Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 1, 2021 (New Series), pp 401–415

Second Order Asymptotics of a Fine-Tuned Purely Sequential Procedure for the Generalized Partition Procedure

Tumulesh K. S. Solanky
Department of Mathematics
University of New Orleans, New Orleans, Louisiana, USA

Received: 24 January 2021; Revised: 01 March 2021; Accepted: 17 March 2021

Abstract

In the area of selection and ranking, partitioning of treatments by comparing them to a control treatment is an important statistical problem. For over seventy years this problem has been investigated by a number of researchers via various statistical designs to specify the partitioning criteria and optimal strategies for data collection. Tong (1969) had proposed a design which had generalized many formulations known at that time. Relying upon Bechhofer's (1954) indifference-zone formulation, Tong (1969) had designated the region between the boundaries for "good" and "bad" treatments as the indifference-zone. Since then the formulation presented in Tong (1969) has been adopted by a number of researchers to study various aspects of the partition problem. However, in "Tong's formulation" the dual role the indifference zone plays, in defining the preference-zone and boundaries for "good" and "bad" treatments, could potentially make the formulation impractical when there is a large difference between the "good" and "bad" treatments. In Solanky and Jhou (2015), a generalization of the "Tongs formulation" was introduced so that the treatments which fall between the "good" and "bad" treatments can be partitioned as a separate identifiable group by introducing two indifference-zones. It was also shown that the formulation design in Tong (1969) is a special case of Solanky and Jhou's (2015) formulation. However, the second-order expansions of the probability of correct partition given in Solanky and Zhou (2015) does not make it clear how close one really gets to the target probability requirement. In this paper, we have proposed a fine-tuned purely sequential procedure which is asymptotically unbiased and guarantees the probability requirement by taking a few additional samples along the lines of Mukhopadhyay and Datta (1995). The "first-order" and "second order" asymptotics of the fine-tuned procedure are derived and it is shown that the second-order expansion of the stopping time has the same order of the remainder as that for the original procedure in Solanky and Zhou (2015). The performance of the proposed fine-tuned procedure is studied via Monte Carlo simulations.

Key words: Fine-tuned purely sequential procedure; Control population; Indifference zone; Probability of correct decision; Normal distribution; Simulations.

AMS Subject Classifications: 62F07, 62L10

Corresponding Author: Tumulesh K. S. Solanky
Email: tsolanky@uno.edu

1. Introduction

The problem of comparisons with a control has intrigued the researchers for the last seventy plus years. It has been investigated under different designs and sampling methodologies. Among the oldest investigations available in the statistical literature, Roessler (1946), Paulson (1952), and Dunnett (1955) are three studies which had introduced the need for statistical methodologies to compare treatments with a control treatment and had formulated designs for such a comparison.

For a somewhat related problem of selecting or isolating the best population, two pioneering papers in 1950s had presented contrasting formulations. The first formulation was in Bechhofer (1954) which introduced the idea of indifference-zone for selecting the best normal population from a group of several normally distributed populations. This formulation, known as "indifference-zone" formulation, had the property of selecting the best population with the pre-determined probability specified by the practitioner for the populations that are in the preference-zone. The region outside of the indifference-zone is referred to as the *preference-zone*. From a practical point of view, the indifference zone was envisioned to be small enough so that the experimenter could be easily be *indifferent* to the treatments that fall inside this region. The other pioneering formulation, introduced in Gupta (1956), did not restrict the selection from the preference-zone but rather the selection was carried out from the entire parameter space. The formulation in Gupta (1956), known as "subset-selection formulation", is designed to select a subset of random size which includes the best treatment with the pre-determined probability of correct selection. Since then, the literature in this area has grown enormously extending Bechhofer's (1954) indifference-zone formulation and Gupta's subset-selection formulation, to solve problems related to selecting the best treatment and by defining the "best-treatment" in many ways to meet the goals of the study. This area of research is broadly known as the area of *selection and ranking* in the statistical literature. Along side with the research in the *selection and ranking* to select the best population, another research problem has also been quite active which is concerned with comparisons of treatments with a specific treatment of choice. What made this research area different from selecting the best treatment is the experimenters requirement that the the population to be selected must be some "specified amount better" than some other treatment typically referred to as a *control* or *standard*. This area of research is typically known as the problem of "comparisons with a control" or the "partition problem" in statistical literature.

1.1. Tong's formulation of partition problem

We will start by presenting the partition problem formulation introduced in Tong (1969) for the populations that follow a normal distribution. The formulation starts by specifying the "good" and "bad" populations based on the input from experts in the area of the application. The region that falls in between these two boundaries is next defined in Tong (1969) as the "indifference zone".

The concept of indifference-zone was introduced in Bechhofer (1954) for selecting the best normal population in order to create a spacing between the best and the rest of the treatments under consideration with the underlying requirement that the experimenter would be "indifferent" to the treatments falling inside the indifference-zone. In other words, any population that is inside the indifference-zone cannot be miss-classified.

This dual role the indifference zone, in specifying the "good" and "bad" populations boundaries and also the "indifference zone" itself, could potentially make the formulation impractical in cases when there is a large difference between "good" and "bad" populations. We will revisit this issue a bit later in this paper.

Let $\pi_0, \pi_1, \cdots, \pi_k$ denote the $(k+1)$ independent and normally distributed populations with respective means $\mu_i,\ i = 0,\ 1, \cdots, k$, and common variance σ^2. Assume that all the parameters $\mu_i,\ i = 0,\ 1, \cdots, k$, and variance σ^2 are unknown. We will denote the population π_0 as the control population with which remaining k populations will be compared.

