Multistage Methodologies for Partitioning a Set of Exponential Populations

Tumulesh K. S. Solanky

Department of Mathematics,
University of New Orleans,
2000 Lakefront,
New Orleans, LA 70148, USA
tsolanky@uno.edu

Abstract. We consider the problem of partitioning a set of k exponential populations with respect to a control population. For this problem some multistage methodologies are proposed and their theoretical properties are derived. Using the Monte Carlo simulation techniques, the small and moderate sample size performance of the proposed procedure are studied.

Keywords. Correct partition, Probability of correct decision, sequential procedure, simulations, two-parameter negative exponential populations.

1 Introduction

The problem of comparisons with a control has been an active research area for over seven decades. The desire of the experimenter to have the best population to be some "specified amount better" than what is already in use has motivated the research in this area. This problem has been studied by many and under different formulations. Tong (1969) formulated the partition problem using Bechhofer's (1954) indifference zone formulation and constructed a two-stage and a purely sequential procedure. Among slightly different formulations of this problem is the problem of selection of the best treatment relative to a control population or selecting a subset of the treatments having means greater than that of a control. For more on such related formulations one is recommended Bechhofer, Santner and Goldsman (1995). For a general overview of sequential methodology and of the partition problem, the reader is recommended Ghosh, Mukhopadhyay and Sen (1997), and, Mukhopadhyay and Solanky (1994).

Suppose that we have $\pi_0, \pi_1, \ldots, \pi_k$, independent and exponentially distributed populations, with density function of $\pi_i,\ i = 0, 1, \cdots, k$ given by $f_X(x) = \sigma^{-1} exp\{-(x-\theta_i)/\sigma\} I(x > \theta_i)$. Assume that the location parameters $\theta_i,\ i = 0, 1, \cdots, k$ and the common scale parameter σ are all unknown. We refer to π_0 as the control population. The goal is to partition the set of treatments $\Omega = (\pi_i : i = 1, 2, \cdots, k)$, into two disjoint and exhaustive subsets, corresponding to "Good" and "Bad" populations compared to the control population with a pre specified probability of correct partition. Given arbitrary but fixed constants δ_1 and δ_2, $\delta_1 < \delta_2$, we define three subsets of Ω along the lines of Bechhofer's (1954) indifference-zone formulation, as:

$$\begin{aligned} \Omega_L &= \{\pi_i : \theta_i \leq \theta_0 + \delta_1,\ i = 1, \cdots, k\}, \\ \Omega_M &= \{\pi_i : \theta_0 + \delta_1 < \theta_i < \theta_0 + \delta_2,\ i = 1, \cdots, k\}, \\ \Omega_R &= \{\pi_i : \theta_i \geq \theta_0 + \delta_2,\ i = 1, \cdots, k\}. \end{aligned} \tag{1}$$

We refer to Ω_R as the set of "good populations" and Ω_L as the set of "bad populations". The set Ω_M would be referred to as the set of "mediocre populations". Adopting the Bechhofer's indifference zone approach, we are interested in the correct partition of the populations in Ω_R and Ω_L. And, we will be indifferent to partition of populations in Ω_M. That is, with high accuracy we want to partition the set Ω into two disjoint subsets P_L and P_R, such that, $\Omega_L \subseteq P_L$ and $\Omega_R \subseteq P_R$. Such a partition is known in the literature as a *correct decision* (CD). In other words, given a pre assigned number P^*, $2^{-k} < P^* < 1$, we seek statistical methodologies $\wp$ to determine P_L and P_R, such that

$$P\{CD|\boldsymbol{\theta}, \sigma^2, \wp\} \geq P^* \qquad \forall\ \boldsymbol{\theta} \in \boldsymbol{R}^{k+1},\ \sigma \in \boldsymbol{R}^+. \tag{2}$$

We will use the following notation in the rest of this paper for convenience:

$$d = (\delta_1 + \delta_2)/2, \quad a = (-\delta_1 + \delta_2)/2, \quad \lambda = \sigma/a, \text{ and,}$$
$$r = \begin{cases} k/2 & \text{if } k \text{ is even;} \\ (k+1)/2 & \text{if } k \text{ is odd.} \end{cases} \tag{3}$$

