TWO-STAGE SAMPLING FOR THE INference OF THE
DIFFERENCE OF TWO MULTIVARIATE NORMAL MEANS

by

M. Aoshima

Technical Report No. 319
August 1996

Prepared Under the Auspices
of
National Science Foundation Grant
DMS 96-26265
Ingram Olkin, Project Director

Department of Statistics
Stanford University
Stanford, California
TWO-STAGE SAMPLING FOR THE INference OF THE DIFFERENCE OF TWO MULTIVARIATE NORMAL MEANS

by

M. Aoshima

Technical Report No. 319
August 1996

Prepared Under the Auspices
of
National Science Foundation Grant
DMS 96-26265
Ingram Olkin, Project Director

Department of Statistics
Stanford University
Stanford, California
TWO-STAGE SAMPLING FOR THE INFERENCE OF THE DIFFERENCE OF TWO MULTIVARIATE NORMAL MEANS

by

Makoto Aoshima
Department of Mathematics and Informatics
Tokyo Gakugei University
Tokyo 184, Japan

Abstract

We consider Takada and Aoshima's (1996) two-stage procedure for the inference of the difference of two multivariate normal means. It is assumed that the covariance matrices of two populations are different only by unknown scalar multipliers. It is shown that Takada and Aoshima's two-stage procedure is more efficient asymptotically than Mukhopadhyay and Abid's (1986) two-stage procedure for the fixed-size confidence region problem. Numerical solutions needed in Takada and Aoshima's two-stage procedure are discussed and it is shown that those are available also for a ranking and selection problem with respect to the mean vectors under the same distribution scheme.

Key Words: asymptotic efficiency, fixed-size confidence region, ranking and selection, Takada and Aoshima's procedure
TWO-STAGE SAMPLING FOR THE INFERENCE OF THE DIFFERENCE OF TWO MULTIVARIATE NORMAL MEANS

by

Makoto Aoshima
Department of Mathematics and Informatics
Tokyo Gakugei University
Tokyo 184, Japan

1. INTRODUCTION

Suppose there are two $p$-variate normal populations $\Pi_i, i = 1, 2$ with the mean vectors $\mu_1, \mu_2$ and the covariance matrices $\sigma_1^2 H, \sigma_2^2 H$ respectively, where $\mu_1, \mu_2, \sigma_1^2$ and $\sigma_2^2$ are all unknown, but $H$ is a known $p \times p$ positive definite matrix. Let $x_1, ..., x_r$ and $y_1, ..., y_s$ be random vectors from each population. The problem to construct exactly a confidence region for $\mu = \mu_1 - \mu_2$ such that

$$P\{(T_{rs} - \mu)'H^{-1}(T_{rs} - \mu) \leq d^2\} \geq 1 - \alpha$$  \hfill (1.1)

for given $d(> 0)$ and given $\alpha(0 < \alpha < 1)$ was discussed by many researchers. It is known that there is no fixed-sample size procedure for this problem and that only two-stage procedure can enjoy exact consistency which is requested in the problem. Mukhopadhyay and Abid (1986) first gave a solution by using a two-stage procedure. However, they did not consider about the expected sum of sample sizes of each population needed the procedure to enjoy the exact consistency. Takada and Aoshima (1996) proposed a different two-stage procedure which is asymptotically 1st-order efficient in the sense of the expected sum of sample sizes of each population. It is an extension of Schwabe’s (1995) two-stage procedure which is defined in the Stein’s (1945) type II scheme to the multivariate case. Takada and Aoshima also proved that it is more economical than a two-stage procedure which is given as a multivariate version of Banerjee’s (1967) two-stage procedure in the Stein’s type I scheme in terms of the sample sizes and the coverage probability. This corresponds to the results given in Aoshima (1994). Takada and Aoshima’s two-stage procedure and their solution to the problem shall be given briefly in Section 2.

