Fall 12-20-2013

A Generalization of The Partition Problem in Statistics

Jie Zhou
jzhou3@uno.edu

A Generalization of The Partition Problem in Statistics

A Dissertation

Submitted to the Graduate Faculty of the
University of New Orleans
in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy
in
Engineering and Applied Science
Mathematics

by

Jie Zhou

M.S. University of New Orleans, 2008
M.S. Southeast University, China, 2005
B.S. Southeast University, China, 2002

December, 2013

Dedication

This thesis is dedicated to my mother, Mrs. Caifeng Wu and my husband, Jiajun.

Acknowledgments

The research of this dissertation was conducted during my Ph.D program at Department of Mathematics, University of New Orleans (UNO). Without the enormous help from the colleagues of Department of Mathematics, UNO and other institutes, this research is impossible.

I would like to express my greatest gratitudes to my advisor, Prof. Tumulesh Solamky, for his expertise in this field, and his continuous support and insightful advice in my Ph.D research and as well my life.

I want to thank the academic committee members: Prof. Linxiong Li, Prof. Xiaorong Li, Prof. Jairo Santanilla and Prof. Dongming Wei for their helpful suggestions on this dissertation.

I would like to thank Department of Mathematics for the support of this research.

Contents

Abstract

In this dissertation, the problem of partitioning a set of treatments with respect to a control treatment is considered. Since early 1950's a number of researchers have worked on this problem and have proposed competing alternative solutions to this statistical problem. In Tong (1979), the author proposed a formulation to solve this problem and since then hundreds of researchers and practitioners have used that formulation for the partition problem. However, Tong's formulation is somewhat rigid and unpractical for the practitioners if the distance between the "good" and the "bad" treatments is large. Under such a scenario, the indifference zone gets quite large and the undesirable feature of the Tong's formulation to partition the populations in the indifference zone, without any penalty, can potentially lead Tong's formulation to produce misleading or unpractical partitions. In this dissertation, a generalization of the Tong's formulation is proposed, under which, the treatments in the indifference zone are not partitioned as "good" or "bad", but are partitioned as a identifiable set. For this generalized partition, a fully sequential and a two-stage procedure is proposed and its theoretical properties are derived. The proposed procedures are also studied via Monte Carlo Simulation studies. The thesis concludes with some non-parametric partition procedures and the study of robustness of the various available procedures in the statistical literature.

Keywords: Control population, Correct partition, Nonparametric procedure, Probability of correct decision, Sequential Procedure, Two-stage procedure, Monte carlo simulations.

Chapter 1

Introduction

1.1 Backgraound

In everyday life, one decides on the best medicine, best fertilizer, best strategy or the best route for a destination, among the several available options. In the statistical literature, such selections have been routinely carried out under the area of multiple comparisons. A commonly used statistical tool called Analysis of Variance (ANOVA) has been used extensively by practitioners to test whether or not the given treatments under consideration are all same or not. Generally, the ANOVA test is followed by some multiple comparisons tests, such as, LSD, Tukey, Scheffe to name a few, to decide which treatments are different from one another. For example in clinical trials, the concern is comparing efficacy of the several essentially different varieties of drugs. If the hypothesis is formulated to test that these different varieties of drugs have the same efficacy, it will not be a realistic hypotheses. This is so because the efficacy will be naturally different for the essentially different varieties of drugs, and, with a sufficiently large sample a researcher can establish this fact at any preassigned level of significance. Thus, the experimenter's problem should not be only testing the efficacy of these drugs are equal or not, but rather to select the "best" one. The definition of the "best" would vary from situation to situation and it is generally for the experts in the area to dictate what best means in a given situation. For example, in some clinical trials. Sometimes, practitioners have even incorrectly used the ANOVA tests to even select the best treatment based on the ranking of the means without realizing that the ANOVA test is designed to test if the given treatments are all same or not. The ANOVA test is not designed to select the best treatment and one cannot associate a probability statement with the selected treatment as being the best via the ANOVA approach. In a pioneer work, Bechhofer (1954) introduced the concept of indifference-zone formulation and formulated some methodologies for the problem of selecting the best treatment from a set of several treatments. The formulation by Bechhofer had the desired property of selecting the best treatment with the pre-specified probability of correct selection. The formulation proposed by Bechhofer (1954) is referred to as the indifference-zone formulation in the statistical literature. Around the same time, Gupta (1956) formulated a strategy which controls the probability of correct selection in the whole parameter space, as opposed to the preference-zone which was the case under Bechhofer's approach. The formulation of Gupta (1956), selects a subset of random size which includes the best treatment with some pre-specified probability. The formulation proposed by Gupta (1956) is referred to as the subset-selection formulation in the statistical literature.

