ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
XIV, 1 (1974)

E. J. DUDEWICZ (Columbus, Ohio)

A NOTE ON SELECTION PROCEDURES WITH UNEQUAL OBSERVATION NUMBERS *

1. Introduction. In a recent article [6] Sitek generalized a selection procedure ** of Gupta and Sobel [5] to the case of unequal observation numbers. Unfortunately, as we shall show in this paper, Sitek's derivation is not correct. An alternative approach, recently given by Dudewicz and Dalal [3], is presented for the same problem. (This new approach has certain superior properties in comparison with that of Gupta [4]. However, it also does not yet cover the case of unequal observation numbers, which is apparently a very difficult problem.) Some suggestions for further work (numerical as well as analytical) on the case of unequal observation numbers are made.

2. Sitek's method for unequal observation numbers. In order to clarify the subtleties which invalidate Sitek's method, it will be helpful if we first state the problem clearly. We have k ($k \geqslant 2$) sources of observations (called *populations*) $\pi_1, \ldots, \pi_k$. Observations from π_i (source i) are normal random variables with mean μ_i and variance σ_i^2 ($1 \leqslant i \leqslant k$), and all observations are independent. Let

$$\mu_{[1]} \leqslant \mu_{[2]} \leqslant \cdots \leqslant \mu_{[k]} \tag{1}$$

denote the (unknown) $\mu_1, \ldots, \mu_k$ in numerical order. Our goal is (based on n_i observations from π_i, $i = 1, 2, \ldots, k$) to select a subset S of $\Pi = \{\pi_1, \ldots, \pi_k\}$ such that with probability at least P^* ($1/k < P^* < 1$) a population with mean $\mu_{[k]}$ is in S. Let $\pi_{(i)}$ denote the population with mean $\mu_{[i]}$, and let $n_{(i)}$ and $\bar{X}_{(i)}$ denote (respectively) the number of observations and sample mean of observations from $\pi_{(i)}$ ($1 \leqslant i \leqslant k$). Then, if $\mathscr{P}$ is any procedure for selecting a subset $S \subseteq \Pi$, our *probability requirement* is that

$$\mathrm{P}(CS \mid \mathscr{P}) \geqslant P^* \tag{2}$$

---

* Supported by the U. S. Army Research Office — Durham.

** This procedure is actually due to Gupta [4].

(CS denotes the event "$\pi_{(k)} \in S$") for all $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$. Since (2) will be clearly satisfied if

$$\inf_{\boldsymbol{\mu}} \mathrm{P}(CS|\mathscr{P}) = P^*, \tag{3}$$

one usually tries to develop a procedure $\mathscr{P}$ in such a way that (3) is satisfied.

Let $\bar{X}_i$ be the sample mean of the n_i observations from π_i $(1 \leqslant i \leqslant k)$, and let

$$\bar{X}_{[1]} \leqslant \bar{X}_{[2]} \leqslant \ldots \leqslant \bar{X}_{[k]} \tag{4}$$

denote $\bar{X}_1, \ldots, \bar{X}_k$ in numerical order, assume $\sigma_1^2 = \ldots = \sigma_k^2 = \sigma^2$ with σ^2 unknown, let s_ν^2 be the usual estimator of σ^2/ν with ν degrees of freedom, and let N be the n_j of that population which yielded the largest sample mean $\bar{X}_{[k]}$. Then Sitek suggests the procedure

$$R\colon \text{ Put } \pi_i \in S \text{ iff } \bar{X}_i \geqslant \bar{X}_{[k]} - qs_\nu\sqrt{1/n_i + 1/N}, \tag{5}$$

where q is a percentage point of a multivariate t-distribution. (The point q is approximated by Sitek in her Section 5.) Unfortunately, Sitek's "proof" that

$$\inf_{\boldsymbol{\mu}} \mathrm{P}(CS|R) \tag{6}$$

equals P^* is incorrect, as we shall now show. We have

$$\begin{aligned} \mathrm{P}(CS|R) &= \mathrm{P}[\bar{X}_{(k)} \geqslant \bar{X}_{[k]} - qs_\nu\sqrt{1/n_{(k)} + 1/N}] \\ &= \mathrm{P}[\bar{X}_{(k)} \geqslant \bar{X}_{(i)} - qs_\nu\sqrt{1/n_{(k)} + 1/N},\ i = 1, \ldots, k-1] \\ &= \mathrm{P}\left[\frac{(\bar{X}_{(i)} - \bar{X}_{(k)}) - (\mu_{[i]} - \mu_{[k]})}{s_\nu\sqrt{1/n_{(k)} + 1/N}} \leqslant q + \frac{\mu_{[k]} - \mu_{[i]}}{s_\nu\sqrt{1/n_{(k)} + 1/N}},\quad i = 1, \ldots, k-1\right]. \end{aligned} \tag{7}$$

