SELECTING THE NORMAL POPULATION WITH THE SMALLEST COEFFICIENT OF VARIATION

Ajit C. Tamhane
Department of IE/MS and Department of Statistics
Northwestern University, Evanston, IL 60208

Anthony J. Hayter
Department of Statistics and Operations Technology
University of Denver, Denver, CO 80208

SYNOPTIC ABSTRACT

We consider the problem of selecting the normal population with the smallest coefficient of variation, which is a natural goal when the means as well as the variances of the populations are unknown and unequal. The indifference-zone approach (Bechhofer 1954) to this problem has been previously considered by Choi, Jeon and Kim (1982). We review their selection procedure and provide tables of sample sizes for it. Next we consider the subset selection approach of Gupta (1956, 1965). We propose a natural selection procedure, derive its least favorable configuration and provide tables of critical constants. An example is given to illustrate the two procedures.

Keywords and Phrases: Indifference-zone approach; Noncentral t-distribution; Subset selection approach; Unequal variances.

1. INTRODUCTION

The problem of selecting the normal population with the largest mean has received much attention in the ranking and selection literature; see Gibbons, Olkin and Sobel (1977) and Gupta and Panchapakesan (1979). When the populations have unknown and unequal variances, procedures for selecting the population with the largest mean under the indifference-zone approach have been proposed by Dudewicz and Dalal (1975) and Rinott (1978). However, in this case, often the experimenter is interested in selecting a population with a large mean and a small variance. One formulation of this problem was studied by Santner and Tamhane (1984), who specified separate indifference zones on the means, μ_i, and the variances, σ_i^2. In this paper we study an alternative formulation in which the μ_i and the σ_i are combined into a single parameter for each population, namely the inverse of the coefficient of variation, $\theta_i = \mu_i/\sigma_i$. We study both the *indifference-zone approach* (Bechhofer 1954) and the *subset selection approach* (Gupta 1956, 1965) to the problem of selecting the normal population with the largest θ_i.

Choi, Jeon and Kim (1982) have offered a different motivation for selecting the normal population with the largest θ_i. They consider a quality control application in which manufactured parts have a lower specification limit, which may be assumed to be zero. If the output of the i-th process $(1 \leq i \leq k)$ is distributed as $N(\mu_i, \sigma_i^2)$ then its fraction defective is $p_i = \Phi(-\theta_i)$, where $\Phi(\cdot)$ is the standard normal c.d.f. Hence the smallest p_i corresponds to the largest θ_i. In passing we note that the univariate case of Alam and Rizvi's (1966) multivariate selection problem corresponds to selecting the normal population with the largest value of $|\theta_i|$.

The paper is organized as follows. The basic notation and assumptions are defined in Section 2. Choi et al.'s (1982) indifference-zone procedure for selecting the largest θ_i is reviewed in Section 3. We provide tables of exact sample sizes for their procedure. In Section 4 we propose a subset selection procedure for the largest θ_i. The proof of the least favorable configuration (LFC) of this procedure is given in the Appendix. Our method of proof of the LFC for the subset selection procedure is different from Choi et al.'s and it also applies to the indifference-zone procedure with only slight modifications (not given here, but available from the authors). We provide tables of critical constants for the subset selection procedure. Section 5 gives a real data example to illustrate the indifference-zone selection and subset selection procedures. Some computational details are provided in Section 6.

2. PRELIMINARIES

Let Π_i denote a normal population with mean μ_i and standard deviation σ_i, and let $\theta_i = \mu_i/\sigma_i$ $(1 \leq i \leq k)$. We assume that the μ_i's and the σ_i's are unknown. We further assume that the μ_i's are known to be nonnegative (in which case only it makes sense to compare the θ_i's). Without loss of generality, suppose that the populations are labeled so that $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_k$ and this labelling is unknown to the experimenter. The population associated with θ_k is referred to as the "best" population and is assumed to be unique (if more than one population is tied for the "best" then one is arbitrarily selected to be the "best"). Finally, we assume that a finite upper bound, $\theta^* > 0$, is specified on all the θ_i. Without such a bound there does not exist a single-stage procedure that achieves the probability requirement stated in (1) below. The reason is that the estimators of the θ_i have noncentral t-distributions whose noncentrality parameters and hence their variances increase beyond limit as the $\theta_i \to \infty$. If too high a value of θ^* is specified then the procedure becomes conservative since too high a sample size is required (see the tables of the sample sizes given in the sequel). On the other hand, If too low a value of θ^* is specified then the procedure becomes anti-conservative and may not meet the specified probability requirement.

3. INDIFFERENCE-ZONE APPROACH

For the indifference-zone approach the experimenter's *goal* is to select the "best" population. This is referred to as the *correct selection (CS)*. The experimenter wants that the following *probability requirement* be satisfied:

$$P(CS) \geq P^* \text{ whenever } \theta_k - \theta_{k-1} \geq \delta^* \text{ and } \theta_k \leq \theta^*, \tag{1}$$

where $\theta^* > 0$, P^* $(1/k < P^* < 1)$ and δ^* $(0 < \delta^* \leq \theta^*)$ are the constants specified by the experimenter.