Based on the guidelines of "good" and "bad" treatments from the practitioner in the area, Tong (1969) defined two appropriate constants δ_1 and δ_2, with the requirement that $\delta_1 < \delta_2$, to split the parameter space Ω into three sets following Bechhofer's (1954) indifference-zone formulation, as follows

$$\begin{array}{rcl} \Omega_{BT} & = & \{\pi_i : \mu_i \leq \mu_0 + \delta_1,\ i = 1, \cdots, k\}, \\ \Omega_{IT} & = & \{\pi_i : \mu_0 + \delta_1 < \mu_i < \mu_0 + \delta_2,\ i = 1, \cdots, k\}, \\ \Omega_{GT} & = & \{\pi_i : \mu_i \geq \mu_0 + \delta_2,\ i = 1, \cdots, k\}. \end{array} \tag{1}$$

Without loss of generality, the set Ω_{GT} will be used to denote "good" populations and the set Ω_{BT} the "bad" populations. From the applications point of view, the values of constants δ_1 and δ_2 are determined based on input from the experts to specify which population should be classified as "good" population compared to control population and which as a "bad" population compared to control population. The partition problem formulated in Tong (1969) was designed to correctly partition the populations which belonged to only Ω_{BT} and Ω_{GT}. Whereas, since the set expressed as Ω_{IT} was the "indifference-zone" set, the mathematical formulation was indifferent to the populations which are in this set. In other words, the populations that are in Ω_{IT} could be partitioned in Ω_{BT} or Ω_{GT} without any penalty. The decision rule proposed in Tong (1969) was to partition or classify the set Ω of the k populations into non-overlapping subsets S_{BT} and S_{GT} of Ω, such that, $\Omega_{GT} \subseteq S_{GT}$ and $\Omega_{BT} \subseteq S_{BT}$. A decision rule which classifies all populations in Ω_{BT} or Ω_{GT} correctly was termed as a "correct decision (CD)". The design constants used in the partition rule used in Tong (1969) are defined below

$$\begin{array}{l} d_1 = \frac{\delta_1+\delta_2}{2}, \quad a_1 = \frac{\delta_2-\delta_1}{2}, \quad \lambda = \frac{\sigma}{a_1}, \\ m_1 = \begin{cases} \frac{k}{2} & \text{for even values of } k, \\ \frac{k+1}{2} & \text{for odd values of } k. \end{cases} \end{array} \tag{2}$$

Mathematically, for any pre-specified probability P^*, where $\frac{1}{2^k} < P^* < 1$, Tong (1969) presented a decision rule $\wp_T$ to obtain sets S_{BT} and S_{GT}, which satisfy the probability requirement

$$P\{CD|\boldsymbol{\mu}, \sigma^2, \wp\} \geq P^* \qquad \forall\ \boldsymbol{\mu} \in \boldsymbol{R}^{k+1},\ \sigma \in \boldsymbol{R}^+, \tag{3}$$

where $\boldsymbol{\mu} = (\mu_0, \mu_1, \cdots, \mu_k)$.

Adopting a sampling design to determine the sample size of N observations from all the

k+1 populations and control, the decision rule to partition the k populations in Ω presented in Tong (1969) was the following:

$$\begin{array}{rcl} S_{BT} & = & \{\pi_i : \bar{X}_{iN} - \bar{X}_{0N} \leq \ d_1, \ i = 1, \cdots, k\}, \\ S_{GT} & = & \{\pi_i : \bar{X}_{iN} - \bar{X}_{0N} \geq \ d_1, \ i = 1, \cdots, k\}, \end{array} \tag{4}$$

where d_1 comes from (2).

For "partition problem" as outlined above, Tong (1969) derived a fixed sample size solution for the case when σ^2 is known by determining the optimal sample size needed to meet the probability condition specified in (3). Tong (1969) also considered the case when the common variance σ^2 unknown by designing a two-stage stopping rule and a purely-sequential in order to meet the probability requirement (3). Tong's (1969) formulation was utilized by Datta and Mukhopadhyay (1998) to construct a fine-tuned purely sequential procedure, an accelerated sequential procedure and a three-stage procedure focusing on the second-order asymptotics for each procedure. Also following Tong's (1969) formulation, with the additional goal of reducing the sampling from populations which can be partitioned based on smaller sample sizes due to being significantly better or worse than the control population, Solanky (2001) constructed an elimination type purely-sequential stopping rule. Using triangular boundaries, Solanky (2001) was designed to reduce the sampling cost by not following "vector-at-a-time" sampling methodology and instead it presented a sampling design that eliminated and partitioned the non-contending populations early during the sampling process. The operational inconvenience of purely-sequential stopping rules was also the focus of Solanky (2006) which designed a two-stage sampling rule which had the desirable property of eliminating "too inferior" or "too superior" populations based on samples collecting during the first stage by tactfully implementing Gupta's subset selection for screening of such populations. In the stage two of sampling in Solanky (2006), only the competing treatments which were shortlisted in stage one were sampled from and partition was implemented using Bechhofer's indifference zone approach. Also relying on the Tong's (1969) formulation, Solanky and Wu (2004) had constructed an "unbalanced sampling design" which allows an experimenter to collect a bigger sample size from the control population while reducing the sample size from the non-control populations from which "vector-at-a-time" sampling methodology is adopted. For references on partition problem for Binomial treatments the reader is referred to Buzaianu (2019).

1.2. Generalization of the partition problem

The formulation presented above due to Tong (1969) was constructed under the Bechhofer's(1954) "indifference-zone formulation" to partition the k populations as either a 'Good" treatments or a "Bad" treatments with respect to the control population. The methodology of Tong (1969) was formulated to partition the populations which may fall under the indifference-zone as either "Good" populations or "Bad" populations without any penalty on the probability requirement specified in (3). This requirement of Tong's (1969) formulation would be intuitively serve the experimenter well as long as the distance between the "Good" populations and "Bad" populations is is not large so that one could be "indifferent" to how the populations which fall inside it are partitioned. However, when there is large gap between the "Good" populations and "Bad" populations boundaries then this could result in misleading results and would also violate the concept of minimum distance

worth measuring as represented by the indifference-zone. One is referred to Solanky and Zhou (2015) for illustrations and further insights.

In Solanky and Jhou (2015), the partition problem was formulated in such a fashion that does not require the region that falls between the "Good" populations and "Bad" populations boundaries to be designated as indifference zone. Instead, by utilizing the creation of two indifference-zones, the generalized formation was able to partition the treatments which fall between the "Good" populations and "Bad" populations as a group by itself. The two indifference-zones thus created were based on the fundamentals of "indifference-zone" as intended in Bechhofer (1954) and they could be as small as the experimenter desires and also met the criteria to be minimum distance worth measuring. Put differently, the Solanky and Jhou (2015) generalized formulation had the capability of creating indifference-zones independently of the boundaries of "Good" populations and "Bad" populations. Next, we represent the mathematical details of the generalized formulation of Solanky and Jhou (2015).