However, in many cases selecting the best treatment may not be good enough for an experimenter to choose it! The experimenter may want the best to be some "specified" amount better than what is already in use (known as Control or Standard). This requirement forced the researchers to seek out alternative formulations and thus the problem of comparisons with a control originated. The problem of comparisons with a control has been investigated by many researchers under different types of formulations, and under different criteria to be satisfied by an acceptable procedure. Among the early investigations, Paulson (1952), Dunnett (1955) and Roessler (1946) provided some of the earlier research related to comparisons with respect to a control population.

For the problem of partition with respect to a control, we address the theoretical and practical aspects of some commonly used sampling methodologies such as the purely sequential procedure and the two-stage procedure, and other multistage sampling methodologies. This thesis was written to consolidate research in the area and to improve upon the methodologies currently available in the statistical literature.

1.2 Current Methodology to Partition Problem

Assume that there are $(k+1)$ independent populations, $\pi_0, \pi_1, \cdots, \pi_k$, with unknown location parameters $\mu_i,\ i = 0,\ 1, \cdots, k$, but common scale parameter σ^2. Denote π_0 as the standard or control population. Given arbitrary but fixed constants δ_1 and δ_2, and $\delta_1 < \delta_2$, define three subsets along the lines of the Bechhofer's (1954) indifference zone formulation, as

$$\begin{aligned}\Omega_B &= \{\pi_i : \mu_i \leq \mu_0 + \delta_1,\ i = 1, \cdots, k\},\\ \Omega_I &= \{\pi_i : \mu_0 + \delta_1 < \mu_i < \mu_0 + \delta_2,\ i = 1, \cdots, k\},\\ \Omega_G &= \{\pi_i : \mu_i \geq \mu_0 + \delta_2,\ i = 1, \cdots, k\}.\end{aligned} \tag{1.2.1}$$

We refer to Ω_G as the set of "good populations" and Ω_B as the set of "bad populations". It is important to note that the choice of the constants δ_1 and δ_2 is generally provided by the experts in the area. We are interested in the correct partition of the populations belonging to two sets. The set Ω_I is considered as the indifference zone set and a correct decision puts no restrictions on the partition of the populations belonging to this set. Next, with high accuracy, we want to partition the set Ω into two disjoint subsets S_B and S_G, such that, $\Omega_B \subseteq S_B$ and $\Omega_G \subseteq S_G$. 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 S_B and S_G, such that

$$P\{CD|\boldsymbol{\mu}, \sigma^2, \wp\} \geq P^* \qquad \forall\ \boldsymbol{\mu} \in \boldsymbol{R}^{k+1},\ \sigma \in \boldsymbol{R}^+. \tag{1.2.2}$$

here $\mu = [\mu_0, \mu_1, \cdots, \mu_k]'$.

For the known σ^2 case, Tong (1969) gave a single-stage procedure for this problem. Tong (1969) considered the following decision rule to partition the set of treatments Ω, based on some appropriately *N* observations from each of the *k* treatments and the control population:

$$\begin{aligned} S_B &= \{\pi_i : \bar{X}_{iN} - \bar{X}_{0N} \leq \ d, \ i = 1, \cdots, k\}, \\ S_G &= \{\pi_i : \bar{X}_{iN} - \bar{X}_{0N} \geq \ d, \ i = 1, \cdots, k\}, \end{aligned} \tag{1.2.3}$$

where $\bar{X}_{iN}$ is the sample mean from $\pi_i, i = 0, 1, \cdots, k$. Let us write

$$\begin{aligned} &d = (\delta_1 + \delta_2)/2, \quad a = (-\delta_1 + \delta_2)/2, \quad \lambda = \sigma/a, \ \text{and,} \\ &m = \begin{cases} k/2 & \text{if } k \text{ is even;} \\ (k+1)/2 & \text{if } k \text{ is odd.} \end{cases} \end{aligned} \tag{1.2.4}$$

