SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853

TECHNICAL REPORT NO. 728

November 1986
(Revised January 1987)

SUBSET SELECTION FOR NORMAL MEANS
IN MULTI-FACTOR EXPERIMENTS

By

Robert E. Bechhofer
Charles W. Dunnett*

Approved for public release; distribution unlimited.

Research partially supported by the U.S. Army Research Office through the Mathematical Sciences Institute of Cornell University and by U.S. Army Research Office Contract DAAL03-86-K-0046 at Cornell University, and the Natural Sciences and Engineering Research Council of Canada at McMaster University.

*Department of Clinical Epidemiology and Biostatistics and Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada L8N 3Z5.

SUBSET SELECTION FOR NORMAL MEANS IN MULTI-FACTOR EXPERIMENTS

Robert E. Bechhofer

Charles W. Dunnett

School of Operations Research
and Industrial Engineering
Cornell University
Ithaca, NY 14853

Dept. of Clinical Epidemiology
and Biostatistics, and Dept.
of Mathematics and Statistics
McMaster University
Hamilton, Ontario, Canada L8N 3Z5

Key Words and Phrases: selection procedures; subset approach;
factorial experiments; multivariate Student t distribution

ABSTRACT

The procedure of Gupta [1956], [1965] for selecting a random sized subset of $k \geq 2$ normal populations which contains the population with the largest population mean when the populations have a common variance is generalized to multi-factor experiments. Two-factor experiments with equal replication on each factor-level combination are discussed in detail. The cases of zero and non-zero interactions between factor levels are considered. For the two-factor, zero interaction case with a common number of observations at each factor-level combination, a table of constants necessary to implement the procedure is provided for experiments having selected levels per factor; the constants are equi-coordinate upper percentage points of a multivariate Student t distribution.

1. INTRODUCTION

A subset selection procedure for normal means arising in single-factor experiments when the populations have a common unknown variance was proposed by Gupta [1956], [1965]. Gupta and Sobel [1957] provided a table of constants needed to implement that procedure and guarantee the subset probability requirement. Earlier, Dunnett [1955] in his Tables 1a and 1b had provided a related table of constants. Recently Bechhofer and Dunnett

[1987] in their Tables A.1-A.4 provided an extensive set of tables which extended and improved the Dunnett tables and also the related tables of Gupta and Sobel.

In the present article we show how Gupta's normal means procedure can be generalized to multi-factor experiments in which there is no interaction between the factor-levels. We derive the equation which must be solved in order to obtain constants needed to implement the procedure and guarantee the subset probability requirement. A set of tables of such constants is provided for experiments having (2,2), (2,3), (2,4), (3,3), (3,4) or (4,4) levels per factor and a common number of observations on each factor-level combination. A much more comprehensive set of such special purpose constants is given as Tables C.1-C.3 in Bechhofer and Dunnett [1987].

In a related paper, Bechhofer and Dunnett [1986] showed how the two-stage indifference-zone selection procedure for single-factor experiments proposed by Bechhofer, Dunnett and Sobel [1954] could be extended to multi-factor experiments.

2. TWO-FACTOR EXPERIMENTS WITH NO INTERACTION

To illustrate the two-factor zero interaction case, we assume that we have $a \cdot b$ populations π_{ij} ($1 \leq i \leq a$, $1 \leq j \leq b$) representing the $a \cdot b$ factor-level combinations, satisfying the fixed effects ANOVA model with $n \geq 1$ observations being taken on each factor-level combination, i.e.,

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk} \quad (1 \leq i \leq a, 1 \leq j \leq b, 1 \leq k \leq n) \quad (2.1)$$

with $\sum_{i=1}^a \alpha_i = \sum_{j=1}^b \beta_j = 0$. Here Y_{ijk} is the k th observation on the i th level of Factor A and the j th level of Factor B, and the ϵ_{ijk} are i.i.d. $N(0, \sigma^2)$ r.v.'s; α_i (β_j) is referred to as the "effect" of the i th level of Factor A (j th level of Factor B) on the population mean. The parameters μ and σ^2 ,

and the α_i and β_j are assumed to be unknown. We denote the ordered values of the α_i and β_j by $\alpha_{[1]} \leq \dots \leq \alpha_{[a]}$ and $\beta_{[1]} \leq \dots \leq \beta_{[b]}$, respectively. It is assumed that the values of the $\alpha_{[i]}$ and $\beta_{[j]}$, and the pairing of the $\alpha_{[i]}$ and $\beta_{[j]}$ with the $a \cdot b$ populations π_{ij} ($1 \leq i \leq a, 1 \leq j \leq b$), are completely unknown. We consider the following goal:

GOAL: To select a subset of the $a \cdot b$ populations
which contains the population associated with $\alpha_{[a]}$ and $\beta_{[b]}$. (2.2)

A correct selection (CS) is said to have been made if the selected subset contains the population associated with $\alpha_{[a]}$ and $\beta_{[b]}$. We restrict consideration to statistical procedures which guarantee the following subset probability of a correct selection ($P\{CS\}$):

PROBABILITY REQUIREMENT:

$$\inf_{\substack{(\alpha_1, \dots, \alpha_a) \\ (\beta_1, \dots, \beta_b) \\ \mu, \sigma^2}} P\{CS\} \geq 1 - \alpha \quad (2.3)$$

Here $1 - \alpha$ with $(1/a \cdot b < 1 - \alpha < 1)$ is specified by the experimenter before the start of experimentation.