Based on the input from experts in the area, the statistical design would start by selecting two design constants δ_1 and δ_4, $\delta_1 < \delta_4$, to define the "Good" populations and "Bad" populations compared to the control population. Next, based on experts understanding of how much distance is worth detecting or the "threshold", one would quantify that information by the design constants δ_2 and δ_3 so that $\delta_2 - \delta_1 (> 0)$ and $\delta_4 - \delta_3 (> 0)$ are the spacing's which will be used to construct two 'indifference-zones". Next, one would define the following sets to split the entire parameter space Ω along the lines of Bechhofer's (1954) "indifference-zone formulation"

$$\begin{array}{rcl} \Omega_B & = & \{\pi_i : \mu_i \leq \mu_0 + \delta_1,\ i = 1, \cdots, k\}, \\ \Omega_{I_1} & = & \{\pi_i : \mu_0 + \delta_1 < \mu_i \leq \mu_0 + \delta_2,\ i = 1, \cdots, k\}, \\ \Omega_M & = & \{\pi_i : \mu_0 + \delta_2 < \mu_i \leq \mu_0 + \delta_3,\ i = 1, \cdots, k\}, \\ \Omega_{I_2} & = & \{\pi_i : \mu_0 + \delta_3 < \mu_i \leq \mu_0 + \delta_4,\ i = 1, \cdots, k\}, \\ \Omega_G & = & \{\pi_i : \mu_i > \mu_0 + \delta_4,\ i = 1, \cdots, k\}. \end{array} \tag{5}$$

One may note that the sets Ω_{I_1} and Ω_{I_2} will serve as the two "indifference-zones". It is also apparent that the sizes of these two "indifference-zones" does not depend on the what a practitioner may have picked to specify the "Good populations" and "Bad populations" compared to the control population. Next, following Solanky and Jhou (2015), we define the constants

$$d_1 = (\delta_1 + \delta_2)/2, \qquad d_2 = (\delta_3 + \delta_4)/2, \qquad a_1 = (\delta_2 - \delta_1)/2, \qquad a_2 = (\delta_4 - \delta_3)/2. \tag{6}$$

We write X_{ij} to denote a random sample of size n from the population π_i where $j = 1, \cdots, n$ and $i = 0, 1, \cdots, k$. Writing $\bar{X}_i$ and S_i^2 for the usual sample mean and sample variance from the ith population π_i as $\bar{X}_i = \frac{\sum_{j=1}^{n} X_{ij}}{n}$, and $S_i^2 = \frac{\sum_{j=1}^{n} \left(X_{ij} - \bar{X}_i\right)^2}{n-1}$, we obtain the pooled estimator of the common variance σ^2 to be denoted as $U(n) = \frac{\sum_{i=0}^{k} S_i^2}{k+1}$. Next, as in Solanky and Jhou (2015), we propose a intuitively defined decision rule $\mathcal{P}_n$ based on the vector-at-a-time sampling strategy of size n from all the populations:

$$\begin{aligned} S_B &= \{\pi_i : \bar{X}_{in} - \bar{X}_{0n} \leq d_1, i = 1, \cdots, k\}, \\ S_M &= \{\pi_i : d_1 \leq \bar{X}_{in} - \bar{X}_{0n} \leq d_2, i = 1, \cdots, k\}, \\ S_G &= \{\pi_i : \bar{X}_{in} - \bar{X}_{0n} \geq d_2, i = 1, \cdots, k\}, \end{aligned} \tag{7}$$

where S_B is the set of populations that are classified as"Bad populations", S_M as "Medium populations", and S_G as"Good populations". As explained earlier, above decision rule does not change the definition of "Good" populations and "Bad" populations compared to the control population which are based on input from experts in the area.

Also, as it is customary for the selection and ranking problems, the decision rules and the probabilities of correct decision are generally derived under the worst possible parametric configurations, which are referred to as the "least favorable configurations (LFC)". And, when the parametric configurations become favorable, then the observed probabilities of correct decision would exceed the target probability values. Using symmetry argument, it is apparent that if Ω_{I_1} and Ω_{I_2} not equal in width then the LFC would become a function of constants δ_2 and δ_3, and hence LFC would cease to exist in a general form. Also, intuitively because there is no penalty for misidentifying the populations that fall inside the two "indifference-zones" Ω_{I_1} and Ω_{I_2}, under the LFC should not have any population that falls inside any of them. Utilizing the above information, in Solanky and Zhou (2015) the LFC was introduced taking advantage of the symmetry in the decision rule (7). We rewrite the spacing's using $a = \delta_4 - \delta_1$, the distance between the "Good" populations and "Bad" populations, and a constant ρ, $0 < \rho < \frac{1}{2}$, which expresses the size of the two "indifference-zones" as a function of a.

(1) $\delta_4 - \delta_3 = \delta_2 - \delta_1 = \rho a$, $0 < \rho < \frac{1}{2}$. Note that ρ is a design constant which is used to define the size of the indifference-zones;

(2) $r_2 + r_3 = \left[\frac{k}{2}\right] = k_1$, $r_1 + r_4 = k - k_1$, $r_2 = \left[\frac{k_1}{2}\right]$, $r_3 = k_1 - r_2$, $r_1 = \left[\frac{k-k_1}{2}\right]$, $r_4 = k - k_1 - r_1$, where r_1, r_2, r_3, and r_4 denotes the number of populations with the respective means: $\mu_0 + \delta_1, \mu_0 + \delta_2, \mu_0 + \delta_3$, and $\mu_0 + \delta_4$, where $[x]$ equals $\frac{x}{2}$ if x is even and $\frac{x+1}{2}$ if x is odd. We denote this parametric configuration as $\boldsymbol{\mu}^0(r_1, r_2, r_3, r_4)$.