Throughout we assume that an i.i.d. random sample with a common sample size n is taken from each population Π_i. It will be seen from the tables at the end of the paper that n required to guarantee (1) is strictly increasing in θ^* for any fixed $\{\delta^*, P^*\}$, which supports the observation above that without a finite upper bound $\theta^* > 0$ on all the θ_i, the required n will be unbounded. Also, care must be exercised in specifying θ^* since an excessively high value would result in an excessively high n, while an excessively low value would result in the probability requirement (1) not being met.

The following single-stage natural selection procedure was proposed by Choi et al. (1982): Take a random sample of size n from each Π_i. Compute the sample mean $\overline{X}_i$, the sample standard deviation S_i, and $\widehat{\theta}_i = \overline{X}_i/S_i$ for the data from Π_i $(1 \leq i \leq k)$. Select the Π_i associated with $\widehat{\theta}_{\max}$ as the best population.

The main design problem is to determine the sample size n that will guarantee the probability requirement (1). Toward this end, Choi et al. (1982) showed that the LFC that minimizes the P(CS) over the so-called *preference zone*, $\{(\theta_1, \theta_2, \ldots, \theta_k) : \theta_k - \theta_{k-1} \geq \delta^*, \theta_k \leq \theta^*\}$, is given by

$$\theta_1 = \ldots = \theta_{k-1} = \theta_k - \delta^*, \theta_k = \theta^*. \tag{2}$$

To prove this result the first step is to show that the infimum of P(CS) subject to $\theta_k - \theta_{k-1} \geq \delta^*$ and $\theta_k = \theta$ (where $\theta \in [\delta^*, \theta^*]$ is fixed) occurs at the slippage configuration: $\theta_1 = \ldots = \theta_{k-1} = \theta_k - \delta^*, \theta_k = \theta$. The proof uses a theorem from Barr and Rizvi (1966) which applies the fact that for $i = 1, 2, \ldots, k$, the $T_i = \widehat{\theta}_i\sqrt{n}$ are independent noncentral t random variables (r.v.'s) with $n-1$ degrees of freedom (d.f.) and noncentrality parameters (n.c.p.) $\lambda_i = \theta_i\sqrt{n}$ (denoted as $T_i \sim t_{n-1}(\lambda_i)$) and that their cumulative distribution functions (c.d.f.'s), $F_\nu(\cdot|\lambda_i)$, form a stochastically increasing family of distributions in λ_i. The second step of the proof is to show that the P(CS) in the slippage configuration is a decreasing function of $\theta_k = \theta$, so the infimum over the preference zone is attained when $\theta = \theta^*$, which is the LFC given in (2). We thus get

$$P_{\text{LFC}}(CS) = \int_{-\infty}^{\infty} \left[F_{n-1}(x|(\theta^* - \delta^*)\sqrt{n})\right]^{k-1} f_{n-1}(x|\theta^*\sqrt{n})dx, \tag{3}$$

where $f_\nu(\cdot|\lambda)$ is the probability density function (p.d.f.) of $t_\nu(\lambda)$. Exact sample sizes, n, calculated using the above expression for selected values of k, P^*, θ^* and δ^* are given in Tables 1 -6.

An excellent approximation to the exact sample sizes can be calculated using the variance stabilizing transformation (Sen 1964)

$$Y_i = \sinh^{-1}(\widehat{\theta}_i/\sqrt{2}) \ \ (1 \le i \le k).$$

This transformation was used by Choi et al. (1982) to show that the large sample approximation to n is given by

$$n = \frac{d^2}{2}\left[\ln\left\{\frac{\theta^* + \sqrt{2+\theta^{*2}}}{(\theta^* - \delta^*) + \sqrt{2 + (\theta^* - \delta^*)^2}}\right\}\right]^{-2}, \tag{4}$$

where $d = d(k, P^*)$ is the solution to the equation

$$\int_{-\infty}^{\infty} [\Phi(x+d)]^{k-1}\, d\Phi(x) = P^*. \tag{5}$$

The values of $d(k, P^*)$ have been tabulated Bechhofer (1954) and Gupta (1963) for selected values of k and P^*. The approximate values of n calculated using the above formula are also given in Tables 1 - 6. One can see that they are always less than or equal to the exact values, and are quite close.

Table 1: Exact and Approximate Values of Sample Size n Per Population for Indifference Zone Selection ($\theta^* = 1.0, \delta^* = 0.5$)

	k						
P^*	2	3	4	5	6	7	8
0.90	18	27	32	36	40	42	45
	17	26	31	35	38	41	43
0.95	29	39	45	50	53	56	58
	28	38	44	48	52	55	57
0.99	56	68	75	80	84	87	89
	56	68	75	79	83	86	89

The upper entry in each cell is the exact n and the lower entry is the approximate n.

Table 2: Exact and Approximate Values of Sample Size n Per Population for Indifference Zone Selection ($\theta^* = 2.0, \delta^* = 0.5$)

	k						
P^*	2	3	4	5	6	7	8
0.90	35	52	64	72	78	83	87
	34	51	61	69	75	79	84
0.95	56	76	89	97	104	110	115
	55	75	86	95	101	107	111
0.99	110	133	147	157	165	171	176
	110	133	146	156	163	169	174

The upper entry in each cell is the exact n and the lower entry is the approximate n.