In the present paper, we first prove that Takada and Aoshima’s two-stage procedure is more efficient than Mukhopadhyay and Abid’s two-stage procedure asymptotically in Section 3. The sample size in two-stage procedures is usually defined by using a percentile point of the distribution by each two-stage sampling. Because of difficulty of the distribution which
is often written by multiple integrals, it is very important to be discussed about a numerical solution to the percentile point in the two-stage sample size. In Section 4, a numerical solution to Takada and Aoshima’s two-stage procedure is given and is compared with the upper bounds of the percentile point which were given in their paper and with the lower bound which was given in Aoshima, Hyakutake and Dudewicz (1996) for $p = 1$, by tabulating those values. Finally, in Section 5, a ranking & selection problem with respect to the mean vectors is discussed as an application of the numerical solution under the same distribution scheme.

2. TAKADA AND AOSHIMA’S PROCEDURE

Take the initial sample $x_1, ..., x_m$ and $y_1, ..., y_m$ of size $m(\geq 2)$ from each population and define the estimators of $\sigma_1^2$ and $\sigma_2^2$ by

$$U_m^2 = \sum_{i=1}^{m} (x_i - \bar{x}_m)'H^{-1}(x_i - \bar{x}_m)/\nu,$$

$$V_m^2 = \sum_{i=1}^{m} (y_i - \bar{y}_m)'H^{-1}(y_i - \bar{y}_m)/\nu,$$  

(2.1)

(2.2)

where $\bar{x}_m = \sum_{i=1}^{m} x_i/m$, $\bar{y}_m = \sum_{i=1}^{m} y_i/m$ and $\nu = p(m - 1)$. The two-stage sample size of each population is defined by

$$R = \max\{m, \left[uU_m(U_m + V_m)/d^2\right] + 1\},$$

$$S = \max\{m, \left[uV_m(U_m + V_m)/d^2\right] + 1\}$$

(2.3)

(2.4)

respectively, where $[x]$ denotes the largest integer less than $x$ and $u(>0)$ is a percentile point of the two-stage distribution defined later. Then Takada and Aoshima gave a solution to the problem by

$$\{\mu : (T_{RS} - \mu)'H^{-1}(T_{RS} - \mu) \leq d^2\},$$

(2.5)

where $T_{RS} = \bar{x}_R - \bar{y}_S$ is computed by combining the additional sample of size $R - m$ or $S - m$ to the initial sample of each population. The coverage probability is given by

$$P\{(T_{RS} - \mu)'H^{-1}(T_{RS} - \mu) \leq d^2\}$$

$$= E\left[F_p\left\{d^2\left(\frac{\sigma_1^2}{R} + \frac{\sigma_2^2}{S}\right)^{-1}\right\}\right]$$

$$\geq \min_{\alpha > 0} E\left[F_p\left\{\frac{uW_1W_2(W_1 + \phi W_2)}{\nu(W_2 + \phi W_1)}\right\}\right] \quad (= H_p(u), \text{say}),$$

(2.6)

(2.7)

where $F_p$ denotes a distribution function of the chi-square random variable with $p$ d.f., $\phi = \sigma_2/\sigma_1$, and $W_1^2$ and $W_2^2$ are independent chi-square random variables with $\nu$ d.f. The
value of \( u \) in the definitions (2.3) and (2.4) is given by solving

\[
H_p(u) = 1 - \alpha
\]

numerically for the value of \( \alpha \) given in the (1.1).

3. ASYMPTOTIC EFFICIENCY

The sample sizes in Mukhopadhyay and Abid’s (1986) two-stage procedure for the problem are defined by

\[
R' = \max\{m, \lfloor hU_m^2/d^2 \rfloor + 1\},
\]

\[
S' = \max\{m, \lfloor hV_m^2/d^2 \rfloor + 1\}
\]

for each population, where \( U_m^2 \) and \( V_m^2 \) are the same ones as in (2.1) and in (2.2). The constant \( h \) is a solution to the equation

\[
E[F_p(h(\sigma_1^2U_m^{-2} + \sigma_2^2V_m^{-2})^{-1})] = 1 - \alpha.
\]

Then a solution to the problem is given by replacing \( R \) and \( S \) with \( R' \) and \( S' \) in (2.5).