Generalizing Gupta [1956], [1965] we propose the following subset selection procedure:

PROCEDURE: "Include in the selected subset all populations
for which

$$\begin{aligned} \bar{y}_{i..} &\geq \bar{y}_{[a]..} - c_A s_v / \sqrt{bn} \\ \text{and} \quad \bar{y}_{.j.} &\geq \bar{y}_{. [b] .} - c_B s_v / \sqrt{an}. \end{aligned} \quad (2.4)$$

Here $\bar{y}_{i..} = \sum_{j=1}^b \sum_{k=1}^n y_{ijk} / bn$ ($1 \leq i \leq a$), $\bar{y}_{[a]..} = \max_{1 \leq i \leq a} \bar{y}_{i..}$,
 $\bar{y}_{.j.} = \sum_{i=1}^a \sum_{k=1}^n y_{ijk} / an$ ($1 \leq j \leq b$), $\bar{y}_{.[b].} = \max_{1 \leq j \leq b} \bar{y}_{.j.}$, and
 s_v^2 is an unbiased estimate of σ^2 based on v d.f. Since by assumption there is no interaction, the sum of squares associated with s_v^2 is based on $v = abn - a - b + 1$ d.f. (If interaction is present, a different goal and method of estimating σ^2 must be adopted. See Section 4.) The constants (c_A, c_B) which depend on $((a, b), v, \alpha)$ are chosen to guarantee (2.3), and may be any solution in (c_A, c_B) of the equation (3.3) in Section 3.

3. DERIVATION OF THE CONSTANTS (c_A, c_B)

Paralleling Gupta [1965], equations (1.12) and (4.3), we have that

$$P\{CS\} = P\left\{U_{(a)} + \frac{c_A S_v}{\sigma} + \frac{\sqrt{bn} (\alpha_{[a]} - \alpha_{[i]})}{\sigma} \geq U_{(i)} \quad (1 \leq i \leq a-1)\right\}$$

and (3.1)

$$V(b) + \frac{c_B S_v}{\sigma} + \frac{\sqrt{an} (\beta_{[b]} - \beta_{[j]})}{\sigma} \geq V_{(j)} \quad (1 \leq j \leq b-1)\}$$

where the $U_{(i)}$ are associated with $\alpha_{[i]}$ ($1 \leq i \leq a$), the $V_{(j)}$ are associated with $\beta_{[j]}$ ($1 \leq j \leq b$), the $U_{(i)}$ and the $V_{(j)}$ are distributed $N(0,1)$, and the $U_{(i)}$ are distributed independently of the $V_{(j)}$ (as shown in Bechhofer [1954],

Section 4) and of S_v/σ which is distributed as $\chi_v/\sqrt{v}$. Thus,

$$P\{CS\} = \int_0^\infty \left[\int_{-\infty}^{+\infty} \prod_{i=1}^{a-1} \Phi\left(u + c_A y + \frac{\sqrt{bn} (\alpha_{[a]} - \alpha_{[i]})}{\sigma}\right) d\Phi(u) \right] \left[\int_{-\infty}^{+\infty} \prod_{j=1}^{b-1} \Phi\left(v + c_B y + \frac{\sqrt{an} (\beta_{[b]} - \beta_{[j]})}{\sigma}\right) d\Phi(v) \right] g_v(y) dy. \quad (3.2)$$

Here $\Phi(\cdot)$ is the cdf of a standard normal r.v., and $q_v(\cdot)$ is the density function of $\chi_v/\sqrt{v}$. Minimizing over the parameter space we have

$$\begin{aligned} \inf_{\substack{\alpha_1, \dots, \alpha_a \\ \beta_1, \dots, \beta_b \\ \sigma}} P\{CS\} = & \int_0^\infty \left[\int_{-\infty}^\infty \phi^{a-1}(u+c_A y) d\Phi(u) \right] \\ & \cdot \left[\int_{-\infty}^{+\infty} \phi^{b-1}(v+c_B y) d\Phi(v) \right] q_v(y) dy. \end{aligned} \quad (3.3)$$

Setting the r.h.s. of (3.3) equal to $1-\alpha$ we solve for c_A, c_B which depend on $((a,b),v,\alpha)$. Tables C.1-C.3 of Bechhofer and Dunnett [1987] give a particular solution $g_1 = c_A/\sqrt{2} = c_B/\sqrt{2}$ to five decimal places for $1-\alpha = 0.80, 0.90, 0.95$, respectively, for $a = 2(1)5$, $b = a(1)7, 10$ and $v = 5(1)30(5)50, 60(20)120, 200, \infty$.