Table 3: Exact and Approximate Values of Sample Size n Per Population for Indifference Zone Selection ($\theta^* = 3.0, \delta^* = 0.5$)

	k						
P^*	2	3	4	5	6	7	8
0.90	64	98	118	133	145	155	163
	63	95	115	129	141	150	157
0.95	105	143	166	182	195	205	214
	104	141	163	178	191	201	209
0.99	206	250	275	293	307	319	329
	207	250	275	293	307	319	329

The upper entry in each cell is the exact n and the lower entry is the approximate n.

Table 4: Exact and Approximate Values of Sample Size n Per Population for Indifference Zone Selection ($\theta^* = 2.0, \delta^* = 1.0$)

	k						
P^*	2	3	4	5	6	7	8
0.90	8	12	15	16	18	19	20
	7	11	13	15	16	17	18
0.95	13	17	20	22	23	25	26
	12	16	18	20	21	23	24
0.99	24	29	32	34	36	37	39
	23	28	31	33	34	36	37

The upper entry in each cell is the exact n and the lower entry is the approximate n.

Table 5: Exact and Approximate Values of Sample Size n Per Population for Indifference Zone Selection ($\theta^* = 4.0, \delta^* = 1.0$)

	k						
P^*	2	3	4	5	6	7	8
0.90	25	38	46	51	56	60	63
	24	36	43	48	52	56	59
0.95	40	55	63	70	74	78	82
	39	52	61	66	71	75	78
0.99	78	95	104	111	117	121	125
	77	93	102	109	114	118	122

The upper entry in each cell is the exact n and the lower entry is the approximate n.

4. SUBSET SELECTION APPROACH

For the subset selection approach the experimenter's *goal* is to select a subset of the k populations that contains the "best" population. This is referred to as the *correct selection (CS)*. Any selection procedure must satisfy the following *probability requirement*:

$$P(CS) \geq P^* \text{ for all } (\theta_1, \theta_2, \ldots, \theta_k) \text{ and } \theta_k \leq \theta^* \quad (6)$$

where $\theta^* > 0$ and P^* ($1/k < P^* < 1$) are prespecified constants.

Analogous to Gupta (1956), we propose the following single-stage natural selection procedure: Take a random sample of size n from each Π_i. Compute $\widehat{\theta}_i = \overline{X}_i / S_i$ for

Table 6: Exact and Approximate Values of Sample Size n Per Population for Indifference Zone Selection ($\theta^* = 6.0, \delta^* = 1.0$)

	k						
P^*	2	3	4	5	6	7	8
0.90	55	83	100	113	123	131	138
	53	80	97	109	118	126	133
0.95	89	121	140	154	164	173	180
	87	118	137	150	161	169	176
0.99	174	211	232	247	259	268	277
	174	211	232	247	259	268	277

The upper entry in each cell is the exact n and the lower entry is the approximate n.

the data from Π_i $(1 \le i \le k)$ and select the subset

$$\mathcal{S} = \left\{ \Pi_i : \widehat{\theta}_i \ge \widehat{\theta}_{\max} - c \right\}, \tag{7}$$

where $c > 0$ is a constant that must be determined to guarantee the probability requirement (6).

We now state a theorem concerning the LFC of the subset selection procedure (7).

Theorem 1 : *The probability of a correct selection of the subset selection procedure (7) is minimized over the entire parameter space at the equal parameter configuration (EPC): $\theta_1 = \cdots = \theta_k = \theta^*$, and this minimum is given by*

$$P_{LFC}(CS) = \int_{-\infty}^{\infty} \left[F_{n-1}(x + b|\theta^* \sqrt{n}) \right]^{k-1} f_{n-1}(x|\theta^* \sqrt{n}) dx, \tag{8}$$

where $b = c\sqrt{n}$.

Proof: The P(CS) of (7) is given by

$$\begin{aligned} P(CS) &= P\{\Pi_k \in \mathcal{S}\} \\ &= P\{\widehat{\theta}_k \ge \widehat{\theta}_i - c \ \ \forall\, i \ne k\} \\ &= P\{\widehat{\theta}_k \sqrt{n} + c\sqrt{n} \ge \widehat{\theta}_i \sqrt{n} \ \ \forall\, i \ne k\} \end{aligned}$$

$$= P\{t_{n-1}(\theta_k\sqrt{n}) + b \geq t_{n-1}(\theta_i\sqrt{n}) \ \ \forall \ \ i \neq k\}$$
$$= \int_{-\infty}^{\infty} \prod_{i=1}^{k-1} F_{n-1}(x + b|\lambda_i) f_{n-1}(x|\lambda_k) dx,$$

where $\lambda_i = \theta_i\sqrt{n}$ $(1 \leq i \leq k)$ are the n.c.p.'s. Because of the stochastically increasing property of the noncentral t-distribution in its n.c.p., it follows that the above P(CS) expression is minimized at the EPC: $\theta_1 = \theta_2 = \cdots = \theta_k = \theta$ (say), where $\theta \leq \theta^*$; see, e.g., equation (11.4) in Gupta and Panchapakesan (1979). We need to find a further minimum with respect to the common value θ. In the lemma given in the Appendix let the U_i be distributed as $\sqrt{\chi^2_\nu/\nu}$ r.v.'s, which makes the X_i i.i.d. noncentral t r.v.'s with ν d.f. and n.c.p. $= \lambda$. Then the lemma shows that the P(CS) expression in the EPC is a decreasing function of the common n.c.p. $\lambda = \theta\sqrt{n}$. Hence it follows that the minimizing value of θ is θ^* and the overall minimum is given by (8). □