Next, using the above partition rule (1.2.3), Tong (1969) showed that the probability of correct decision for the normally distributed populations can be expressed as

$$\underset{\mu \in R^{k+1}}{Inf} \ P\,[CD] = \int_{-\infty}^{\left(\frac{1}{2}N\right)^{\frac{1}{2}}\big/\lambda} \cdots \int_{-\infty}^{\left(\frac{1}{2}N\right)^{\frac{1}{2}}\big/\lambda} \frac{|\Sigma|^{\frac{1}{2}}}{(2\pi)^{\frac{k}{2}}} \exp\left(-\frac{y'\Sigma^{-1}y}{2}\right) dy_1 \cdots dy_k, \tag{1.2.5}$$

where $y' = (y_1, \cdots, y_k)$ has a multivariate normal distribution with mean 0, the covariance matrix Σ is given by

$$\Sigma = \begin{pmatrix} 1 & & \frac{1}{2} & -\frac{1}{2} & \cdots & -\frac{1}{2} \\ & \ddots & & \vdots & \ddots & \vdots \\ \frac{1}{2} & & 1 & -\frac{1}{2} & \cdots & -\frac{1}{2} \\ -\frac{1}{2} & \cdots & -\frac{1}{2} & 1 & & \frac{1}{2} \\ \vdots & \ddots & \vdots & & \ddots & \\ -\frac{1}{2} & \cdots & -\frac{1}{2} & \frac{1}{2} & & 1 \end{pmatrix},$$

and the infimum is attained if $\mu_1 = \mu_2 = \cdots = \mu_m = \mu_0 + \delta_1$ and $\mu_{m+1} = \mu_{m+2} = \cdots = \mu_k = \mu_0 + \delta_2$. In the statistical literature, this parameter configuration is known as the *least favorable* *configuration* (LFC). Next, suppose *b* is a constant satisfying

$$P^* = \int_{-\infty}^{b} \cdots \int_{-\infty}^{b} \frac{|\Sigma|^{\frac{1}{2}}}{(2\pi)^{\frac{k}{2}}} \exp\left(-\frac{y'\Sigma^{-1}y}{2}\right) dy_1 \cdots dy_k. \qquad (1.2.6)$$

Then, if we take a sample of size *N*, where $N \geqslant 2\lambda^2 b^2$, and partition the *k* treatments according to the partition rule (1.2.3), we have

$$P\left[CD\right] \geqslant P^*, \quad \forall \mu \in R^{k+1}, \quad \sigma \in R^+. \qquad (1.2.7)$$

The values of *b* have been tabulated in the Table 1 of Tong (1969) and also in Chapter 10 of Gibbons et al. (1977). The single-stage procedure is designed for the known σ^2 case. For the unknown σ^2 case, Tong (1969) also constructed a two-stage and a purely sequential procedure. Recently, for the unknown σ^2 case, Datta and Mukhopadhyay (1998) have constructed a fine-tuned purely sequential procedure and some other multistage methodologies, emphasizing the second-order asymptotics. In order to minimize the sampling from too inferior or too superior populations, which is an in-built feature of the vector-at-a-time sampling design, Solanky (2001) has constructed an elimination type fully-sequential procedure which reduces the sampling cost considerably. However, the sequential procedures are known to be operationally inconvenient and rather cumbersome to use, as decisions and computations need to be carried out after each stage of the sampling process. With that as the motivation, Solanky (2006) constructed a two-stage procedure with elimination which eliminates too inferior or too superior populations after the stage one of the sampling process and in the stage two only continues sampling from the competing treatments. For this problem, Solanky and Wu (2004) considered an unbalanced sampling design which exploits collecting a larger sample size from the control population in order to reduce the sample size from the competing treatments.

Chapter 2

A Generalization of the Partition Problem

2.1 Introduction of the New Methodology

The partition methodology of Tong (1969), as described in (1.2.3) is designed to partition the given k populations into two sets: "Good populations" and "Bad populations" utilizing the Bechhofer(1954) indifference zone formulation. This methodology Means that the populations in the indifference zone can be partitioned as "Good population" or as a "Bad population" without any penalty and without changing the probability of correct decision. This feature of the Tong (1969) methodology is considered undesirable and intuitively it can make the methodology unattractive to the practitioners. Let's explain this via an illustration. Suppose in some clinical trials dealing with curability of a disease it is reasonable to assume at least 60% curability is "Good" effectiveness and less than 10% curability is "Bad" effectiveness. One will note that in this fictitious illustration, following the Tong's (1969) formulation (1.2.1), any drug with effectiveness between 10% to 60% would belong to the indifference zone. And, the decision rule of Tong (1969) is designed to partition all the k populations as either "Good" or "Bad" as defined in (1.2.3). Now, consider a drug which has 55% curability, it is possible that following Tong's (1969) rule this drug may get partitioned as a "Bad population" and on the other hand a drug which has say 12% curability may get partitioned as a "Good population". And, such a partition would not alter the probability of correct decision. Intuitively, this ambiguity is due to the fact that Tong's (1969) procedure is designed to partition all the k populations, including the ones in the indifference zone, as either a