2 Known σ case

We assume that we observe random variables $X_{0j}, X_{1j}, \cdots, X_{kj}$ from $\pi_0, \pi_1, \cdots, \pi_k,\ j = 1, \cdots, n$, respectively, in a sequential framework, where n is to be determined below. Assuming that σ is known, we denote

$$T_i = min_{1 \le j \le n}(X_{ij}),\ i = 0, 1, \cdots, n. \tag{4}$$

Along the lines of Desu et al. (1977), we define a pooled estimator of σ as

$$\hat{\sigma} = \Sigma_{i=0}^{k} \Sigma_{j=1}^{n} (X_{ij} - T_i)/((k+1)(n-1)). \tag{5}$$

Note that $2(k+1)(n-1)\hat{\sigma}/\sigma$ is independent of T_i and has a chi-squared distribution with $2(k+1)(n-1)$ degrees of freedom. Next, we Consider the decision rule $\wp$ defined as:

$$\begin{aligned} P_L &= \{\pi_i :\ T_i - T_0 < d,\ i = 1, \cdots, k\}, \\ P_R &= \{\pi_i :\ T_i - T_0 > d,\ i = 1, \cdots, k\}. \end{aligned} \tag{6}$$

Next, observe that for a mean vector $\boldsymbol{\theta}$ to be a least favorable configuration under $\wp$, the set Ω_M must be empty, and, all the populations in Ω_L and Ω_R must have common means $\theta_0 + \delta_1$ and $\theta_0 + \delta_2$, respectively. Let $\boldsymbol{\theta}^0(r')$ be the configuration such that $\theta_i = \mu_0 + \delta_2$ and $\theta_j = \mu_0 + \delta_1$, $0 < i \le r'$, $r' < j \le k$ for some r' such that $0 < r' \le k$. Then, we have

$$P\Big[CD|\boldsymbol{\theta}^0(r'),\, \sigma,\, \wp\Big]$$

$$= P\Big[T_j - T_0 < d,\, T_i - T_0 > d,\, 0 < i \le r',\, r' < j \le k\Big],$$

$$= P\Big[\tfrac{T_j - \theta_j}{\sigma/n} - \tfrac{T_0 - \theta_0}{\sigma/n} < \tfrac{d - \theta_j + \theta_0}{\sigma/n},\ \tfrac{T_i - \theta_i}{\sigma/n} - \tfrac{T_0 - \theta_0}{\sigma/n} > \tfrac{d - \theta_i + \theta_0}{\sigma/n},\ 0 < i \le r',\, r' < j \le k\Big].$$

Next, we write $Y_j = \frac{T_j - \theta_j}{\sigma/n}, r' < j \le k,\ Y_i = \frac{T_i - \theta_i}{\sigma/n},\ 0 < i \le r',\ Y_0 = \frac{T_0 - \theta_0}{\sigma/n}$. Note that $Y_i,\ Y_j$ and $Y_0, 0 < i \le r',\, r' < j \le k$ all have standard exponential distributions. Note that for $r' < j \le k$, $\frac{d - \theta_j + \theta_0}{\sigma/n} \le \frac{d - \delta_1}{\sigma/n}$, and for $0 < i \le r$, $\frac{d - \theta_i + \theta_0}{\sigma/n} \ge \frac{d - \delta_2}{\sigma/n}$. Next using the notations from (3), we have

$$P\Big[CD|\boldsymbol{\theta}^0(r'),\, \sigma,\, \wp\Big] \ \ge\ P\Big[Y_j - Y_0 \ <\ \frac{an}{\sigma},\, Y_0 - Y_i \ <\ \frac{an}{\sigma},\, 0 < i \le r',\, r' < j \le k\Big]. \tag{7}$$

Let the $(k \times k)$ covariance matrix $\Sigma_{r'} = (\sigma_{ij})$ be given by

$$\sigma_{ij} \ = \ \begin{cases} 1 \quad \text{for } \ i = j, \\ 1 \quad \text{for } \ i \ne j, \ \text{ and, } \ 0 < i,\, j \le r' \text{ or } r' < i, j \le k, \\ -1 \text{ for } \ 0 < i \le r', \ \text{ and, } \ r' < j \le k. \end{cases} \tag{8}$$