In Tables 7, 8 and 9 we list the critical constants $b = c\sqrt{n}$ for $P^* = 0.90, 0.95$ and 0.99, respectively, for selected values of k, θ^* and n. It is worth noting that if n is small then c may exceed θ^*. Therefore, $\theta_{\max} - c \leq \theta^* - c \leq 0$ while $\theta_i \geq 0 \ \forall \ i$. Hence $\theta_i \geq \theta_{\max} - c \ \forall \ i$, which means that all Π_i will be included in the subset with high probability. For example, from Table 7 for $P^* = 0.90$ we see that for $k = 8, \theta^* = 5, n = 10$, we have $b = 15.870$ and hence $c = 15.870/\sqrt{10} = 5.019 > \theta^* = 5$. However, for $k = 8, \theta^* = 5, n = 20$, we have $b = 13.096$ and hence $c = 13.096/\sqrt{20} = 2.928 < \theta^* = 5$. The point here is that if n is not sufficiently large then a specified P^* may not be achieved unless all populations are included in the subset. Thus the selection procedure may not be an effective screening procedure. The example in Section 5 illustrates this point.

5. EXAMPLE

Vardeman (1994) gave the data (originally from Pellicane, 1990) shown in Table 10 on the strengths of wood joints made using eight commercially available construction adhesive glues. Eight wood joints were tested for each glue. If there is too much variability in the joint strengths for the glue with the highest average joint strength, then we may want to choose another glue with somewhat lower average, but also

Table 7: Critical Constants $b = c\sqrt{n}$ for Subset Selection ($P^* = 0.90$)

k	θ^*	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
2	1	2.513	2.352	2.305	2.283	2.270
	2	3.671	3.377	3.292	3.252	3.229
	3	5.050	4.607	4.480	4.420	4.384
	4	6.509	5.915	5.745	5.664	5.617
	5	8.003	7.258	7.045	6.943	6.884
3	1	3.276	2.991	2.908	2.867	2.843
	2	4.869	4.340	4.187	4.112	4.068
	3	6.732	5.940	5.711	5.601	5.534
	4	8.693	7.636	7.331	7.183	7.095
	5	10.699	9.377	8.994	8.809	8.699
4	1	3.727	3.351	3.242	3.189	3.157
	2	5.587	4.890	4.688	4.590	4.532
	3	7.744	6.704	6.404	6.258	6.171
	4	10.011	8.624	8.224	8.030	7.914
	5	12.327	10.592	10.093	9.850	9.706
5	1	4.051	3.603	3.473	3.410	3.372
	2	6.106	5.275	5.036	4.919	4.850
	3	8.477	7.240	6.885	6.712	6.609
	4	10.965	9.317	8.844	8.615	8.478
	5	13.505	11.447	10.856	10.570	10.399
6	1	4.304	3.795	3.648	3.577	3.534
	2	6.514	5.573	5.301	5.170	5.091
	3	9.055	7.653	7.252	7.058	6.942
	4	11.718	9.853	9.319	9.061	8.907
	5	14.438	12.105	11.441	11.117	10.925
7	1	4.513	3.951	3.790	3.711	3.664
	2	6.852	5.812	5.516	5.372	5.285
	3	9.534	7.990	7.550	7.337	7.209
	4	12.343	10.288	9.704	9.420	9.251
	5	15.209	12.644	11.912	11.558	11.348
8	1	4.691	4.082	3.908	3.823	3.773
	2	7.141	6.015	5.696	5.540	5.447
	3	9.944	8.273	7.799	7.570	7.432
	4	12.878	10.655	10.026	9.721	9.538
	5	15.870	13.096	12.310	11.929	11.700

Table 8: Critical Constants $b = c\sqrt{n}$ for Subset Selection ($P^* = 0.95$)

k	θ^*	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
2	1	3.348	3.069	2.990	2.952	2.931
	2	4.959	4.437	4.290	4.221	4.181
	3	6.853	6.068	5.847	5.743	5.683
	4	8.848	7.798	7.503	7.364	7.283
	5	10.888	9.574	9.204	9.030	8.928
3	1	4.122	3.683	3.560	3.501	3.467
	2	6.196	5.372	5.143	5.033	4.969
	3	8.599	7.368	7.024	6.861	6.765
	4	11.121	9.480	9.021	8.805	8.675
	5	13.696	11.646	11.071	10.798	10.639
4	1	4.584	4.032	3.879	3.806	3.763
	2	6.943	5.911	5.624	5.489	5.409
	3	9.656	8.118	7.691	7.489	7.370
	4	12.500	10.451	9.882	9.613	9.454
	5	15.403	12.841	12.131	11.796	11.596
5	1	4.918	4.277	4.101	4.017	3.967
	2	7.485	6.290	5.960	5.804	5.712
	3	10.425	8.647	8.157	7.924	7.786
	4	13.497	11.137	10.489	10.174	9.991
	5	16.647	13.686	12.873	12.485	12.255
6	1	5.180	4.465	4.270	4.177	4.122
	2	7.913	6.582	6.217	6.045	5.943
	3	11.033	9.056	8.514	8.257	8.106
	4	14.295	11.666	10.946	10.604	10.402
	5	17.620	14.339	13.439	13.013	12.761
7	1	5.397	4.618	4.406	4.306	4.246
	2	8.268	6.820	6.246	6.239	6.129
	3	11.538	9.390	8.802	8.526	8.362
	4	14.954	12.010	11.319	10.950	10.732
	5	18.435	14.872	13.899	13.439	13.169
8	1	5.582	4.747	4.521	4.413	4.350
	2	8.572	7.022	6.601	6.402	6.284
	3	11.968	9.671	9.047	8.751	8.578
	4	15.515	12.463	11.633	11.244	11.009
	5	19.130	15.323	14.284	13.798	13.508