To discuss about the efficiency asymptotically, we assume that the initial sample size \( m = m(d) \) is chosen such that

\[
m(d) \to \infty, \quad d^2m(d) \to 0 \quad \text{as} \quad d \to 0.
\]

We can see from (2.7) and from (3.3) that \( u = u(m, p, \alpha) \to \chi_p^2(\alpha) \) and \( h = h(m, p, \alpha) \to 2\chi_p^2(\alpha) \) as \( d \to 0 \) under this assumption, where \( \chi_p^2(\alpha) \) denotes the upper \( \alpha \) point of the chi-square distribution. Then the following theorem is proved by using a usual way.

**Theorem 3.1** Takada and Aoshima’s two-stage procedure is more efficient than Mukhopadhyay and Abid’s two-stage procedure asymptotically, that is,

\[
\lim_{d \to 0} \frac{E(R + S)}{E(R' + S')} = \frac{1}{2} + \left( \phi + \frac{1}{\phi} \right)^{-1} \leq 1.
\]

Note that the efficiency of those is same asymptotically only when \( \sigma_1 = \sigma_2 \). Especially, when \( \phi \approx 0 \) (\( \phi \approx \infty \)), the expected sum of sample sizes of Takada and Aoshima’s two-stage procedure is asymptotically half of that of Mukhopadhyay and Abid’s two-stage procedure.
4. NUMERICAL SOLUTIONS TO $u$

Because of difficulty of obtaining directly the solution to $u$ in (2.8), Takada and Aoshima (1996) proposed two different methods to obtain upper bounds of the value of $u$. One method is available for any $p$ and an upper bound is given as a solution to the equation

$$2 \int_0^\infty F_p(\frac{ux}{\nu})(1 - F'_\nu(x))f_\nu(x)dx = 1 - \alpha$$

(4.1)

for given $m, p$ and $\alpha$, where $f_\nu(x) = F'_\nu(x)$. Another method is available only for $p = 1, 2$ and an upper bound is given as a solution to the equation

$$2F_{p, \nu}(\frac{u}{p}) - F_{p, \nu+1}(\frac{(\nu + 1)u}{\nu p}) = 1 - \alpha$$

(4.2)

for given $m, p$ and $\alpha$, where $F_{m,n}$ denotes a distribution function of the $F$ random variable with $(m, n)$ d.f. It gives better approximations to the values of $u$ than the first method for $p = 1, 2$. Table I gives those upper bounds of values of $u$ for $m = 5(1)10, 15, 30$ when $\alpha = .05$ by using (4.2) for $p = 1, 2$ and by using (4.1) for $p = 3, 4, 5$. The values in parentheses given below the line for $p = 1$ are derived by Table I in Aoshima, Hyakutake and Dudewicz (1996) which gives lower bounds of the values of $u$ when $p = 1$. They obtained the table by computing multiple integrals in (2.8) via the Monte Carlo method. Here we shall seek the value of $u$ directly from (2.8) by using a numerical approach.