It can be shown that g_1 is the one-sided upper equicoordinate $100(1-\alpha)$ percentage point of a central p -variate Student t -distribution based on v d.f. with a correlation matrix $P = \{\rho_{ij}\}$ where $p_1+p_2 = p$, $p_1 = a-1$, $p_2 = b-1$ and

$$\rho_{ij} = \begin{cases} 1 & \text{if } i = j \quad (1 \leq i, j \leq p) \\ 1/2 & \text{if } i \neq j \quad (1 \leq i, j \leq p_1) \\ 1/2 & \text{if } i \neq j \quad (p_1+1 \leq i, j \leq p) \\ 0 & \text{if } (1 \leq i \leq p_1, p_1+1 \leq j \leq p) \\ 0 & \text{if } (p_1+1 \leq i \leq p, 1 \leq j \leq p_1). \end{cases} \quad (3.4)$$

That is, it is the solution of the equation

$$P\left\{ \max_{1 \leq i \leq p} T_i \leq g_1 \right\} = \int_{-\infty}^{g_1} \dots \int_{-\infty}^{g_1} f_v(t_1, \dots, t_p; P) dt_1 \dots dt_p = 1-\alpha$$

where $f_v(t_1, \dots, t_p; P)$ is the p -variate Student t density function with P given by (3.4).

An abbreviated table of $g_1((a, b), v, \alpha)$ is given as Table I, below. It is abstracted from Table C.3 of Bechhofer and Dunnett [1987].

Table I
Values of g_1 for $1-\alpha = 0.95$
and selected (a, b)

$v^{-1/}$	$a = 2$ $b = 2$	$a = 2$ $b = 3$	$a = 2$ $b = 4$	$a = 3$ $b = 3$	$a = 3$ $b = 4$	$a = 4$ $b = 4$
5	2.532	2.784	2.951	2.978	3.114	3.231
6	2.417	2.647	2.799	2.823	2.946	3.052
7	2.340	2.555	2.698	2.719	2.834	2.933
8	2.285	2.490	2.626	2.645	2.754	2.848
9	2.243	2.441	2.572	2.590	2.695	2.784
10	2.211	2.403	2.530	2.547	2.648	2.735
11	2.186	2.373	2.496	2.513	2.611	2.695
12	2.164	2.348	2.469	2.485	2.581	2.663
13	2.147	2.327	2.446	2.461	2.556	2.636
14	2.132	2.310	2.426	2.442	2.535	2.613
15	2.119	2.295	2.410	2.425	2.516	2.594
16	2.108	2.282	2.396	2.410	2.501	2.577
17	2.099	2.270	2.383	2.397	2.487	2.563
18	2.090	2.260	2.372	2.386	2.475	2.550
19	2.083	2.251	2.362	2.376	2.464	2.538
20	2.076	2.243	2.354	2.367	2.454	2.528
22	2.064	2.230	2.338	2.352	2.438	2.510
24	2.055	2.219	2.326	2.339	2.424	2.496
26	2.047	2.209	2.316	2.329	2.413	2.483
28	2.040	2.201	2.307	2.320	2.403	2.473
30	2.034	2.194	2.299	2.312	2.394	2.464
35	2.022	2.180	2.284	2.296	2.378	2.446
40	2.014	2.170	2.273	2.285	2.365	2.433
60	1.993	2.147	2.247	2.258	2.337	2.403
80	1.984	2.135	2.234	2.245	2.323	2.388
∞	1.955	2.101	2.197	2.207	2.282	2.344

^{1/}Table C.3 of Bechhofer and Dunnett [1987] gives g_1 -values only for $v \geq 5$.

4. TWO-FACTOR SELECTION WITH INTERACTION PRESENT

If interaction is present in the two-factor experiment, i.e., if in place of (2.1) we have

$$Y_{ijk} = \mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \quad (4.1)$$

$$(1 \leq i \leq a, 1 \leq j \leq b, 1 \leq k \leq n)$$

with $\sum_{i=1}^a \alpha_i = \sum_{j=1}^b \beta_j = 0$ and $\sum_{i=1}^a (\alpha\beta)_{ij} = \sum_{j=1}^b (\alpha\beta)_{ij} = 0$, $(\alpha\beta)_{ij} \neq 0$ (all i, j), then the goal of the experiment should be re-defined. One possibility is

GOAL: To select a subset of the ab populations which contains the population associated with

$$\max_{i,j} \mu_{ij} = \mu_{[ab]}. \quad (4.2)$$

Then the problem reduces to the one solved by Gupta (1956) with k (representing the number of populations) replaced by ab .