Table 9: Critical Constants $b = c\sqrt{n}$ for Subset Selection ($P^* = 0.99$)

k	θ^*	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$
2	1	5.235	4.533	4.347	4.261	4.212
	2	7.983	6.659	6.306	6.142	6.048
	3	11.128	9.156	8.627	8.382	8.241
	4	14.416	11.793	11.088	10.760	10.572
	5	17.768	14.492	13.606	13.202	12.966
3	1	6.073	5.113	4.864	4.750	4.684
	2	9.359	7.564	7.093	6.876	6.751
	3	13.087	10.422	9.721	9.396	9.209
	4	16.975	13.437	12.502	12.070	11.821
	5	20.956	16.516	15.356	14.815	14.503
4	1	6.582	5.449	5.159	5.027	4.951
	2	10.201	8.092	7.545	7.293	7.148
	3	14.284	11.164	10.350	9.975	9.759
	4	18.550	14.399	13.316	12.813	12.529
	5	22.881	17.697	16.357	15.734	15.367
5	1	6.952	5.687	5.366	5.219	5.135
	2	10.817	8.467	7.862	7.585	7.426
	3	15.165	11.689	10.791	10.380	10.142
	4	19.688	15.083	13.885	13.344	13.021
	5	24.303	18.550	17.063	16.384	15.969
6	1	7.246	5.871	5.524	5.367	5.277
	2	11.304	8.758	8.107	7.809	7.638
	3	15.838	12.097	11.134	10.691	10.436
	4	20.606	15.619	14.328	13.743	13.404
	5	25.463	19.206	17.609	16.866	16.450
7	1	7.490	6.021	5.653	5.486	5.391
	2	11.709	8.996	8.306	7.991	7.810
	3	16.450	12.436	11.411	10.943	10.672
	4	21.350	16.045	14.700	14.066	13.710
	5	26.381	19.753	18.047	17.281	16.822
8	1	7.699	6.148	5.762	5.587	5.486
	2	12.059	9.198	8.474	8.144	7.954
	3	16.942	12.720	11.648	11.156	10.872
	4	22.050	16.428	14.995	14.339	13.967
	5	27.125	20.191	18.419	17.588	17.150

Table 10: Sample Means ($\overline{X}_i$), Standard Deviations (S_i) and Inverses of Coefficients of Variation ($\widehat{\theta}_i = \overline{X}_i/S_i$) for Wood Joint Strength Data (Units are kN)

Glue i	1	2	3	4
$\overline{X}_i$	1821	1968	1439	616
S_i	214	435	243	205
$\widehat{\theta}_i$	8.509	4.524	5.922	2.865
Glue i	5	6	7	8
$\overline{X}_i$	1354	1424	1694	1669
S_i	135	191	225	551
$\widehat{\theta}_i$	10.080	7.455	7.529	3.029

lower variance. Therefore the goal of selecting the glue with the smallest coefficient of variation or the largest θ_i is reasonable.

First consider the indifference-zone selection goal. Suppose that based on past experience with similar data θ^* is specified to be 10. Also suppose that $\delta^* = 2$ and $P^* = 0.90$. From Table I in Bechhofer (1954) we find that $d = d(8, 0.90) = 2.8691$. Then using the formula (4), we obtain

$$n = \frac{(2.8691)^2}{2}\left[\ln\left\{\frac{10+\sqrt{2+10^2}}{(10-2)+\sqrt{2+(10-2)^2}}\right\}\right]^{-2} = 84.74 \text{ or } 85.$$

Here we have $n = 8$, which is too small. In fact, $n = 8$ guarantees P^* slightly less than 0.35 as can be verified using $d = d(8, 0.35) = 0.8897$ from the same table in Bechhofer (1954) which corresponds to $n = 8.15$. Glue ♯5 with $\widehat{\theta}_{\max} = 10.080$ will be selected as the "best" glue, but only with confidence slightly less than 35% if $\theta_{\max}$ is ≤ 10 and exceeds other θ_i's by at least $\delta^* = 2$.

Next consider the subset selection goal with $P^* = 0.90$. The value of b is not tabled for the combination $k = 8, n = 8, \theta^* = 10$ and $P^* = 0.90$. Using our program, this value is calculated to be 34.584 and hence $c = 34.584/\sqrt{8} = 12.227$. Now, $\widehat{\theta}_{\max} - c = 10.080 - 12.227 = -2.147$ and all $\widehat{\theta}_i$ exceed this negative lower bound; therefore all glues are selected in the subset and there is no screening. This is due to the fact that with a small sample size of 8 per glue, we cannot guarantee with

90% confidence that the "best" glue is in the subset unless we include all glues in the subset. To eliminate any glue we must settle for a lower P^*. For example, to eliminate Glue ♯4 with the smallest $\widehat{\theta}_i = 2.865$, the critical constant c must be no more than $10.080 - 2.865 = 7.215$; the corresponding value of P^* is calculated to be 0.655.