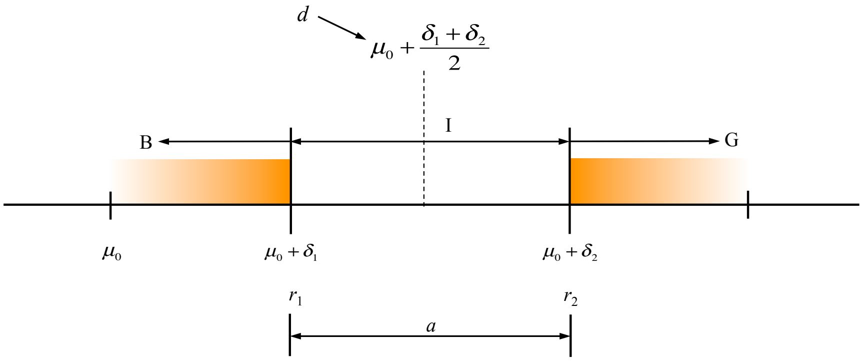

Figure 2.1: Depiction of the Partition Problem Based on Tong's Method

"Good populations" or as a "Bad populations". In this thesis, the partition problem is generalized so that the experimenter has essentially the choice of not partitioning the populations in the indifference zone as either "Good populations" or as "Bad populations", but rather such populations can be partitioned as a separate identifiable group. In addition, under the proposed generalization, there would be some penalty associated with incorrect partition of the populations belonging to the indifference zone. In this chapter, first we introduce such a generalization of the Tong's (1969) methodology and then we will design a fully-sequential sampling methodology to carry out the partitioning of the k populations. Following this, the first and second-order theoretical properties are derived and verified using Monte Carlo Simulation studies. We will also provide the values of the design constants which are needed to implement the fully-sequential sampling methodology.

In the Figure (2.1), we have visualized the partition rule constructed in Tong (1969). Next, in the Figure (2.2), we have depicted the conceptual visualization of the proposed generalization of the Tong's (1969) partitioning methodology using the location parameter of the normal distribution to define "Good populations", "Bad populations", and the "Medium or Indifferent populations".

As before, suppose there are *(k+1)* independent normally distributed populations, $\pi_0, \pi_1, \cdots, \pi_k$, with unknown location parameters $\mu_i, i = 0, 1, \cdots, k$, and common scale parameter σ^2. Denote π_0 as the standard or the control population. Given arbitrary but fixed constants δ_1, δ_2, δ_3 and δ_4,

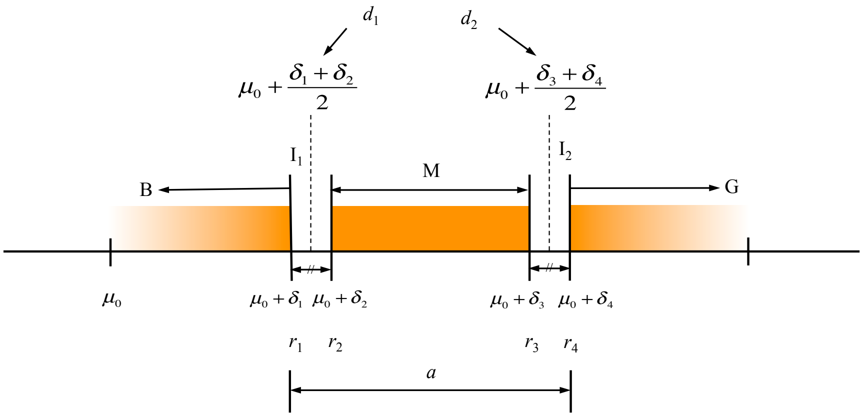

Figure 2.2: Depiction of the Generalized Partition Problem

$\delta_1 < \delta_2 < \delta_3 < \delta_4$, define five subsets of Ω along the lines of Bechhofer's (1954) indifference-zone formulation, as

$$\begin{aligned}
\Omega_B &= \{\pi_i : \mu_i \leq \mu_0 + \delta_1,\ i = 1, \cdots, k\}, \\
\Omega_{I_1} &= \{\pi_i : \mu_0 + \delta_1 < \mu_i \leq \mu_0 + \delta_2,\ i = 1, \cdots, k\}, \\
\Omega_M &= \{\pi_i : \mu_0 + \delta_2 < \mu_i \leq \mu_0 + \delta_3,\ i = 1, \cdots, k\}, \qquad (2.1.1) \\
\Omega_{I_2} &= \{\pi_i : \mu_0 + \delta_3 < \mu_i \leq \mu_0 + \delta_4,\ i = 1, \cdots, k\}, \\
\Omega_G &= \{\pi_i : \mu_i > \mu_0 + \delta_4,\ i = 1, \cdots, k\}.
\end{aligned}$$

