NOTES ON SOBEL'S INDIFFERENCE ZONE APPROVAL

TO A SELECTION PROBLEM

by

D. J. Daley

October 1987

Mimeo Series #1732

DEPARTMENT OF STATISTICS
Chapel Hill, North Carolina

NOTES ON SOBEL'S INDIFFERENCE ZONE APPROACH TO A SELECTION PROBLEM

D. J. Daley

Department of Statistics (IAS)
Australian National University
and
Statistics Department
University of North Carolina at Chapel Hill
(Visiting)

A simplified explicit notation is used to describe an approach to a selection problem involving k populations with unknown means $\mu_1, \ldots, \mu_k$. The task is the identification of the subset Ω_G of $\Omega \equiv \{1,\ldots,k\}$ comprised of the indices with the t largest means. The Strong Law of Large Numbers establishes the acceptability of a modification to the notion of "correct selection". Combinatorial considerations and approximations indicate that the Equal Parameter Configuration does not necessarily yield a lower bound on the probability of this modified event of correct selection.

§1 Introduction

This is a self-contained note arising as an amplification and extension of a contribution to the discussion of a paper by Sobel (1987) given at a workshop on Inference Procedures in Ranking and Selection held in Sydney in August 1987. It presents a notation that I found useful in focussing attention on the main results in the paper and in giving the solution to a problem that Sobel indicated.

We are given k populations that are indexed by $\Omega = \{1,\dots,k\}$. Population i is associated with independent identically distributed random variables with unknown mean μ_i. The general question is the following:

> On the basis of the vector $\bar{X}_n$ of means of n i.i.d. r.v.s from each of the k populations, identify the set Ω_G of populations with the t largest means.

Mnemonically, the suffix G stands for "Good": the complementary set can be regarded as the "Bad" set,

$$\Omega_B \equiv \Omega \setminus \Omega_G = \{i: \mu_i \leq \mu_{(k-t)}\},$$

where the usual order statistics notation $\mu_{(i)}$ is used to denote the i^{th} smallest component of the vector μ. Strictly, and usefully, we should and shall write $\Omega_G(\mu)$ and $\Omega_B(\mu)$, as these subsets of $\Omega = \{1,\dots,k\}$ are defined in terms of the components of any k-vector, and not just μ.

For any $z \in \mathbb{R}^k$, let $R(i;z)$ denote the index of the i^{th} smallest component of z, with an appropriate modification in the case of any coincident components, a situation which will not concern us. Mnemonically, $R(.;z)$ is the Rank Index function, and for componentwise distinct z, it is a one-one function of the finite set Ω onto itself. Clearly $z_{R(i;z)} = z_{(i)}$, and

$$\Omega_G(z) = \{R(i;z): i = k-t+1, \dots, k\}. \tag{1.1}$$

We shall also have need of the possibly smaller sets, defined for any non-negative c and d by

$$\Omega_G(z;c) \equiv \{i\colon z_i \geq z_{(k-t)} + c\} \quad \text{and} \quad \Omega_B(z;d) = \{i\colon z_i \leq z_{(k-t)} - d\},$$

observing that each of these set functions is ordered with respect to set inclusion for monotonic increasing c and d.

In order to be definite and simplify discussion, assume that the vector $\mu = (\mu_1, \ldots, \mu_k)$ has mutually distinct components, and that the r.v.s have a common continuous distribution apart from the location parameters μ_i.

A most pertinent observation is that when μ has distinct components, the Strong Law of Large Numbers implies that

$$R(i;\bar{X}_n) \longrightarrow_{a.s.} R(i;\mu) \qquad (n \longrightarrow \infty). \tag{1.2}$$

Indeed, for any collection of well-defined location parameters, we should have the sample location parameters converging to the population parameters, and the appropriate analogue of (1.2) would hold.

The important consequence of (1.2) is that with probability one, the event of "correct selection", which is identified with $\{\Omega_G(\bar{X}_n) = \Omega_G(\mu)\}$, occurs for sufficiently large n, subject only to the strictness of the inequality $\mu_{(k-t)} < \mu_{(k-t+1)}$. In practice, more interest tends to focus on the probability of this event, i.e., on

$$P(CS) = P(CS|\mu) \equiv \Pr\{\Omega_G(\bar{X}_n) = \Omega_G(\mu)\} \ . \tag{1.3}$$

The calculation of (1.3) with specific distributions leads to various computational problems which are not of concern here. There is one general result available, as follows from a simple combinatorial argument, that for fixed n,

$$P(CS|\mu) \longrightarrow 1/\begin{bmatrix} k \\ t \end{bmatrix} \tag{1.4}$$

as μ converges (e.g. in terms of euclidean distance) towards any point of the Equal Configuration Set

$$\mathrm{EPC} \equiv \{z \in \mathbb{R}^k\colon z_{(1)} = \dots = z_{(k)}\}. \tag{1.5}$$

It is arguable that it is desirable to have a definition of "correct selection" that is not affected by proximity of μ to the EPC in the sense that for such μ, any subset should then constitute an acceptable selection. Sobel's paper builds on earlier work with Chen (1987a, b) in using the idea of the acceptability in practice of knowing that correct identification occurs for the populations with means μ_i exceeding the best of the "Bad" set by δ say, i.e., that

$$\Omega_G(\bar{X}_n) \supseteq \Omega_G(\mu;\delta). \tag{1.6}$$

We shall now paraphrase Sobel's notion of what "correct selection" entails.