As explained in Solanky and Zhou (2015), the condition (1) above forces the two indifference-zone's Ω_{I_1} and Ω_{I_2} to be of same size, whereas, the condition (2) adds symmetry to the problem and essentially allows equal number of populations at the four boundaries $\mu_0 + \delta_1, \mu_0 + \delta_2, \mu_0 + \delta_3$, and $\mu_0 + \delta_4$. Solanky and Zhou (2015) had shown that LFC is when $r_1 = r_2 = r_3 = r_4$. It is easy to note that as ρ becomes large and approaches $\frac{1}{2}$, the size of the Ω_{I_1} and Ω_{I_2} increases and the size of "Medium populations" which are classified as a separate group under generalized partition methodology becomes smaller. And, in this case the generalized partition rule would approach the formulation presented in Tong (1969). The constant ρ determines the "threshold" or "minimum distance worth detecting" along the lines of Bechhofer (1954). Whereas, in the formulation of Tong (1969) the "indifference zone" played the dual role of defining the "Good" and "Bad" populations and the role of the "threshold" or the "minimum distance worth detecting". As derived in Solanky and Jhou (2015), for the known common variance σ^2 case, the probability for the correct decision for (7) can be shown to be

$$P\Big[CD|\boldsymbol{\mu}^0(r_1,r_2,r_3,r_4),\sigma^2\Big] \geq P^* \tag{8}$$

if the sample size from the k populations and the control population at least $\frac{8b^2\sigma^2}{(\rho a)^2}$ $(=C)$ and the design constant $b = b(k, P^*)$ is obtained by solving:

$$\int_{-b}^{b}\int_{-b}^{b}...\int_{-b}^{b}(2\pi)^{-\frac{k_1}{2}}|\Sigma|^{-\frac{1}{2}}\exp(-\frac{1}{2}y'\Sigma^{-1}y)dy_1...dy_{\frac{k_1}{2}} = \frac{P^*+1}{2}. \tag{9}$$

In the statistical literature, $C = \frac{8b^2\sigma^2}{(\rho a)^2}$, is referred to as the fixed-sample solution to the partition problem. The values of design constant $b = b(k, P^*)$ are available in Solanky and Jhou (2015).

In section 2, we have proposed a fine-tuned purely sequential procedure that is asymptotically unbiased and guarantees the probability requirement by taking a few additional samples along the lines of Mukhopadhyay and Datta (1995) and Woodroofe (1991). The "first-order" and "second order" asymptotics of the proposed procedure are obtained and it is shown that the second-order expansion of the stopping time has the same order of the remainder as that for the original procedure in Solanky and Zhou (2015). The probability of the correct decision for the fine-tuned version of the purely-sequential procedure for the generalized partition problem will be shown to be at least P^* up to the second-order approximation and will not have additional terms.

2. Fine-Tuned Purely Sequential Procedure

Next, we propose a "fine-tuned purely sequential procedure" for the generalized partition problem described in (5). The procedure starts with initial sample size of $m_0\,(\geq 2)$ observations. Next, adopting the "vector-at-a-time" sampling design and it takes one observation at a time updating the statistic $U(n)$ after each observation until the first time the condition in the condition below is satisfied

$$N \equiv N\,(a) = \text{Inf}\left\{n \geq m_0 : n + \epsilon\ \geq \frac{8b^2U(n)}{(\rho a)^2}\right\}, \tag{10}$$

where the constant $\epsilon = \epsilon\,(k, P^*)$ is defined in (21) and $b = b(k, P^*)$ in (9). We first verify that the stopping rule (10) will terminate with probability one. Note that for given values of constants $\boldsymbol{\mu}$, σ^2, m_0, ρ and a, we can write

$$\begin{aligned} P\,(N<\infty) &= 1-\lim_{n\to\infty}P\,(N>n) \\ &\geqslant 1-\lim_{n\to\infty}P\left\{n+\epsilon<\frac{8b^2U(n)}{(\rho a)^2}\right\} \\ &= 1,\ \text{since } U(n)\xrightarrow{P}\sigma^2 \text{ as } n\to\infty. \end{aligned}$$

Hence, with probability one the proposed "fine-tuned purely sequential procedure" (10) will terminate. Based on the totality of all samples of size N, one computes the sample means $\bar{X}_{0N}, \bar{X}_{1N}, ..., \bar{X}_{kN}$ and implements the decision rule (7). Next, we will derive some theoretical

properties to establish the performance of the "fine-tuned purely sequential procedure" (10) asymptotically. It is easy to verify that as a becomes small, the sample size as determined by the stopping rule (10) becomes larger. We start with presenting first-order asymptotics "fine-tuned purely sequential procedure" (10).

Theorem 1: For N as defined by the"fine-tuned purely sequential procedure" (10), the following properties are satisfied :

(i) $\frac{N}{C} \xrightarrow{P} 1$ as $a \to 0$:

(ii) $E(\frac{N}{C}) \to 1$ as $a \to 0$:

(iii) $\frac{N-C}{C^{\frac{1}{2}}} \xrightarrow{L} N(0, \frac{2}{k+1})$ as $a \to 0$:

(iv) $\liminf_{a\to 0} P\left(CD|\mathcal{P}_N\right) > P^*$ under the LFC;

where $C = \frac{8b^2\sigma^2}{(\rho a)^2}$, $b = b(k, P^*)$ is defined in (9) and has been tabulated in Solanky and Zhou (2015), and the constant $\epsilon = \epsilon(k, P^*)$ is defined in (21).