6. COMPUTATIONAL DETAILS

All computations were implemented in R, version 1.8.0, base package. The p.d.f. and the c.d.f. of the noncentral t-distribution are built-in functions 'dt' and 'pt' in R; the actual algorithms are described in Becker, Chambers and Wilks (1988) and Lenth (1989). The integrals were computed by the 'integrate' function in R, which is based on QUADPACK routines 'dqags' and 'dqagi' by Piessens et al. (1983) available from Netlib. An estimate of the modulus of the absolute error is provided for each evaluation of integral.

Instead of directly solving the equation

$$f(b) = \int_{-\infty}^{\infty} \left[F_{n-1}(x + b|\theta^* \sqrt{n})\right]^{k-1} f_{n-1}(x|\theta^* \sqrt{n}) dx - P^* = 0,$$

we minimized $f^2(b)$. The 'optim' function in R was used for minimization. The 'optim' function was called with default optimizing method, which is the Nelder and Mead (1965) method. The convergence criterion used was $f^2(b) < 10^{-9}$.

ACKNOWLEDGMENTS

The authors thank Mr. Kunyang Shi, a statistics doctoral student at Northwestern, for computing the tables of the critical constants for the subset selection procedure. We also thank two referees for their useful comments and suggestions. This research was partially supported by the National Heart, Lung and Blood Institute Grant 1 R01 HL082725-01A1 and the National Security Agency Grant H98230-07-1-0068.

A. APPENDIX

Lemma 1 : *Let $Y_1, Y_2, \ldots, Y_k$ be i.i.d. $N(\lambda, 1)$ r.v.'s and let $U_1, U_2, \ldots, U_k$ be i.i.d. nonnegative continuous r.v.'s independent of the Y_i's. Define $X_i = Y_i/U_i$ $(1 \le i \le k)$ and let*

$$h(\lambda) = P\{X_k + c \ge X_1, \ldots, X_{k-1}\}. \tag{A.9}$$

Then for $c, \lambda > 0$, $h(\lambda)$ is decreasing in λ.

Proof: The c.d.f. of X_i can be written as

$$\begin{aligned} F(x|\lambda) &= P\{Y_i \le xU_i\} \\ &= \int_0^\infty g(u) \int_{-\infty}^{xu} (2\pi)^{-1/2} \exp\left\{-\frac{1}{2}(y-\lambda)^2\right\} dy\, du, \end{aligned} \tag{A.10}$$

where $g(u)$ is the p.d.f. of U_i. By differentiating (A.10) with respect to x we find that the p.d.f. of X_i is given by

$$f(x|\lambda) = (2\pi)^{-1/2} \int_0^\infty \exp\left\{-\frac{1}{2}(xu-\lambda)^2\right\} ug(u)du.$$

Therefore we can write (A.9) as

$$h(\lambda) = (2\pi)^{-k/2} \int_{R_1(\boldsymbol{u})} \int_{R_2(\boldsymbol{x})} \exp\left\{-\frac{1}{2}\sum_{i=1}^{k}(x_i u_i - \lambda)^2\right\} \prod_{i=1}^{k} dx_i u_i g(u_i) du_i,$$

where the integrations are over the regions

$$R_1(\boldsymbol{u}) = \left\{\boldsymbol{u} = (u_1, \ldots, u_k) \in \mathcal{R}^k : u_i \ge 0 \ \ (1 \le i \le k)\right\}$$

and

$$R_2(\boldsymbol{x}) = \left\{\boldsymbol{x} = (x_1, \ldots, x_k) \in \mathcal{R}^k : x_k + c > x_i \ \ (1 \le i \le k)\right\}.$$

We have

$$\frac{dh(\lambda)}{d\lambda} = (2\pi)^{-k/2} \int_{R_1(\boldsymbol{u})} \int_{R_2(\boldsymbol{x})} \exp\left\{-\frac{1}{2}\sum_{i=1}^{k}(x_i u_i - \lambda)^2\right\} \left\{\sum_{i=1}^{k}(x_i u_i - \lambda)\right\}$$

$$\begin{aligned}
&\times \prod_{i=1}^{k} dx_i\, u_i g(u_i) du_i \\
&= (2\pi)^{-k/2} \int_{R_1(\boldsymbol{u})} \int_{R_2(\boldsymbol{x})} \exp\left\{-\frac{1}{2}\sum_{i=1}^{k}(x_i u_i - \lambda)^2\right\} \\
&\quad \times \{(k-1)(x_1 u_1 - \lambda) + (x_k u_k - \lambda)\} \prod_{i=1}^{k} dx_i\, u_i g(u_i) du_i \\
&= A + B \ \ \text{(say)},
\end{aligned}$$