Let us define $z_j = Y_j - Y_0$ for $r' < j \le k$ and $z_i = Y_0 - Y_i$ for $0 < i \le r$. Note that the vector $Z' = (z_1, ..., z_k)$ has a symmetric multivariate Laplace distribution with mean vector zero and the covariance matrix $\Sigma_{r'}$ as defined above. For details on Multivariate Laplace Distribution one may look at Kotz et al. (2001). Note that (7) gives the infimum of the probability of correct decision under $\wp$ for

the set of all configurations such that there are r' populations in Ω_R and $k-r'$ in Ω_L. Next, along the lines of Tong (1969), one can derive the Least Favorable Configuration (LFC) under the decision rule $\wp$ as: $\mu_1 = \cdots = \mu_r = \mu_0 + \delta_2$, and, $\mu_{r+1} = \cdots = \mu_k = \mu_0 + \delta_1$, where r is defined in (3). We will refer to the LFC as $\boldsymbol{\theta}^0$. Next, along the lines of (8) with r in place of r', we define the covariance matrix Σ as:

$$\Sigma = \begin{pmatrix} 1 & & 1 & -1 & \cdots & -1 \\ & \ddots & & \vdots & \ddots & \vdots \\ 1 & & 1 & -1 & \cdots & -1 \\ -1 & \cdots & -1 & 1 & & 1 \\ \vdots & \ddots & \vdots & & \ddots & \\ -1 & \cdots & -1 & 1 & & 1 \end{pmatrix}. \tag{9}$$

Next, let $b = b(P^*, k)$ be the solution of the equation

$$P^* = \int_{-\infty}^{b} \cdots \int_{-\infty}^{b} 2(2\pi)^{-\frac{k}{2}} |\Sigma|^{-\frac{1}{2}} (Y'\Sigma^{-1}Y'/2)^{\nu/2} K_\nu(\sqrt{2Y'\Sigma^{-1}Y'}) \prod_{i=1}^{k} dy_i, \tag{10}$$

where $\nu = (2-k)/2$ and $K_\nu(.)$ is the modified Bessel function of the third kind given by $K_\nu(u) = \frac{1}{2}\frac{u}{2}^\nu \int_0^\infty t^{-\nu-1} exp(-t - \frac{u^2}{4t} dt, u > 0$. Then, one can immediately note that

$$P\Big[CD|\boldsymbol{\theta}, \sigma, \wp\Big] \geq P^*,$$

provided n satisfies

$$n \geq \frac{b\sigma}{a} \ (= n^*, \text{ say}). \tag{11}$$

In other words, if σ is known, and one collects a sample of size n^* from each of $\pi_0, \pi_1, \ldots, \pi_k$, and, uses the decision rule $\wp$ given by (6) to partition the k populations, then the probability requirement (2) is achieved. However, if σ is unknown, then there does not exist a single-stage procedure which can achieve the probability requirement (2). For general details on non-existence of a single-stage procedure, one may refer to Dudewicz (1971).

For normal populations case, Tong (1969) gave a single-stage procedure for the partition problem when the common variance known. The single-stage procedure provided above is an extension of Tong's (1969) single-stage procedure. For normal populations case when the common variance is unknown, Tong (1969) constructed a two-stage and a purely sequential procedure. Datta and Mukhopadhyay (1998) studied this problem further and constructed a fine-tuned purely sequential procedure and some other multistage methodologies, emphasizing the second-order asymptotics. Solanky (2001) has constructed an elimination type procedure for the normal populations partition problem which takes samples of unequal sizes. The reader is also recommended to look at Chen and Rollin (2004), Aoshima and Takada (2000), and Solanky (2006), who have studied various aspects of the partition problem. Many other additional references to the partition problem are available in the articles mentioned in this paragraph. Next, We construct a purely sequential procedure for this problem. The theoretical properties of the proposed procedures are derived and verified using Monte Carlo simulation studies.

3 Purely sequential procedure

The purely sequential procedure starts with observations $X_{0j}, X_{1j}, \cdots, X_{kj}$, $j = 1, \cdots, m$, where m ($\geq$2) is the starting sample size from $\pi_0, \pi_1, \cdots, \pi_k$. After this, the sampling continues with one observation from $\pi_0, \pi_1, \cdots, \pi_k$, at each step, according to the stopping rule

$$N = inf\{n \geq m : \ n \geq \frac{b\hat{\sigma}}{a}\}, \tag{12}$$

where $\hat{\sigma}$ is an estimator of σ based on a sample of size n along the lines of (5).