Let us write

$$d_1 = (\delta_1 + \delta_2)/2, \qquad d_2 = (\delta_3 + \delta_4)/2, \qquad a_1 = (\delta_2 - \delta_1)/2, \qquad a_2 = (\delta_4 - \delta_3)/2,$$

to denote some constants which will be used to denote several midpoints and distances in this chapter. It is important to note that the generalization outlined above relies upon the construction of two indifference-zones Ω_{I_1} and Ω_{I_2}. However, the size of these two indifference-zones will not depend upon the experimenters choice of δ_1 and δ_4 and the experimenter will have full control over how large or how small these two indifference-zones could be without impacting the definition of "Good populations" and "Bad populations".

2.2 A Single-Stage Procedure

Based on a sample of size n, let X_{ij} denote the jth observation from the population π_i with density function

$$f\left(X_{ij}\right)=\frac{1}{\sqrt{2\pi\sigma}}\exp^{-\frac{\left(X_{ij}-\mu_i\right)^2}{2\sigma^2}}, \quad i=0,1,\cdots,k; \quad j=1,\cdots,n.$$

Where the parameter μ_i is the mean for population π_i, and σ^2 is the common population variance for all π_i's, $i=0,1,\cdots,k$. Define

$$\bar{X}_i=\sum\nolimits_{j=1}^{n} X_{ij}/n, \qquad S_i^2=\sum\nolimits_{j=1}^{n}\left(X_{ij}-\bar{X}_i\right)^2/\left(n-1\right),$$

for $i=0,1,\cdots,k$, $j=1,\cdots,n$. Based on a sample of size n, a natural estimator for σ^2 is given by

$$S_n^2=\frac{\sum_{j=0}^{k}S_j^2}{k+1}. \tag{2.2.2}$$

However, note that through out this section, we will assume that σ^2 is a known parameter. Next, along the lines of Tong (1969), we propose the partition rule based on the difference of sample means as the following:

$$\begin{array}{rcl} S_B & = & \{\pi_i : \bar{X}_{iN}-\bar{X}_{0N}\leq d_1, i=1,\cdots,k\}, \\ S_M & = & \{\pi_i : d_1\leq \bar{X}_{iN}-\bar{X}_{0N}\leq d_2, i=1,\cdots,k\}, \\ S_G & = & \{\pi_i : \bar{X}_{iN}-\bar{X}_{0N}\geq d_2, i=1,\cdots,k\}, \end{array} \tag{2.2.3}$$

where S_B is the set of "Bad populations", S_M is the set of "Medium or Indifferent populations", and S_G is the set of "Good populations". Note that without altering the definition of the "Good populations" or "Bad populations", which are selected by the practitioners, now one can control the size of the "Medium or Indifferent populations" by selecting the appropriate values of the constants δ_2 and δ_3. These two constants also control the size of the two indifference-zones as illustrated in the Figure (2.2).

Next, we will consider a parametric configuration which is most unfavorable for the partition problem on hand. Such parametric configuration is known as the *least favorable configuration* (LFC) in the statistical literature. It is clear that for a mean vector to be a *LFC* under the procedure (2.2.3), the set Ω_{I_1} and the set Ω_{I_2} must be empty. Let us redefine the design constants to introduce some symmetry which would play a key role in establishing the LFC. We write:

(1) $\delta_2 - \delta_1 = \delta_4 - \delta_3 = ra$, where r is a known number and $0 < r < \frac{1}{2}$, $a = \delta_4 - \delta_1$,

(2) $r_2 + r_3 = \left[\frac{k}{2}\right] = k'$, $r_1 + r_4 = k - k'$, $r_2 = \left[\frac{k'}{2}\right]$, $r_3 = k' - r_1$, $r_1 = \left[\frac{k-k'}{2}\right]$, $r_4 = k - k' - r_1$, 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.