Proof: We start the proof by noting that the N is well-defined and is a non-decreasing function of a. Next, along the lines of Theorem 2.4.1 of Mukhopdahyay and Solanky (1994) we use the Lemma 1 of Chow and Robbins (1965) to note $N \xrightarrow{P} \infty$ as $a \to 0$. Also, $U(N) \xrightarrow{P} \sigma^2$ and $U(N-1) \xrightarrow{P} \sigma^2$ as $a \to 0$. Next, inequality (2.4.3) from Mukhopadhyay and Solanky (1994) for the stopping rule (10) simplifies to

$$\frac{8b^2U(N)}{(\rho a)^2} - \epsilon \leq N \leq m_0 - \epsilon + \frac{8b^2U(N-1)^2}{(\rho a)^2}. \tag{11}$$

We can rewrite (11) as

$$\frac{8b^2U(N)}{(\rho a)^2C} - \frac{\epsilon}{C} \leq \frac{N}{C} \leq \frac{m_0 - \epsilon}{C} + \frac{8b^2U(N-1)^2}{(\rho a)^2C},$$

which simplifies to

$$\frac{U(N)}{\sigma^2} - (\frac{\epsilon\rho^2}{8b^2\sigma^2})a^2 \leq \frac{N}{C} \leq (\frac{(m_0-\epsilon)\rho^2}{8b^2\sigma^2})a^2 + \frac{U(N-1)}{\sigma^2}.$$

Above with $U(N) \xrightarrow{P} \sigma^2$ and $U(N-1) \xrightarrow{P} \sigma^2$ as $a \to 0$ completes the part (i) of the Theorem. Next, as in Mukhopadhyay and Solanky's (1994) Theorem 3.5.1, we rewrite the estimator $U(n)$ by using the "Helmert's orthogonal transformation" to obtain R_1', R_2', $\cdots$ which are i.i.d. $(k+1)^{-1}\sigma^2\chi^2_{k+1}$ to express $U(n) = \frac{1}{n-1}\sum_{i=1}^{n-1} R_i'$. Considering only one side of (11), and writing $R_* = \sup_{n\geq 2}\left\{\frac{1}{n-1}\sum\limits_{i=1}^{n-1} R_i'\right\}$, we have

$$N \leq m_0 - \epsilon + \frac{8b^2R_*}{(\rho a)^2}.$$

This can be rewritten as

$$\frac{N}{C} \leq \frac{m_0 - \epsilon}{C} + \frac{R_*}{\sigma^2}, \tag{12}$$

and when a is small enough so that $\frac{1}{C}$ become less than 1, the inequality (12) simplifies to $\frac{N}{C} \leq m_0 - \epsilon + \frac{R_*}{\sigma^2}$. By "Wiener's (1939) dominated ergodic theorem" we have $E(R_*) < \infty$. By applying the "dominated convergence theorem" and using Theorem's (i) result, have the proof for part (ii). The proof for part (iii) of the Theorem can be derived by comparing the "fine-tuned purely sequential procedure" (10) with (3.5.1) of Mukhopadhyay and Solanky (1994). The details are omitted for brevity.

Next, to prove the part (iv), note that without loss of generality we will assume that $\pi_i, i = 1, ..., r_1$ have means $\mu_0 + \delta_1$; $\pi_j, j = r_1 + 1, ..., r_1 + r_2$ have means $\mu_0 + \delta_2$; $\pi_l, l = r_1 + r_2 + 1, ..., r_1 + r_2 + r_3$ have means $\mu_0 + \delta_3$; and $\pi_m, m = r_1 + r_2 + r_3 + 1, ..., k$ have means $\mu_0 + \delta_4$. Next, the probability of correct decision can be written as

$$\begin{aligned} &P\Big(CD|\mathcal{P}_N; \boldsymbol{\mu}^0(r_1, r_2, r_3, r_4),\ \ \sigma^2\Big) \\ &\quad = P\Big[\bar{X}_i - \bar{X}_0 < d_1,\ d_1 < \bar{X}_j - \bar{X}_0 < d_2,\ d_1 < \bar{X}_l - \bar{X}_0 < d_2,\ \bar{X}_m - \bar{X}_0 > d_2, \\ &\quad 0 < i \leq r_1,\ \ r_1 < j \leq r_1 + r_2,\ \ r_1 + r_2 < l \leq r_1 + r_2 + r_3,\ \ r_1 + r_2 + r_3 < m \leq k\Big]. \end{aligned}$$

As in Solanky and Zhou (2015), the P(CD) can be tactfully written as probability expressions for the correct partition of the populations which are "Good" or "Bad populations" and probability expressions for the correct partition of the populations which are "Medium populations". Assuming that the practitioner considers the correct partition of these two groups of populations as equal, then the P(CD) expression would simplify to:

$$\begin{aligned} P(CD|\mathcal{P}_N) \ \ \geq \ \ (-1) + 2P\Big\{&d_1 < \bar{X}_j - \bar{X}_0 < d_2, r_1 + 1 \leq j \leq r_1 + r_2; \\ &d_1 < \bar{X}_l - \bar{X}_0 < d_2, r_1 + r_2 + 1 \leq l \leq r_1 + r_2 + r_3\Big\}. \end{aligned}$$

Above can be simplified as

$$\begin{aligned} P(CD|\mathcal{P}_N) \ \ \geq \ \ (-1) + 2P\Bigg\{&\frac{d_1 - \delta_2}{\sqrt{\frac{\sigma^2}{N}}} < \frac{\left(\bar{X}_j - \mu_j\right)}{\sqrt{\frac{\sigma^2}{N}}} - \frac{\left(\bar{X}_0 - \mu_0\right)}{\sqrt{\frac{\sigma^2}{N}}} < \frac{d_2 - \delta_2}{\sqrt{\frac{\sigma^2}{N}}}; \\ &r_1 + 1 \leq j \leq r_1 + r_2; \quad \frac{\rho a}{\sqrt{\frac{\sigma^2}{N}}} < \frac{\left(\bar{X}_l - \mu_l\right)}{\sqrt{\frac{\sigma^2}{N}}} \\ &-\frac{\left(\bar{X}_0 - \mu_0\right)}{\sqrt{\frac{\sigma^2}{N}}} < \frac{d_2 - \delta_3}{\sqrt{\frac{\sigma^2}{N}}},\ \ r_1 + r_2 + 1 \leq l \leq r_1 + r_2 + r_3\Bigg\}. \end{aligned}$$

Denoting, $b = \frac{\rho a/2}{\sqrt{\frac{\sigma^2}{N}}}$, and $c = \frac{(2a - 3\rho a)/2}{\sqrt{\frac{\sigma^2}{N}}}$, then the above can be rewritten as

$$\begin{aligned} P(CD|\mathcal{P}_N) \ \ \geq \ \ (-1) + 2P\{&-b < Z_j - Z_0 < c, r_1 + 1 \leq j \leq r_1 + r_2; \\ &-c < Z_l - Z_0 < b, r_1 + r_2 + 1 \leq l \leq r_1 + r_2 + r_3\}. \end{aligned}$$

