225.9	84.0	38.1
2	5	6	1	187.1	89.1	52.3	245.5	115.7	66.7	#	177.4	100.2
			2	146.7	64.9	36.2	192.0	83.8	45.7	296.8	127.5	67.8
			3	133.3	57.0	30.9	174.2	73.3	38.8	268.9	111.0	57.1
			4	126.6	53.0	28.3	165.3	68.0	35.4	255.0	102.8	51.8
			10	114.5	45.9	23.7	149.4	58.6	29.2	229.9	88.0	42.2
2	5	10	1	191.4	93.8	57.4	249.8	120.4	72.0	#	181.9	105.6
			2	150.8	69.2	40.7	196.2	88.1	50.2	300.9	131.8	72.4
			3	137.3	61.1	35.2	178.3	77.4	43.1	273.0	115.2	61.4
			4	130.6	57.1	32.5	169.4	72.1	39.6	259.0	106.9	56.0
			10	118.5	49.9	27.7	153.4	62.6	33.3	233.9	92.0	46.2

#: $n > 300$.

distance when covariance matrices are unknown and estimated from the samples. The range of parameters is $P^* = 0.90, 0.95, 0.99$; $t = 1, 2$; $k = t + 1(1)5$; $p = 2, 6, 10$; $\delta_1 = 1, 2, 3, 4, 10$ and $\delta_2 = 1.5, 2.0, 3.0$. Linear interpolation to get values for $p = 4, 8$ will be exact when the result is rounded up to the next highest integer, except for two values which will be one unit too large. The smoothness of the tabled values further suggests that interpolation for odd p and for other values of δ_1 will be good also.

This table expands considerably on the early version which appears as Table S.3 in Gibbons *et al.* (1977) in terms of coverage of parameters and flexibility of use. This was made possible by improved computing and analytic techniques, and by increased computer resources for a computation-intensive problem. Numerical examples which use these tabled values (for $t = 1$) are also given by Gibbons *et al.* (1977).

REFERENCES

Akima, H. (1970). A new method of interpolation and smooth curve fitting based on local procedures, *J. Assoc. Comput. Mach.*, **17**, 589–602.

Alam, K. and Rizvi, M. H. (1966). Selection from multivariate normal populations, *Ann. Inst. Statist. Math.*, **18**, 307–318.

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.

Bechhofer, R. E. (1985). Selection and ranking problems—some personal reminiscences, and thoughts about its past, present, and future, *Amer. J. Math. Management Sci.*, **5**, 201–234.

Dudewicz, E. J. and Koo, J. O. (1982). *The Complete Categorized Guide to Statistical Selection and Ranking Procedures*, Series in Mathematical and Management Sciences, Vol. 6, American Sciences Press, Columbus, Ohio.

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

Gupta, S. S. (1956). On a decision rule for a problem in ranking means, Mimeograph Series No. 150, Institute of Statistics, University of North Carolina, Chapel Hill, North Carolina.

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

Gupta, S. S. and Panchapakesan, S. (1985). Subset selection procedures: review and assessment, *Amer. J. Math. Management Sci.*, **5**, 235–311.

Mahalanobis, P. C. (1930). On tests and measures of group divergence, *Journal of the Asiatic Society of Bengal*, **26**, 541–588.

Seber, G. A. F. (1963). The non-central chi-squared and non-central beta distributions, *Biometrika*, **50**, 542–544.