Note that the requirement (1) above forces the two indifference-zone's to be symmetric and the length of the two indifference-zones have been expressed in terms of the distance between the "Good populations" and the "Bad populations" via the constant r. When r is close to $\frac{1}{2}$, the size of the two indifference-zone's becomes small and the size of "Medium or Indifferent populations" gets larger. And when r is close to 0, the size of the two indifference-zone's becomes small and the size of "Medium or Indifferent populations" is smaller. As noted earlier, the constant r does not depend on the definition of the "Good populations" and the "Bad populations" and thus allows the experimenter to control the precision of the partition without altering the baseline requirements.

It is also important to note that without the symmetry requirement (1) above, there does not exist any general solution to this partition problem. This is so because, without this requirement, the partition probability would actually depend on the specific parametric configuration and there would not be any parametric configuration that is LFC as such.

Under the requirement (2), we are forcing the number of populations to be situated on all the four boundaries and in equal number. Intuitively, the symmetry requirement in condition (1) above, ensures that this would be the parametric configuration which is the LFC. The issue of

LFC is visited again in later this Chapter and we have shown via simulations that this parametric configuration described in condition (2) above is indeed the LFC.

Theorem 1 *Assuming σ^2 is known, the generalized partition problem* (2.1.1) *has*

$$P\Big[CD|\boldsymbol{\mu}^0(r_1,r_2,r_3,r_4),\sigma^2\Big] \geq P^* \tag{2.2.4}$$

for the partition rule (2.2.3)*, provided that the sample size is at least $n^* = \frac{8b^2\sigma^2}{(ra)^2}$. The constant $b = b(k,P^*)$ is the solution of an integral equation* (2.2.9).

Proof. Without the loss of generality assume that the first r_1 populations have the mean $\mu_0+\delta_1$ the second r_2 populations have the mean $\mu_0+\delta_2$, the third set of r_3 populations have the mean $\mu_0+\delta_3$, and the last set of r_4 populations have the mean $\mu_0+\delta_4$. let us denote this parametric configuration as $\boldsymbol{\mu}^0(r_1,r_2,r_3,r_4)$. Then, the probability of *correct decision* can be expressed as

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

Next, under the LFC the above expression simplifies to:

$$\begin{aligned}
&P\Big[CD|\boldsymbol{\mu}^0(r_1,r_2,r_3,r_4),\ \ \sigma^2\Big] = P\Big[\ \big((\bar{X}_i-\mu_i)-(\bar{X}_0-\mu_0)\big)\Big/\sqrt{\frac{2\sigma^2}{n}} < (d_1-\delta_1)\Big/\sqrt{\frac{2\sigma^2}{n}},\\
&(d_1-\delta_2)\Big/\sqrt{\frac{2\sigma^2}{n}} < \big((\bar{X}_j-\mu_j)-(\bar{X}_0-\mu_0)\big)\Big/\sqrt{\frac{2\sigma^2}{n}} < (d_2-\delta_2)\Big/\sqrt{\frac{2\sigma^2}{n}},\\
&(d_1-\delta_3)\Big/\sqrt{\frac{2\sigma^2}{n}} < \big((\bar{X}_m-\mu_m)-(\bar{X}_0-\mu_0)\big)\Big/\sqrt{\frac{2\sigma^2}{n}} < (d_2-\delta_3)\Big/\sqrt{\frac{2\sigma^2}{n}},\\
&\big((\bar{X}_l-\mu_l)-(\bar{X}_0-\mu_0)\big)\Big/\sqrt{\frac{2\sigma^2}{n}} > (d_2-\delta_4)\Big/\sqrt{\frac{2\sigma^2}{n}},\\
&1\leq i\leq r_1, r_1+1\leq j\leq r_1+r_2,, r_1+r_2+1\leq m\leq r_1+r_2+r_3, r_1+r_2+r_3+1\leq l\leq k\Big]
\end{aligned}$$