§2. Modified definition of "correct selection"

At the outset it is helpful to distinguish between the idea of correct selection in terms of the parameters μ of the model on the one hand, and (Sobel's) Procedure R as a practical tool using the mean vector $\bar{X}_n$ on the other.

In terms of the model, choose $\delta > 0$ and define a subset of $\mathbb{R}^k$ called the Preference Zone

$$\mathrm{PZ}(\delta) \equiv \{z \in \mathbb{R}^k\colon z_{(k-t+1)} > z_{(k-t)} + \delta\}. \tag{2.1}$$

Observe that $\Omega_G(\mu;\delta) = \Omega_G(\mu)$ for $\mu \in \mathrm{PZ}(\delta)$, so that in terms of $\bar{X}_n$ we should try and identify all t populations as "Good". As the complement of $\mathrm{PZ}(\delta)$ define the Indifference Zone

$$\mathrm{IZ}(\delta) \equiv \mathbb{R}^k \setminus \mathrm{PZ}(\delta) = \{z \in \mathbb{R}^k\colon z_{(k-t+1)} \leq z_{(k-t)} + \delta\}, \tag{2.2}$$

with the idea of still being concerned to identify all the populations of $\Omega_G(\mu;\delta)$ but of accepting a larger set that should contain the rest of Ω_G as a subset. (For later use, partition the parameter space $\mathbb{R}^k$ into the t+1 zones

$$IZ_j(\delta) \equiv \{z \in \mathbb{R}^k\colon z_{(k-j)} \leq z_{(k-t)} + \delta < z_{(k-j+1)}\} \quad (j = 0,\dots,t), \tag{2.3}$$

defining $z_{(k+1)} = \infty$. Observe that then $PZ(\delta) = IZ_t(\delta)$, and that $\Omega_G(\mu;\delta)$ contains exactly j points for every $\mu \in IZ_j(\delta)$.)

In terms of $\bar{X}_n$, a procedure based on $\{\bar{X}_n \in PZ(\delta)\}$ can be expected to be too stringent to identify correctly the subset $\Omega_G(\mu;\delta)$ for every $\mu \in PZ(\delta)$. On the other hand, using $\Omega_G(\bar{X}_n)$ would represent no change from the original formulation. These extremes suggest as a compromise the use of a set $\Omega_G(\bar{X}_n;c)$ for some c in $0 < c < \delta$. Similarly a more broadly based set than $\Omega_G(\bar{X}_n) \setminus \Omega_G(\bar{X}_n;c)$ is needed to reflect any weakening of the identification constraint from (1.3), so introduce the sample Indifference set, defined for non-negative c and d (still to be prescribed) by

$$\Omega_I(\bar{X}_n;c,d) \equiv \Omega \setminus (\Omega_G(\bar{X}_n;c) \cup \Omega_B(\bar{X}_n;d)). \tag{2.4}$$

Sobel's Procedure R defines "correct selection" as the intersection of the events

$$\{\Omega_G(\bar{X}_n;c) \supseteq \Omega_G(\mu;\delta)\} \tag{2.5}$$

and

$$\{(\Omega_G(\bar{X}_n;c) \cup \Omega_I(\bar{X}_n;c,d)) \supseteq \Omega_G(\mu)\}, \tag{2.6'}$$

equivalently

$$\{\Omega_B(\bar{X}_n;d) \subseteq \Omega_B(\mu)\}. \tag{2.6"}$$

We shall use the notation CS_{Imod} to describe explicitly this Indifference set modification to the event of correct selection, so that

$$CS_{Imod} \equiv \{\Omega_G(\bar{X}_n;c) \supseteq \Omega_G(\mu;\delta)\} \cap \{\Omega_B(\bar{X}_n;d) \subseteq \Omega_B(\mu)\}. \tag{2.7}$$

As for $P(CS|\mu)$, we note explicitly the dependence of the probability of correct selection as just defined on μ, though it does indeed depend also on n, c, d and δ: we write $P(CS_{Imod}|\mu)$ for what Sobel designated $P(CS_1|PZ)$ for $\mu \in PZ(\delta)$ and $P(CS_2|IZ)$ for $\mu \in IZ(\delta)$.