Theorem 1. *For the purely sequential procedure (12) and using the decision rule (6) based on a sample of size N from $\pi_0, \pi_1, \cdots, \pi_k$, we have as $a \to 0$:*

$$(i) \quad N/n_c^* \to 1 \text{ w.p. } 1;$$
$$(ii) \quad E(N/n_c^*) \to 1;$$
$$(iii) \liminf P(CD)] \geq P^* \text{ for all } \boldsymbol{\mu} \in \boldsymbol{R}^{k+1};$$

where $n_c^* = \frac{b\sigma}{a}$ and b comes from (10).

Proof: The proof follows along the lines of the proof of the Theorem 4.4.1 of Mukhopadhyay and Solanky (1994). The details are omitted for brevity.

4 Monte carlo simulations

In this section, we will study the performance of the purely sequential procedure (12) via Monte Carlo simulation studies. The purely sequential procedure (12) was simulated for $m = 10$, $k = 10$ and $P^* = .95$, under a LFC. Without loss of generality we took $\sigma = 1$ for the purpose of generating populations. We took $\delta_1 = -\delta_2$, giving $a = \delta_2 (= \delta$, say). Next, using $n^* = \frac{b\sigma}{a}$, we computed the values of δ corresponding to $n^* = 25, 50, 75,$ and 100. Then, each procedure was independently repeated 1000 times. The performance of the purely sequential procedure (12) is summarized in the Figure 1 via ploting the average value of the observed P(CS) against the optimal sample sizes. The Figure 1 shows that even for small sample sizes, the purely sequential procedure achieves the target value 0.95 quite precisely.

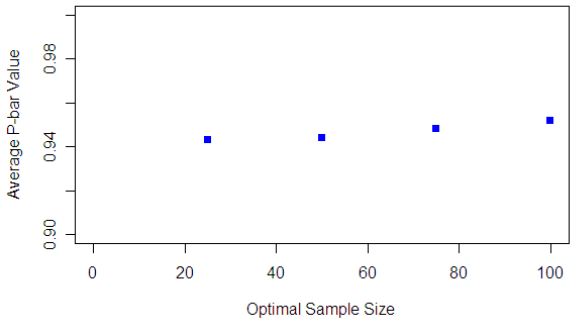

Fig. 1. Performance of the purely sequential procedure (12)
for: k=10, m=10, P^*=.95

References

Aoshima, M. and Takada, Y. (2000). Second order properties of a two stage procedure for comparing several treatments with a control. *Journal of Japan Statistical Society*, **30**, No. 1, 27-41.

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., Santner, T.J, and Goldsman, D.M. (1995). *Design and analysis of Experiments for statistical selection, screening, and multiple comparisons*, John Wiley & Sons, Inc., New York.

Chen, P. and Rollin, L.M. (2004). A two-stage selection and testing design for comparing several normal means with a standard. *Sequential Analysis*, **23**, 75-101.

Datta, S. and Mukhopadhyay, N. (1998). Second-order asymptotics for multistage methodologies in partitioning a set of normal populations having a common unknown variance. *Statistics and Decisions*, **16**, No. 2, 191-205.

Dudewicz, E.J. (1971). Non-existence of a single-sample selection procedure whose P(CS) is independent of the variance. *South African Statistics Journal*, **5**, No. 4, 37-39.

Ghosh, M., Mukhopadhyay, N. and Sen, P. K. (1997). *Sequential Estimation*, John Wiley & Sons, Inc., New York.

Kotz, S., Kozubowski, T.J. and Podgórski, K. (2001). *The Laplace Distribution and Generalizations: A Revisit With Applications to Communications, Economics, Engineering, and Finance*, Birkhäuser, Boston.

Mukhopadhyay, N. and Solanky, T. K. S. (1994). *Multistage selection and ranking procedures*, Marcel Dekker, New York.

Solanky, T.K.S. (2001). A sequential procedure with elimination for partitioning a set of normal populations having a common unknown variance. *Sequential Analysis*, **20**, No. 4, 279-292.

Solanky, T.K.S. (2006). A two-stage procedure with elimination for partitioning a set of normal populations with respect to a control. *Sequential Analysis*, **25**, No. 3, 297-310.

Tong, Y. L. (1969). On partitioning a set of normal populations by their locations with respect to a control. *Annals of Mathematical Statistics*, **40**, 1300-1324.