$$\begin{aligned}
&= P\Big[\big((\bar{X}_i-\mu_i)-(\bar{X}_0-\mu_0)\big)\Big/2\sqrt{\frac{2\sigma^2}{n}} < ra\Big/2\sqrt{\frac{2\sigma^2}{n}},\\
&\quad -ra\Big/2\sqrt{\frac{2\sigma^2}{n}} < \big((\bar{X}_j-\mu_j)-(\bar{X}_0-\mu_0)\big)\Big/2\sqrt{\frac{2\sigma^2}{n}} < (2a-3ra)\Big/2\sqrt{\frac{2\sigma^2}{n}},\\
&\quad -(2a-3ra)\Big/2\sqrt{\frac{2\sigma^2}{n}} < \big((\bar{X}_m-\mu_m)-(\bar{X}_0-\mu_0)\big)\Big/2\sqrt{\frac{2\sigma^2}{n}} < ra\Big/2\sqrt{\frac{2\sigma^2}{n}},\\
&\quad \big((\bar{X}_l-\mu_l)-(\bar{X}_0-\mu_0)\big)\Big/2\sqrt{\frac{2\sigma^2}{n}} > -ra\Big/2\sqrt{\frac{2\sigma^2}{n}},\\
&\quad 1\le i\le r_1, r_1+1\le j\le r_1+r_2, r_1+r_2+1\le m\le r_1+r_2+r_3, r_1+r_2+r_3+1\le l\le k\Big]\\
&= P\Big[Y_i < ra\Big/2\sqrt{\frac{2\sigma^2}{n}}, 1\le i\le r_1, r_1+r_2+r_3+1\le i\le k,\\
&\quad -(2a-3ra)\Big/2\sqrt{\frac{2\sigma^2}{n}} < Y_j < ra\Big/2\sqrt{\frac{2\sigma^2}{n}}, r_1+1\le j\le r_1+r_2+r_3\Big], \qquad (2.2.5)
\end{aligned}$$

where, $Y_i = ((\bar{X}_i-\mu_i)-(\bar{X}_0-\mu_0))\Big/\sqrt{\frac{2\sigma^2}{n}}$, for $0<i\le r_1$, and $r_1+r_2<i\le r_1+r_2+r_3$, $Y_i = -((\bar{X}_i-\mu_i)-(\bar{X}_0-\mu_0))\Big/\sqrt{\frac{2\sigma^2}{n}}$, for $r_1<i\le r_1+r_2$, and $r_1+r_2+r_3<i\le k$. Note that under the parameter configuration $\boldsymbol{\mu}^0(r_1,r_2,r_3,r_4)$, Y_i has the standard normal distribution, $i=1,\cdots,k$. Let us define the $(k\times k)$ covariance matrix $\Sigma_Y=(\sigma_{ij})$ as

$$\begin{aligned}
\sigma_{ij} &= 1 \text{ for } i=j\\
&= 1/2 \text{ for } i\neq j, \text{ and } i,\, j\in[1,r_1]\cup[r_1+r_2+1,r_1+r_2+r_3],\\
&\quad \text{or } i,\, j\in[r_1+1,r_1+r_2]\cup[r_1+r_2+r_3+1,k]\\
&= -1/2 \text{ for } i\in[1,r_1]\cup[r_1+r_2+1,r_1+r_2+r_3],\ j\in[r_1+1,r_1+r_2]\cup[r_1+r_2+r_3+1,k].
\end{aligned}$$

Let us denote $ra\Big/2\sqrt{\frac{2\sigma^2}{n}}=b$ and $(2a-3ra)\Big/2\sqrt{\frac{2\sigma^2}{n}}=c$. Note that the probability of *correct decision* can be simplified as

$$P\Big[CD\Big] = P\Big[Y_i<b, 1\le i\le r_1, r_1+r_2+r_3+1\le i\le k, -c<Y_j<b, r_1+1\le j\le r_1+r_2+r_3\Big].$$

If we express

$$A = Y_i < b; \;\; 1 \leq i \leq r_1, \;\; r_1 + r_2 + r_3 + 1 \leq i \leq k,$$
$$B = -c < Y_j < b; \;\; r_1 + 1 \leq j \leq r_1 + r_2 + r_3,$$

then the probability of *correct decision* can be stated as

$$P\left[CD\right] = P\left[A \cap B\right] = P\left[A\right] + P\left[B\right] - P\left[A \cup B\right] \geqslant P\left[A\right] + P\left[B\right] - 1 \geqslant P^* \qquad (2.2.6)$$

and the equality holds if $P\left[A \cup B\right] = 1$.