It follows from (1.2) that, provided $0 < c < \delta$ and $d > 0$, $\Pr\{\Omega_G(\bar{X}_n;c) \supseteq \Omega_G(\mu;\delta)\} \longrightarrow 1$ and $\Pr\{\Omega_B(\bar{X}_n;d) \subseteq \Omega_B(\mu)\} \longrightarrow 1$ for $n \longrightarrow \infty$, such convergence being uniform over $\mu \in \mathbb{R}^k$ conditional on fixed $c < \delta$ and $d > 0$. This is a more general result than Sobel's demonstration in the case of normally distributed r.v.s (and hence, $\bar{X}_n$ is normally distributed) that, for given δ, P_1 and P_2, there exist positive c in $0 < c < \delta$, positive δ, and an integer n_0 such that for all $n \geq n_0$,

$$P(CS_{Imod}|\mu) \geq P_1 \quad (\text{all } \mu \in PZ(\delta)) \quad \text{and} \quad P(CS_{Imod}|\mu) \geq P_2 \quad (\text{all } \mu \in IZ(\delta)).$$

The probability of correct selection is computable more simply for $\mu \in PZ(\delta)$ than $\mu \in IZ(\delta)$, because in the former case $CS_{Imod} = \{\Omega_G(\bar{X}_n;c) = \Omega_G(\mu;\delta)\}$ as the latter set coincides with $\Omega_G(\mu)$ and the former set has at most t elements.

§3. Is the EPC the worst case in the IZ?

In terms of specific computations and obtaining bounds on $P(CS_{Imod})$ it is common to look at points on the "boundaries" of various regions. The uniformity property at the end of the last section is heavily dependent on the use of $c < \delta$ and $d > 0$. This will become more apparent as we elucidate the remark at the end of the Introduction to Sobel's paper about there being "open questions associated with [whether] the Equal Parameter Configuration is the worst case in the Indifference Zone for $t > 1$".

The study of "Worst Cases" in ranking and selection problems investigates e.g. $P(CS)$ when the parameter μ takes values such as $\mu \in EPC$. Further sets of parameter values are appropriate in this context: define for each $j = 0,\ldots,t$ the Slippage Parameter Configurations $SPC_j(\delta)$ consisting of $z \in \mathbb{R}^k$ for which

$$z_{(i)} = \begin{cases} z_{(k-t)} & \text{for } i = 1,\dots,k-t, \\ z_{(k-t)} + \delta - 0 & \text{for } i = k-t+1,\dots,k-j, \\ z_{(k-t)} + \delta & \text{for } i = k-j+1,\dots,k; \end{cases}$$

and the Equal Parameter Configurations $EPC_j(\delta)$ consisting of z for which

$$z_{(i)} = \begin{cases} z_{(k-t)} & \text{for } i = 1,\dots,k-j, \\ z_{(k-t)} + \delta & \text{for } i = k-j+1,\dots,k. \end{cases}$$

It is clear that these points are on the boundary of $IZ_j(\delta)$ and are the closest such points to $SPC_t(\delta)$ and $EPC_0(\delta)$ respectively. The EPC set at (1.4) coincides with $EPC_0(\delta)$.

For $\mu \in IZ(\delta)$ arbitrarily close to $EPC_j(\delta)$, and δ large in relation to the precision of $\bar{X}_n$, we shall have $\Omega_G(\bar{X}_n;\delta-\epsilon_n)$ for large enough n and small enough ϵ_n almost certainly containing $\Omega_G(\mu;\delta)$, so that for such μ, $P(CS_{Imod}|\mu) \approx Pr\{\Omega_B(\bar{X}_n;d) \subseteq \Omega_B(\mu)\}$, which for $d = 0+$ makes $P(CS_{Imod}|\mu)$ arbitrarily close to $1/\begin{bmatrix} k-j \\ t-j+1 \end{bmatrix}$. On the other hand, for μ sufficiently close to $SPC_j(\delta)$, the inclusion relation for Ω_B sets in $P(CS_{Imod}|\mu)$ will almost certainly be satisfied but the set $\Omega_G(\bar{X}_n;\delta-\epsilon_n)$ will contain about j points; then $P(CS_{Imod}|\mu) \approx 1/\begin{bmatrix} t \\ j \end{bmatrix}$. Elementary arithmetic shows that there exist integers $k > t > j > 0$ such that the latter of these approximations may be smaller or larger than $1/\begin{bmatrix} k \\ t+1 \end{bmatrix}$ which is the limit of $P(CS_{Imod}|\mu)$ for $\mu \in IZ(\delta) \longrightarrow EPC_0(\delta)$ (it is only the combinatorics of the Ω_B sets that determine this probability for such μ).

REFERENCES

CHEN, Pinyuen and SOBEL, Milton (1987a) An integrated formulation for selecting the t best of k normal populations. Comm. Stat. 16, 121-146.

----- (1987b) A new formulation for the multinomial selection problem. Comm. Stat. 16, 147-180.

SOBEL, M. (1987) A new formulation for selecting the t best of k normal populations. Technical Report No.23, Program in Statistics and Applied Probability, University of California, Santa Barbara.