where

$$\begin{aligned}
A &= (k-1)(2\pi)^{-k/2} \int_{R_1(\boldsymbol{u})} \int_{R_2(\boldsymbol{x})} \exp\left\{-\frac{1}{2}\sum_{i=1}^{k}(x_i u_i - \lambda)^2\right\} (x_1 u_1 - \lambda) \\
&\quad \times \prod_{i=1}^{k} dx_i\, u_i g(u_i) du_i \\
&= (k-1)(2\pi)^{-k/2} \int_{R_1(\boldsymbol{u})} \int_{x_k=-\infty}^{\infty} \int_{x_2=-\infty}^{x_k+c} \cdots \int_{x_{k-1}=-\infty}^{x_k+c} \exp\left\{-\frac{1}{2}\sum_{i=2}^{k}(x_i u_i - \lambda)^2\right\} \\
&\quad \times \left[\int_{x_1=-\infty}^{x_k+c} (x_1 u_1 - \lambda) \exp\left\{-\frac{1}{2}(x_1 u_1 - \lambda)^2\right\} dx_1\right] \prod_{i=2}^{k} dx_i \prod_{i=1}^{k} u_i g(u_i) du_i \\
&= (k-1)(2\pi)^{-k/2} \int_{R_1(\boldsymbol{u})} \int_{x_k=-\infty}^{\infty} \int_{x_2=-\infty}^{x_k+c} \cdots \int_{x_{k-1}=-\infty}^{x_k+c} \exp\left\{-\frac{1}{2}\sum_{i=2}^{k}(x_i u_i - \lambda)^2\right\} \\
&\quad \times \left(-\frac{1}{u_1}\right) \exp\left\{-\frac{1}{2}\left[(x_k + c)u_1 - \lambda\right]^2\right\} \prod_{i=2}^{k} dx_i \prod_{i=1}^{k} u_i g(u_i) du_i, \qquad \text{(A.11)}
\end{aligned}$$

where we have used the fact that

$$\int (x_1 u_1 - \lambda) \exp\left\{-\frac{1}{2}(x_1 u_1 - \lambda)^2\right\} dx_1 = \left(-\frac{1}{u_1}\right) \exp\left\{-\frac{1}{2}(x_1 u_1 - \lambda)^2\right\}. \qquad \text{(A.12)}$$

Next,

$$\begin{aligned}
B &= (2\pi)^{-k/2} \int_{R_1(\boldsymbol{u})} \int_{R_2(\boldsymbol{x})} \exp\left\{-\frac{1}{2}\sum_{i=1}^{k}(x_i u_i - \lambda)^2\right\} (x_k u_k - \lambda) \prod_{i=1}^{k} dx_i\, u_i g(u_i) du_i \\
&= (k-1)(2\pi)^{-k/2} \int_{R_1(\boldsymbol{u})} \int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{x_1} \cdots \int_{x_{k-1}=-\infty}^{x_1} \exp\left\{-\frac{1}{2}\sum_{i=1}^{k-1}(x_i u_i - \lambda)^2\right\} \\
&\quad \times \left[\int_{x_k=x_1-c}^{\infty} (x_k u_k - \lambda) \exp\left\{-\frac{1}{2}(x_k u_k - \lambda)^2\right\} dx_k\right] \prod_{i=1}^{k-1} dx_i \prod_{i=1}^{k} u_i g(u_i) du_i
\end{aligned}$$

$$= (k-1)(2\pi)^{-k/2} \int_{R_1(\boldsymbol{u})} \int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{x_1} \cdots \int_{x_{k-1}=-\infty}^{x_1} \exp\left\{-\frac{1}{2}\sum_{i=1}^{k-1}(x_i u_i - \lambda)^2\right\}$$
$$\times \left(\frac{1}{u_k}\right) \exp\left\{-\frac{1}{2}\left[(x_1 - c)u_k - \lambda\right]^2\right\}$$
$$\times \prod_{i=1}^{k-1} dx_i \prod_{i=1}^{k} u_i g(u_i) du_i, \qquad \text{(A.13)}$$

where we have again used (A.12).

Relabeling $x_k + c$ as x_1 in (A.11) and combining it with (A.13) we obtain

$$\frac{dh(\lambda)}{d\lambda} = (k-1)(2\pi)^{-k/2} \int_{R_1(\boldsymbol{u})} \int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{x_1} \cdots \int_{x_{k-1}=-\infty}^{x_1}$$
$$\times \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{k-1}(x_i u_i - \lambda)^2 + \{(x_1 - c)u_k - \lambda\}^2\right]\right\} \left(\frac{1}{u_k} - \frac{1}{u_1}\right)$$
$$\times \prod_{i=1}^{k-1} dx_i \prod_{i=1}^{k} u_i g(u_i) du_i.$$

Define

$$C(x_1) = \int_{u_2=0}^{\infty} \cdots \int_{u_{k-1}=0}^{\infty} \int_{x_2=-\infty}^{x_1} \cdots \int_{x_{k-1}=-\infty}^{x_1} \exp\left\{-\frac{1}{2}\sum_{i=2}^{k-1}(x_i u_i - \lambda)^2\right\}$$
$$\times \prod_{i=2}^{k-1} u_i g(u_i) du_i dx_i.$$

Clearly, $C(x_1)$ is a nondecreasing function of x_1. Then

$$\frac{dh(\lambda)}{d\lambda} = (k-1)(2\pi)^{-k/2} \int_{u_1=0}^{\infty} \int_{u_k=0}^{\infty} \int_{x_1=-\infty}^{\infty} C(x_1) g(u_1) g(u_k)(u_1 - u_k)$$
$$\times \exp\left\{-\frac{1}{2}\left[(x_1 u_1 - \lambda)^2 + \{(x_1 - c)u_k - \lambda\}^2\right]\right\} dx_1 du_1 du_k.$$