In order to obtain a simpler closed form expression for $P(CD)$ note that for $\rho < \frac{1}{2}$, $b < c$, and hence $-b > -c$ giving the $P(CD)$ as

$$\begin{aligned} P(CD|\mathcal{P}_N) &\geq (-1) + 2P(-b < Z_j - Z_0 < b, \quad r_1 + 1 \leq j \leq r_1 + r_2, \quad -b < Z_l - Z_0 < b, \\ &\quad r_1 + r_2 + 1 \leq l \leq r_1 + r_2 + r_3) \\ &= (-1) + 2P(-b + Z_0 < Z_i < b + Z_0, \quad r_1 + 1 \leq i \leq r_1 + r_2 + r_3) \\ &= (-1) + 2E[\int_{-\infty}^{+\infty} \{\Phi(z+b) - \Phi(z-b)\}^{r_2+r_3} \phi(z) dz | Z_0 = z]. \end{aligned} \tag{13}$$

Rewriting the first result in the Theorem, we have $\sqrt{N}\frac{\rho a}{\sqrt{8\sigma}} \xrightarrow{P} b$ as $a \to 0$, and next using the "dominated convergence theorem" completes the proof of part (iv). □

Next, for the the "fine-tuned purely sequential procedure" (10) the "second-order expansions" are presented to asymptotically quantify the number of observations the procedure will sample compared to the optimal same size. We will also show that the proposed procedure is asymptotically unbiased version of the procedure in Solanky and Zhou (2015) in achieving the targeted value of probability requirement.

Theorem 2: The"fine-tuned purely sequential procedure" (10), have the following properties as $a \to 0$:

(i) $E(N) = C - \epsilon + \frac{\nu - 2}{k+1} + \circ(1)$, provided that $m_0 \geq 1 + \frac{2}{k+1}$;

(ii) $E(N^\omega) = C^\omega + [\omega(\frac{\nu-2}{k+1} - \epsilon) + \frac{1}{2}\omega(\omega-1)\frac{2}{k+1}]C^{\omega-1} + \circ(1)$, provided that (a) $m_0 \geq 1 + (3-\omega)\frac{2}{k+1}$ for $\omega \in (\infty, 2) - \{-1, 1\}$; (b) $m_0 \geq 1 + \frac{2}{k+1}$ for $\omega = 1$ and $\omega \geq 2$; (c) $m_0 \geq 1 + \frac{4}{k+1}$ for $\omega = -1$.

(iii) $P(CD|\mathcal{P}_N; \boldsymbol{\mu}^0(r_1, r_2, r_3, r_4), \ \sigma^2) > P^* + o(\frac{1}{C})$ provided that $m_0 > \frac{5}{k+1} + 1$ for the LFC;

where $C = \frac{8b^2\sigma^2}{(\rho a)^2}$, and the constant $\nu(k)$ is defined in equation (16).

Proof: First note that as in Woodroofe (1977) and Theorem 2.4.8 of Mukhopadhyay and Solanky (1994), the pooled estimator $U(n)$ of σ^2 can be written as sum of i.i.d. random variables as $U(n) = \frac{1}{n-1}\sum_{i=1}^{n-1} R_i'$, where R_1', R_2', $\cdots$ are i.i.d. $\frac{1}{k+1}\sigma^2\chi^2_{k+1}$ random variables. Let's write $R_i = (k+1)\sigma^{-2}R_i'$, with R_i being i.i.d. χ^2_{k+1}. Using this the purely sequential procedure could be rewritten as

$$N = \text{Inf}\left\{n \geq m_0 : n + \epsilon \ \geq \frac{8b^2}{(\rho a)^2}(n-1)^{-1}\frac{\sigma^2}{k+1}\sum_{i=1}^{n-1} R_i\right\}.$$

This can rewritten as

$$N = \text{Inf}\left\{n \geq m_0 : C^{-1}(n+\epsilon)(n-1)(k+1) \geq \sum_{i=1}^{n-1} R_i\right\}. \tag{14}$$

Next, following Mukhopadhyay and Solanky's (1994) Theorem 2.4.8 we express N as $T+1$, for the stopping rule T defined as

$$T = \text{Inf}\left\{ n \geq m-1 : C^{-1} n^2 (k+1) \left(1 + \frac{\epsilon+1}{n}\right) \geq \sum_{i=1}^{n} R_i \right\}. \tag{15}$$

The stopping rule T has been expressed in the general form of "equation (2.4.7) in Mukhopadhyay and Solanky (1994)" with the values: $L_0 = \epsilon + 1$; $\delta = 2$; $h^* = \frac{k+1}{C}$; $\theta = E(R_1) = k+1$; $r^2 = E(R_1^2) - \theta^2 = 2(k+1)$; $\beta^* = \frac{1}{\delta - 1}$ which simplifies to 1; the value of $n_0^* = (\frac{\theta}{h^*})^{\beta^*}$ simplifies to C; and the constant $p = \frac{\beta^{*2} r^2}{\theta^{-2}}$ becomes $\frac{2}{k+1}$. Also, note that with R_i being i.i.d. χ^2_{k+1}, we can express

$$P(R_i \leq r) = \int_0^r \frac{e^{\frac{-x}{2}} x^{\frac{k+1}{2}-1}}{2^{\frac{k+1}{2}} \Gamma(\frac{k+1}{2})}\, dx.$$

Since $e^{\frac{-x}{2}} \leq 1$, replacing $e^{\frac{-x}{2}}$ by 1 and carrying out the integral one will obtain

$$P(R_i \leq r) \leq \frac{2}{k+1} \frac{1}{2^{\frac{k+1}{2}} \Gamma(\frac{k+1}{2})} r^{\frac{k+1}{2}}.$$