However, it is not clear (as Sitek implies in lines 12-19 of p. 359) that (7) is minimized when $\mu_{[1]} = \ldots = \mu_{[k]}$, since N is a random variable dependent upon $\bar{X}_{(1)}, \ldots, \bar{X}_{(k)}$. For any i $(1 \leqslant i \leqslant k)$,

$$\begin{aligned} \mathrm{P}[N = n_{(i)}] &= \mathrm{P}[\bar{X}_{(i)} = \max(\bar{X}_{(1)}, \ldots, \bar{X}_{(k)})] = \mathrm{P}[\bar{X}_{(j)} < \bar{X}_{(i)}, j \neq i] \\ &= \int_{-\infty}^{\infty} \left[\prod_{j \neq i} \Phi\left(\sqrt{n_{(j)}/n_{(i)}}\,x + \frac{\mu_{[i]} - \mu_{[j]}}{\sigma/\sqrt{n_{(j)}}}\right)\right] \varphi(x)\,dx, \end{aligned} \tag{8}$$

where $\Phi(\cdot)$ and $\varphi(\cdot)$ are the distribution function and density function of a normal random variable with mean zero and variance one. *Even if* one assumes (7) to be minimized when $\mu_{[1]} = \ldots = \mu_{[k]}$, one finds that infimum (6) to be equal to

$$
\begin{aligned}
(9)\quad & \mathrm{P}_{\mu_{[1]}=\ldots=\mu_{[k]}}(CS\,|\,R) \\
&= \mathrm{P}\left[\frac{\bar{X}_{(i)}-\bar{X}_{(k)}}{s_\nu\sqrt{1/n_{(k)}+1/N}} \leqslant q,\ i=1,\ldots,k-1\right] \\
&= \mathrm{P}\left[\frac{\bar{X}_{(i)}-\bar{X}_{(k)}}{s_\nu\sqrt{1/n_{(i)}+1/n_{(k)}}} \leqslant q\sqrt{\frac{1/n_{(k)}+1/N}{1/n_{(k)}+1/n_{(i)}}},\ i=1,\ldots,k-1\right] \\
&= \mathrm{P}\left[T_i \leqslant q\sqrt{\frac{1/n_{(k)}+1/N}{1/n_{(k)}+1/n_{(i)}}},\ i=1,\ldots,k-1\right],
\end{aligned}
$$

where $(T_1, \ldots, T_{k-1})$ has the multivariate t-distribution but with correlation matrix (ϱ_{ij}) given by

$$
(10)\qquad \varrho_{ij} = \frac{1}{\sqrt{(1+n_{(k)}/n_{(i)})(1+n_{(k)}/n_{(j)})}}.
$$

Sitek gave (10) with $n_{(k)}$ replaced by N, which is incorrect. Now (9) cannot be evaluated since $n_{(1)}, \ldots, n_{(k)}$ are not known: knowledge of the $n_{(i)}$'s implies knowledge of which population has each mean $\mu_{[i]}$ $(1 \leqslant i \leqslant k)$. If we knew this, no experiment would be necessary.

3. Another method for $\sigma_1^2, \ldots, \sigma_k^2$ unequal. In Section 2 we saw that Sitek's attempt to generalize Gupta's procedure R (to the case of unequal observations) was unsuccessful. Even had it succeeded, it would still have assumed $\sigma_1^2 = \ldots = \sigma_k^2 = \sigma^2$ with σ^2 unknown. While this homoscedasticity assumption is sometimes valid, often treatments are sufficiently diverse in character that their variances are substantially unequal. For this situation Dudewicz and Dalal [3] propose the procedure

$$
(11)\qquad \mathscr{P}_E\colon \text{Put } \pi_i \in S \ \text{ iff } \ \tilde{\bar{X}}_i \geqslant \tilde{\bar{X}}_{[k]} - d,
$$

and they show that $\mathrm{P}(CS\,|\,\mathscr{P}_E)$ is independent of $\sigma_1^2, \ldots, \sigma_k^2$ and that

$$
(12)\qquad \inf_{\mu} \mathrm{P}(CS\,|\,\mathscr{P}_E) = P^*.
$$

The details of their procedure are as follows. Take an initial sample of size n_0 $(n_0 \geqslant 2)$ $X_{i1}, \ldots, X_{in_0}$ from π_i, and write

$$
(13)\qquad \bar{X}_i(n_0) = \sum_{j=1}^{n_0} X_{ij}/n_0, \qquad s_i^2 = \sum_{j=1}^{n_0} (X_{ij} - \bar{X}_i(n_0))^2/(n_0-1),
$$

$$
(14)\qquad n_i = \max\left\{n_0+1,\ \left[\left(\frac{s_i h}{d}\right)^2\right]\right\},
$$