Now,

$$(x_1 u_1 - \lambda)^2 + \{(x_1 - c)u_k - \lambda\}^2 = (u_1^2 + u_k^2)\left[x_1 - \frac{\lambda(u_1 + u_k) + cu_k^2}{u_1^2 + u_k^2}\right]^2$$
$$+ \frac{\{\lambda(u_1 - u_k) + cu_1 u_k\}^2}{u_1^2 + u_k^2}.$$

Also, define

$$\begin{aligned} D(u_1, u_k) &= \left[\exp\left\{-\frac{[\lambda(u_1 - u_k) + cu_1u_k]^2}{2(u_1^2 + u_k^2)}\right\}\right] \\ &\quad \times \int_{-\infty}^{\infty} C(x_1)\exp\left\{-\frac{u_1^2 + u_k^2}{2}\left[x_1 - \frac{\lambda(u_1 + u_k) + cu_k^2}{u_1^2 + u_k^2}\right]^2\right\} dx_1. \end{aligned}$$

Therefore we have

$$\frac{dh(\lambda)}{d\lambda} = (k-1)(2\pi)^{-k/2}\int_{u_k=0}^{\infty}\int_{u_1=u_k}^{\infty}(u_1-u_k)g(u_1)g(u_k)[D(u_1,u_k)-D(u_k,u_1)]du_1du_k.$$

Since $c, \lambda \geq 0$ then $u_1 \geq u_k \geq 0$,

$$\begin{aligned} D(u_1, u_k) - D(u_k, u_1) &= \sqrt{\frac{2\pi}{u_1^2 + u_k^2}}\left(\left[\exp\left\{-\frac{[\lambda(u_1 - u_k) + cu_1u_k]^2}{2(u_1^2 + u_k^2)}\right\}\right] E[C(X_1)]\right. \\ &- \left.\left[\exp\left\{-\frac{[\lambda(u_1 - u_k) + cu_1u_k]^2}{2(u_1^2 + u_k^2)}\right\}\right] E[C(X_k)]\right) \\ &\leq \sqrt{\frac{2\pi}{u_1^2 + u_k^2}}\left(\left[\exp\left\{-\frac{[\lambda(u_1 - u_k) + cu_1u_k]^2}{2(u_1^2 + u_k^2)}\right\}\right]\right. \\ &\quad \left.\times \{E[C(X_1)] - E[C(X_k)]\}\right) \\ &\leq 0, \end{aligned}$$

where X_1 and X_k are normal random variables with means

$$\frac{\lambda(u_1 + u_k) + cu_k^2}{u_1^2 + u_k^2} \text{ and } \frac{\lambda(u_1 + u_k) + cu_k^2}{u_1^2 + u_k^2},$$

respectively, and variances $1/(u_1^2 + u_k^2)$. The last inequality follows from the fact that $C(x)$ is nondecreasing. Thus $dh(\lambda)/d\lambda \leq 0$ and the lemma is proved. □

REFERENCES

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

Barr, D. R. and Rizvi, M. H. (1966), "An introduction to ranking and selection procedures," *J. Amer. Statist. Assoc.*, **61**, 640 -646.

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.

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988), *The New S Language*, Wadsworth & Brooks/Cole.

Choi, C. H., Jeon, J. W. and Kim, W-C (1982), "Selection problems in terms of coefficients of variation," *Journal of the Korean Statistical Society*, **XI**, 12-24.

Dudewicz, E. J. and Dalal, S. R. (1975), "Allocation of observations in ranking and selection with unequal variances," *Sankhya, Ser. B*, **37**, 28-78.

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

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

Gupta, S. S. (1963), "Probability integrals of multivariate normal and multivariate t," *Ann. Math. Statist.*, **34**, 792–828.

Gupta, S. S. (1965), "On some multiple decision (selection and ranking) rules," *Technometrics*, **7**, 225–245.

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

Lenth, R. V. (1989), "Algorithm AS 243 - Cumulative distribution function of the non-central t-distribution," *Appl. Statist.*, **38**, 185–189.

Nelder, J. A. and Mead, R. (1965), "A simplex algorithm for function minimization," *Computer J.*, **7**, 308–313.

Pellicane, P. (1990), "Behavior of rubber-based elastomeric construction adhesive," *J. Testing and Eval.*, **18**, 256 -264.

Piessens, R., deDoncker-Kapenga, E., Uberhuber, C. and Kahaner, D. (1983), *'Quadpack: A Subroutine Package for Automatic Integration*, Springer Verlag.

Rinott, Y. (1978), "On two-stage procedures and related probability inequalities," *Communications in Statistics, Ser. A*, **8**, 799-811.

Santner, T. J. and Tamhane, A. C. (1984), "Designing experiments for selecting a normal population with a large mean and a small variance," *Design of Experiments: Ranking and Selection* (Eds. T. J. Santner and A. C. Tamhane), Marcel-Dekker, (1984), 179–198.

Sen, P. K. (1964), "Tests for the validity of the fundamental assumption in dilution (-direct) assays," *Biometrics*, **20**, 770–784.

Vardeman, S. B. (1994), *Statistics for Engineering Problem Solving*, Boston: PWS Publishing Co.