Also, note that for the random variable R, we have

$P(R_i \leq r) \leq H r^h$, where $H = \frac{2}{k+1} \frac{1}{2^{\frac{k+1}{2}} \Gamma(\frac{k+1}{2})}$ and $h = (k+1)\big/2$. Next, using the nonlinear renewal theory from "Mukhopadhyay and Solanky (section 2.4.2)" we define the constant $\nu = \nu(k)$ as below:

$$\nu = \nu(k) = \frac{k+3}{2} - \sum_{n=1}^{\infty} \frac{1}{n} E\left[\max\left\{0, \chi^2_{(k+1)n} - 2(k+1)n\right\}\right]. \tag{16}$$

Next, following "equation (2.4.10) in Mukhopadhyay and Solanky (1994)" the constant $\eta = \frac{\beta^* \nu}{\theta} - L_0 \beta^* - \frac{\delta \beta^{*2} r^2}{2\theta^2}$ would become

$$\eta = (k+1)^{-1}(\nu - 2) - (\epsilon + 1). \tag{17}$$

Since $T = N - 1$ and therefore $E(T) = E(N) - 1$, next with η as defined above in (17), applying the "Theorem 2.4.8(v) of Mukhopadhyay and Solanky (1994)" would lead to

$$E(N) = C - \epsilon + \frac{\nu - 2}{k+1} + \circ(1),$$

provided that $m > 1 + 2(k+1)^{-1}$. This completes the proof of part (i).

The proof of the part (ii) follows by expressing the stopping variable T in the form of "Mukhopadhyay and Solanky (1994)'s equation (2.4.7)" and applying the "Theorem 2.4.8 of Mukhopadhyay and Solanky (1994)". The details are omitted for brevity. One may note that the part (i) is a special case of part (ii) when $\omega = 1$.

For part (iii), note that from (13) we can express the probability of correctly partitioning all k populations using the decision rule $\mathcal{P}_N$ under the LFC as

$$P(CD|\mathcal{P}_N) \;\geq\; (-1) + 2E[\int_{-\infty}^{\infty} \{\Phi(z+b) - \Phi(z-b)\}^{k_1} \phi(z)dz | Z_0 = z].$$

Next, we write $b = 2^{1/2}x$ and define a function $\beta(x)$ as

$$\beta(x) = \int_{-\infty}^{\infty} \left\{\Phi(z + 2^{1/2}x) - \Phi(z - 2^{1/2}x)\right\}^{k_1} \phi(z)dz.$$

It is easy to verify that

$$\beta'(x) = \int_{-\infty}^{\infty} k_1 2^{1/2} \left\{\Phi(z + 2^{1/2}x) - \Phi(z - 2^{1/2}x)\right\}^{k_1 - 1} (\phi(z + 2^{1/2}x) + \phi(z - 2^{1/2}x))\phi(z)dz$$

$$\begin{aligned}\beta''(x) \;=\; & \int_{-\infty}^{\infty} 2k_1(k_1-1) \left\{\Phi(2^{1/2}x + z) - \Phi(-2^{1/2}x + z)\right\}^{k_1-2} \\ & (\phi(2^{1/2}x + z) + \phi(-2^{1/2}x + z))^2 - 2k_1 \left\{\Phi(2^{1/2}x + z) - \Phi(-2^{1/2}x + z)\right\}^{k_1-1} \\ & \left((2^{1/2}x + z)\phi(2^{1/2}x + z) + (2^{1/2}x - z)\phi(-2^{1/2}x + z))\right) \phi(z)dz.\end{aligned}$$

Next, we write $g(x)$ for the function $\beta(bx^{\frac{1}{2}})$ and then we can express

$$\begin{aligned} g'(x) &= \frac{1}{2} b x^{-\frac{1}{2}} \beta'(bx^{\frac{1}{2}}) \\ g''(x) &= \frac{1}{4} b^2 x^{-1} \beta''(bx^{\frac{1}{2}}) - \frac{1}{4} b x^{-\frac{3}{2}} \beta'(bx^{\frac{1}{2}}) \end{aligned} \tag{18}$$

and

$$|g''(x)| \leq a_1 x^{-\frac{1}{2}} + a_2 x^{-1} + a_3 x^{-\frac{3}{2}},$$

where $a_1, \quad a_2 \quad a_3$ are positive constants. For the fine-tuned purely sequential procedure (10), one can verify that the distribution of N does not depend on the mean vector and $P(N < \infty) = 1$. Hence, by using "Theorem 3.2.1 of Mukhopadhyay and Solanky (1994)", we have

$$\inf_{a \to 0} P[CD|\mathcal{P}_N; \boldsymbol{\mu}^0(r_1, r_2, r_3, r_4), \;\; \sigma^2] > (-1) + 2E(g(\frac{N}{C})). \tag{19}$$

Next, taking a series expansion of the function $g(x)$ at $x = 1$ for random $Z \in (1, \frac{N}{C})$, we can write

$$g(x) = g(1) + g'(1)(x-1) + \frac{1}{2} g''(Z(x))(x-1)^2.$$

As shown above, we can express $|g''(x)| \leq \sum_{i=1}^{3} \frac{a_i}{x^{\alpha_i}}$, with $a_i > 0$ and $\alpha_1 > 0$, by applying "Lemma 3.5.1 of Mukhopadhyay and Solanky (1994)", we can show that for $m_0 > \frac{5}{k+1} + 1$, one will obtain

$$E(g(\frac{N}{C})) = g(1) + \frac{1}{C} g'(1) E(N - C) + \frac{1}{2C^2} E(g''(Z(x))(N - C)^2).$$

With $\alpha_1 = \frac{1}{2}, \alpha_2 = 1$, and $\alpha_3 = \frac{3}{2}$ and maximum of $(\alpha_1, \alpha_2, \alpha_3) = \frac{3}{2}$, the "Lemma 3.5.1 of Mukhopadhyay and Solanky (1994)" leads to

$$E(g(\frac{N}{C})) = g(1) + \frac{1}{C}g'(1)E(N-C) + \frac{1}{k+1}\frac{1}{C}g''(1) + o(\frac{1}{C}) \tag{20}$$

Following "Mukhopadhyay and Datta (1995)'s Theorem 2.1" with the function $g(.)$ defined as above, and ϵ as:

$$\epsilon = \epsilon\,(k, P^*) = (k+1)^{-1}[\nu - 2 + g''(1)(g'(1))^{-1}], \tag{21}$$

the $E(N)$ would simplify to

$$E(N) = C - (k+1)^{-1}[g''(1)(g'(1))^{-1}] + o(1), \text{ provided } m_0 > 1 + \frac{2}{k+1}. \tag{22}$$