where $h = h_k(P^*)$ is the unique solution of the equation

$$\int_{-\infty}^{\infty} \big(F_{n_0}(z+h)\big)^{k-1} f_{n_0}(z)\,dz = P^*, \tag{15}$$

where $F_{n_0}(\cdot)$ and $f_{n_0}(\cdot)$ are, respectively, the distribution function and density function of a Student t random variable with $n_0-1 \geqslant 1$ degrees of freedom, and $[y]$ denotes the smallest integer not less than y $(i = 1, \ldots, k)$. Take $n_i - n_0$ additional observations $X_{i,n_0+1}, \ldots, X_{in_i}$ from π_i, and write

$$\tilde{\tilde{X}}_i = \sum_{j=1}^{n_i} a_{ij} X_{ij} \quad (1 \leqslant i \leqslant k), \tag{16}$$

where the a_{ij}'s $(j = 1, \ldots, n_i;\ i = 1, \ldots, k)$ are any numbers such that

$$\sum_{j=1}^{n_i} a_{ij} = 1, \quad a_{i1} = \ldots = a_{in_0}, \quad s_i^2 \sum_{j=1}^{n_i} a_{ij}^2 = (d/h)^2. \tag{17}$$

The procedure $\mathscr{P}_E$ also has the property of monotonicity. An additional feature of $\mathscr{P}_E$ is that it satisfies (12) (the probability requirement) irrespective of the prior choice of $d > 0$. This allows one to choose d to make the expected size of the selected subset $E\big(\#(S)\big)$, suitably small in any specified configuration $\mu_{[1]}, \ldots, \mu_{[k]}$ (e. g. $\mu_{[1]} = \ldots = \mu_{[k-1]} = \mu_{[k]} - \delta^*$ for some $\delta^* > 0$). Tables and graphs to allow easy implementation of this approach are under development by Dudewicz and Chen [2].

4. General comments on unequal observation numbers. As we have seen, selection problems with unequal observations are inherently very complex due to the fact that in such situations one does not know the association between $n_{(1)}, \ldots, n_{(k)}$ and $\pi_1, \ldots, \pi_k$. Even in the earliest work on selection problems, Bechhofer [1] faced a related problem (see his p. 24) but was unable to resolve it other than for $k = 2$ populations, and that was when assuming $\sigma_1^2, \ldots, \sigma_k^2$ were known. Dudewicz and Dalal [3] would have liked to allow different initial sample sizes $n_{01}, \ldots, n_{0k}$ but were unable to do so in general. In our opinion, the problem definitely merits consideration because of its practical importance. Useful methods may be: (1) numerical solution for "typical" cases to check out conjectures about actual or approximate solutions (e. g. one might conjecture that for suitably high P^* one can obtain an approximate lower bound on $\mathrm{P}(CS)$ in most procedures by assuming a common sample size $n = \min(n_1, \ldots, n_k)$); and (2) analytical study *via* bounds (e. g. from the Bonferroni or Ljapunov Inequalities) on $\mathrm{P}(CS)$.

References

[1] R. E. Bechhofer, *A single-sample multiple decision procedure for ranking means of normal populations with known variances*, Ann. Math. Statist. 25 (1954), p. 16-39.

[2] E. J. Dudewicz and H. J. Chen, *Subset selection of normal populations with unknown unequal variances*, paper in preparation.

[3] E. J. Dudewicz and S. R. Dalal, *Allocation of observations in ranking and selection with unequal variances*, submitted for publication. (Abstract, *Optimizing methods in statistics*, edited by J. S. Rustagi, Academic Press Inc., New York 1971, p. 471-474.)

[4] S. S. Gupta, *On a decison rule for a problem in ranking means*, Institute of Statistics, Mineograph Series No. 150 (May 1956), University of North Carolina, Chapel Hill, North Carolina.

[5] — and M. Sobel, *On a statistic which arises in selection and ranking problems*, Ann. Math. Statist. 28 (1957), p. 957-967.

[6] M. Sitek, *Application of the selection procedure R to unequal observation numbers*, Zastosow. Matem. 12 (1972), p. 355-371.

DIVISION OF STATISTICS
THE OHIO STATE UNIVERSITY
COLUMBUS, OHIO 43210, U.S.A.

Received on 29. 11. 1972

E. J. DUDEWICZ (Columbus, Ohio)

UWAGI O ZASADACH WYBORU PRZY NIEJEDNAKOWYCH LICZEBNOŚCIACH OBSERWACJI

STRESZCZENIE

W pracy [6] Sitek uogólniła zasadę wyboru podaną w [5] na przypadek niejednakowych liczebności obserwacji. W tej nocie autor wykazuje, że rozumowanie Sitek nie jest poprawne. Autor przedstawia inne podejście do tego zagadnienia, opublikowane wcześniej w [3]. Nota zawiera także sugestie dotyczące dalszych, zarówno numerycznych, jak i analitycznych, badań nad tym problemem.