<table>
<thead>
<tr>
<th>$p \backslash m$</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>15</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.22</td>
<td>8.24</td>
<td>7.17</td>
<td>6.52</td>
<td>6.07</td>
<td>5.76</td>
<td>4.95</td>
<td>4.33</td>
</tr>
<tr>
<td></td>
<td>(7.73)</td>
<td>(6.60)</td>
<td>(6.00)</td>
<td>(5.57)</td>
<td>(5.34)</td>
<td>(5.11)</td>
<td>(4.58)</td>
<td>(4.20)</td>
</tr>
<tr>
<td>2</td>
<td>10.19</td>
<td>9.12</td>
<td>8.48</td>
<td>8.05</td>
<td>7.75</td>
<td>7.53</td>
<td>6.93</td>
<td>6.42</td>
</tr>
<tr>
<td>3</td>
<td>12.80</td>
<td>11.80</td>
<td>11.17</td>
<td>10.73</td>
<td>10.41</td>
<td>10.17</td>
<td>9.47</td>
<td>8.79</td>
</tr>
<tr>
<td>5</td>
<td>15.76</td>
<td>14.90</td>
<td>14.35</td>
<td>13.95</td>
<td>13.66</td>
<td>13.43</td>
<td>12.76</td>
<td>12.09</td>
</tr>
</tbody>
</table>

5. APPLICATION

We consider a ranking & selection problem as an application of the numerical solution in Section 4. The problem is to select one population from $\Pi_i, i = 1, 2$ as the "best" in an
experimenter-specified sense with respect to the mean vectors. The $2p-$dimensional space of $M = [\mu_1, \mu_2]$ can be divided into two disjoint zones by

$$D_j = \{ M : g(M) = j \}, \quad j = 1, 2, \quad (5.1)$$

where $g(M)$ is an experimenter-specified function of $M$. The "$g(M) = j$" means that the experimenter would prefer $\Pi_j$ with $\mu_j$ as the best. We induce a distance function from $M$ to the boundary of $D_{g(M)}$ by

$$d(M) = \inf_{C \in \mathcal{D}_{g(M)}} \{ \sup_{a' H \tilde{a} = 1} |a'(\mu_1 - \mu_2 - c_1 + c_2)|^2 \}^{1/2}$$

$$= \inf_{C \in \mathcal{D}_{g(M)}} \{ (\mu_1 - \mu_2 - c_1 + c_2)' H^{-1} (\mu_1 - \mu_2 - c_1 + c_2) \}^{1/2}, \quad (5.2)$$

where $C = [c_1, c_2]$. Then the probability requirement is given by

$$P(\text{CS}) \geq P^* \text{ whenever } d(M) \geq d^* \quad (5.3)$$

for given $P^*(1/2 < P^* < 1)$ and given $d^*(> 0)$. When $\sigma_1^2$ and $\sigma_2^2$ are known,

$$P(\text{CS}) = P\{(\bar{x}, \bar{y}) \in D_{g(M)}\}$$

$$\geq P\{(\bar{x} - \bar{y}, - \mu_1 + \mu_2)' H^{-1} (\bar{x} - \bar{y}, - \mu_1 + \mu_2) \leq d^2(M)\}$$

$$\geq P\{(T_{rs} - \mu)' H^{-1} (T_{rs} - \mu) \leq d^*2\}$$

$$= F_p \{ d^*2 (\frac{\sigma_1^2}{r} + \frac{\sigma_2^2}{s})^{-1} \}. \quad (5.5)$$

Hence we obtain by using Takada and Aoshima's two-stage procedure that

$$P(\text{CS}) \geq E \left[ F_p \left\{ d^*2 (\frac{\sigma_1^2}{R} + \frac{\sigma_2^2}{S})^{-1} \right\} \right] \quad (5.6)$$

when $\sigma_1^2$ and $\sigma_2^2$ are unknown. We can see that the right hand side of the above is the same one as in (2.6). The discussions in Sections 2 and 4 are available and the numerical solution given in Section 4 is useful also for this problem.

ACKNOWLEDGMENTS

Research of this author was supported by Grant-in-Aid for Encouragement of Young Scientists, the Ministry of Education, Science and Culture, Japan, under Contract Number 721-7025-08780217.
REFERENCES

Aoshima, M. (1994). In heteroscedastic simultaneous inference, the heteroscedastic method II (based on Healy's two-stage sampling) is more economical than the original HM (which is based on Chatterjee's sampling). *Amer. J. Math. Manage. Sci.*, 14, 167–196.