Another possible goal is the following:

GOAL: To select for each $j = 1, \dots, b$ a subset of the a populations associated with $\mu_{1j}, \mu_{2j}, \dots, \mu_{aj}$ which contains the population associated with

$$\max_{1 \leq i \leq a} \mu_{ij} = \mu_{[a]j}. \quad (4.3)$$

For this goal, a correct selection will be considered to have been made when $\mu_{[a]j}$ is included in the j th subset simultaneously for all $1 \leq j \leq b$. We would wish to guarantee the same probability requirement as for the no-interaction case, which is given by (2.3), but relative to this alternative definition of a correct selection (CS). For this definition we propose the following subset selection procedure:

PROCEDURE: "Include in the j -th selected subset all populations for which

$$\bar{y}_{ij\cdot} \geq \bar{y}_{[a]j\cdot} - c^* s_v / \sqrt{n} \quad (1 \leq j \leq b)."$$
 (4.4)

Here the sum of squares associated with s_v^2 is based on $ab(n-1)$ d.f. The constant c^* required to implement this procedure depends on $((a,b),v,\alpha)$ and is chosen to satisfy the probability requirement (2.3). The value of c^* is given by the equicoordinate $100(1-\alpha)$ -percentage point of an $(a-1)b$ -variate central Student t distribution with correlation matrix $\{\rho_{ij}\}$ defined by $\rho_{ij} = 1/2$ for $i \neq j$ when $1 + (a-1)(d-1) \leq i, j \leq (a-1)d$ for $d = 1, 2, \dots, b$ and $\rho_{ij} = 0$ for $i \neq j$ otherwise. For the special case $b = 2$, the required constant can be obtained from Table C of Bechhofer and Dunnett (1987) where it is referred to as g_1 . For $b > 2$, the required tables have not been computed.

Remark 4.1: In the event that the experimenter is unsure whether or not interaction is present, he may be tempted to conduct a preliminary test for interaction. However, at the present time we can give no guidance to the experimenter as to how to employ the results of such a test.

5. GENERALIZATION TO COMPLETE MULTI-FACTOR EXPERIMENTS WITHOUT INTERACTION

The generalization to balanced multi-factor experiments with zero interaction between levels of different factors is straightforward. We illustrate the method for the 3-factor case with factors denoted by A , B and C and the number of levels for each denoted by a , b and c . Then, replacing the right-side of equation (3.3) by

$$\begin{aligned}
& \int_0^{\infty} \left[\int_{-\infty}^{+\infty} \phi^{a-1}(u+c_A y) d\Phi(u) \right] \\
& \cdot \left[\int_{-\infty}^{+\infty} \phi^{b-1}(v+c_B y) d\Phi(v) \right] \\
& \cdot \left[\int_{-\infty}^{+\infty} \phi^{c-1}(w+c_C y) d\Phi(w) \right] q_v(y) dy, \quad (5.1)
\end{aligned}$$

this is set equal to $1-\alpha$ and solved for (c_A, c_B, c_C) . It would be reasonable to set $c_A = c_B = c_C = g^*$ (say) and solve for g^* which depends on $((a, b, c), v, \alpha)$. However, no tables for g^* have been computed.

ACKNOWLEDGEMENT

This research was partially supported by the U.S. Army Research Office through the Mathematical Sciences Institute of Cornell University and by U.S. Army Research Office Contract DAAL03-86-K-0046 at Cornell University, and the Natural Sciences and Engineering Research Council of Canada at McMaster University.

REFERENCES

- Bechhofer, R.E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. Annals of Mathematical Statistics 25, 16-39.
- Bechhofer, R.E. and Dunnett, C.W. (1986). Two-stage selection of the best factor-level combination in multi-factor experiments: common unknown variance. Statistical Design: Theory and Practice, Proceedings of a Conference in Honor of Walter T. Federer, Biometrics Unit, Cornell University.
- Bechhofer, R.E. and Dunnett, C.W. (1987). Percentage points of multivariate Student t distributions. Submitted for publication to Selected Tables in Mathematical Statistics.
- Bechhofer, R.E., Dunnett, C.W. and Sobel, M. (1954). A two-sample multiple-decision procedure for ranking means of normal populations with a common unknown variance. Biometrika 41, 170-176.

Dunnett, C.W. (1955). A multiple comparisons procedure for comparing several treatments with a control. Journal of the American Statistical Association 50, 1096-1121.

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, N.C.

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

Gupta, S.S. and Sobel, M. (1957). On a statistic which arises in selection and ranking problems. Annals of Mathematical Statistics 28, 957-967.