Using the equations (20) and (22), one can obtain

$$E(g(\frac{N}{C})) = g(1) + o(\frac{1}{n^*}), \text{ provided } m_0 > 1 + \frac{5}{k+1}.$$

That is,

$$\inf_{a\to 0} P[CD|\mathcal{P}_N; \boldsymbol{\mu}^0(r_1, r_2, r_3, r_4),\ \ \sigma^2] > (-1) + 2g(1) + o(\frac{1}{n^*}), \text{ provided } m_0 > 1 + \frac{5}{k+1}.$$

Note that $g(1) = \frac{1}{2}(P^* + 1)$ to conclude part(iii) of the theorem. □

3. Simulation Study for the Fine-tuned Purely Sequential Procedure

In this section using the Monte Carlo simulation study, the "fine-tuned purely sequential procedure" (10) is replicated independently 5,000 times by picking different values of design constants to study how the asymptotic values provided in the Theorems 2.1 and 2.2 compare with the observed values when the procedure is simulated for small and moderate sample sizes. For the simulation results presented in the Table 1, we selected $k = 8$ and $P^* = 0.95$ and generated two independent normal populations at each of the four boundaries to generate data from LFC, as outlined in the Section 1. In order to obtain the value of constant b for given choice of k and P^*, we used the "Table 1 in Solanky and Zhou (2015)". For example, the value of constant b for $k = 8$ and $P^* = 0.95$ is 2.6959. By picking the values of the "optimal sample sizes" (=C) as 25, 50, 100, 200, 400, and 800 and taking $r = \frac{1}{3}$, we obtained the values of constant a which are reported in the Table 1. In a practical application, the values δ_1 and δ_4 would be chosen so as to reflect the "Good populations" and "Bad populations" based on the situation. Note that the difference between the "Good populations" and "Bad populations" is $a = \delta_4 - \delta_1$. In the Table 1, by picking $r = \frac{1}{3}$ we have divided the indifference-zone of Tong (1969) into three non-overlapping regions of equal size. The "middle" of these three regions would serve as the region which we will serve as Ω_M and the other two as indifference-zones Ω_{I_1} and Ω_{I_1} as defined in (5).

The procedure (10) starts with the initial sample size of $m_0 = 5$ observations from each of the 8 populations and the control population. Then, each additional observation is

Table 1: Performance of the Fine-Tuned Purely Sequential Procedure (10)

$(k = 8,\ P^* = 0.95,\ \rho = \frac{1}{3},\ \sigma = 1.0,\ m_0 = 5)$

C	ρa	a	$\bar{N}$	$std(\bar{N})$	$\bar{P}$	$std(\bar{P})$
25	1.5250	4.575	25.3246	0.0347	0.9710	0.0024
50	1.0784	3.2351	50.2660	0.0481	0.9720	0.0023
100	0.7625	2.2875	100.2612	0.0682	0.9776	0.0021
200	0.5392	1.6175	200.2736	0.0943	0.9746	0.0022
400	0.3813	1.1437	400.2050	0.1327	0.9772	0.0021
800	0.2696	0.8088	800.2242	0.1890	0.9750	0.0022

collected according to the following stopping rule (10). In the Table 1, the average value of the stopping time N and the proportion of the times all the 8 populations were partitioned correctly are reported as $\bar{N}$ and $\bar{P}$ respectively. The Table also reports the standard deviation of these two reported statistics as $std(\bar{N})$ and $std(\bar{P})$. For all the cases considered in the Table 1, the $\bar{P}$ is larger than the desired probability of correct decision which was selected to be 0.95. Also, the $\bar{N}$ matches the optimal sample size values $(=C)$ indicating that the "fine-tuned purely sequential procedure" (10) over-samples by less than half an observation on the average. The overall findings in the Table 1 confirm the theoretical results which were derived asymptotically in the Theorems 2.1 and 2.2 for the "fine-tuned purely sequential procedure" (10) are met even for small and moderate sample sizes.

Acknowledgements

It is my honor to receive an invitation to prepare this paper for the special issue of *Statistics and Applications* in honor of Professor Aloke Dey. I heartily thank the Guest Editors.

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.

Buzaianu, E. M. (2019). Selection among Bernoulli populations in comparison with a standard. *Sequential Analysis*, **38**, 184-198.

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**, 191-205.

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

Gupta, S. S. (1956). On a decision rule for a problem in ranking means. *Ph.D. dissertation, University of North Carolina*, Chapel Hill.

Paulson, E. (1952). On the comparison of several experimental categories with a control. *Annals of Mathematical Statistics*, **23**, 239-246.

Mukhopadhyay, N. and Datta, S. (1995). On fine-tuning a purely sequential procedure and associated second-order properties. *Sankhya: The Indian Journal of Statistics*, **57**, Series A, Pt.1: 100-117.

Mukhopadhyay, N. and Solanky, T. K. S. (1994). *Multistage Selection and Ranking Procedures*, New York: Dekker.

Roessler, E. B. (1946). Testing the significance of observations compared with a control. *Proceedings of American Society for Horticultural Science*, **47**, 249-251.

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**, 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**, 297-310.

Solanky, T. K. S. and Wu, Y. (2004). On unbalanced multistage methodologies for the partition problem, in *Proceeding of International Sri Lankan Statistical Conference: Version of Futuristic Methodologies*, B. Silva and N. Mukhopadhyay, eds., pp. 447-466, Sri Lanka: University of Peradeniya ; Melbourne: RMIT University.

Solanky, T. K. S. and Zhou, J. (2015). A generalization of the partition problem. *Sequential Analysis*, **34**, 483 – 503.

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.

Woodroofe, M. (1977). Second order approximations for sequential point and interval estimation. *Annals of Statistics*, **5**, 984-995.

Woodroofe, M. (1991). The role of renewal theory in sequential analysis. *Chapter 6, Handbook of Sequential Analysis*, (eds. B.K. Ghosh and P.K. Sen), 145-167. Marcel Dekker, Inc., New York.