In the above expression, the two probability expressions can be expressed as:

$$P\left[A\right] = \int_{-\infty}^{b}\int_{-\infty}^{b} ... \int_{-\infty}^{b} (2\pi)^{-(k-k')/2} |\Sigma_a|^{-1/2} \exp\left(-\frac{1}{2} y' \Sigma_{_a}^{-1} y\right) \prod_{i=1}^{(k-k')/2} dy_i \qquad (2.2.7)$$

$$P\left[B\right] = \int_{-c}^{b}\int_{-c}^{b} ... \int_{-c}^{b} (2\pi)^{-k'/2} |\Sigma_b|^{-1/2} \exp\left(-\frac{1}{2} y' \Sigma_{_b}^{-1} y\right) \prod_{i=1}^{k'/2} dy_i \qquad (2.2.8)$$

where the two covariance matrices Σ_a and Σ_b are given by

$$\Sigma_a = \begin{pmatrix} \begin{pmatrix} 1 & \cdots & \frac{1}{2} \\ \vdots & \ddots & \vdots \\ \frac{1}{2} & \cdots & 1 \end{pmatrix}_{r_1 \times r_1} & \begin{pmatrix} -\frac{1}{2} & \cdots & -\frac{1}{2} \\ \vdots & \ddots & \vdots \\ -\frac{1}{2} & \cdots & -\frac{1}{2} \end{pmatrix}_{r_1 \times r_4} \\ \begin{pmatrix} -\frac{1}{2} & \cdots & -\frac{1}{2} \\ \vdots & \ddots & \vdots \\ -\frac{1}{2} & \cdots & -\frac{1}{2} \end{pmatrix}_{r_4 \times r_1} & \begin{pmatrix} 1 & \cdots & \frac{1}{2} \\ \vdots & \ddots & \vdots \\ \frac{1}{2} & \cdots & 1 \end{pmatrix}_{r_4 \times r_4} \end{pmatrix}_{(k-k') \times (k-k')}$$

$$\Sigma_b = \begin{pmatrix} \begin{pmatrix} 1 & \cdots & \frac{1}{2} \\ \vdots & \ddots & \vdots \\ \frac{1}{2} & \cdots & 1 \end{pmatrix}_{r_2 \times r_2} & \begin{pmatrix} -\frac{1}{2} & \cdots & -\frac{1}{2} \\ \vdots & \ddots & \vdots \\ -\frac{1}{2} & \cdots & -\frac{1}{2} \end{pmatrix}_{r_2 \times r_3} \\ \begin{pmatrix} -\frac{1}{2} & \cdots & -\frac{1}{2} \\ \vdots & \ddots & \vdots \\ -\frac{1}{2} & \cdots & -\frac{1}{2} \end{pmatrix}_{r_3 \times r_2} & \begin{pmatrix} 1 & \cdots & \frac{1}{2} \\ \vdots & \ddots & \vdots \\ \frac{1}{2} & \cdots & 1 \end{pmatrix}_{r_3 \times r_3} \end{pmatrix}_{k' \times k'}$$

One will note that, since $0 < r < \frac{1}{2}$, we have

$$-a\Big/\sqrt{\frac{2\sigma^2}{n}} < (3r-2)\,a/2\sqrt{\frac{2\sigma^2}{n}} < -a\Big/4\sqrt{\frac{2\sigma^2}{n}},$$

and $0 < ra\Big/2\sqrt{\frac{2\sigma^2}{n}} < a\Big/4\sqrt{\frac{2\sigma^2}{n}}$. Combining these two, one can obtain

$$-ra\Big/2\sqrt{\frac{2\sigma^2}{n}} > -a/4\sqrt{\frac{2\sigma^2}{n}} > (3r-2)\,a\Big/2\sqrt{\frac{2\sigma^2}{n}}.$$

That is, $-b > -c$, so we can claim that

$$P\left[B\right] > \int_{-b}^{b}\int_{-b}^{b}...\int_{-b}^{b}(2\pi)^{-k'/2}|\Sigma_b|^{-1/2}\exp\left(-\frac{1}{2}y'\Sigma_{_b}^{-1}y\right)\prod_{i=1}^{k'/2}dy_i = P_2.$$

Hence, if $P\left[A\right] + P_2 - 1 > P^*$, then $P\left[A\right] + P\left[B\right] - 1 > P^*$. Note that $P(A)$ is associated with correct classification of the populations belonging to the set of “Good” or “Bad” populations, whereas, $P(B)$ is associated with correct classification of the populations belonging to the set of “Medium of Indifferent” populations. If the correct probability equally distributed, that is, $P(A) = P(B)$ then $P\left[A\right] = P_2 \geqslant \frac{P^*+1}{2}$. Therefore, $P_2 > \frac{P^*+1}{2}$, $P\left[A\right] \geqslant \frac{P^*+1}{2}$. Next, let b be the solution of the integral equation:

$$\int_{-b}^{b}\int_{-b}^{b}\dots\int_{-b}^{b}(2\pi)^{-k'/2}|\Sigma_b|^{-1/2}\exp\left(-\frac{1}{2}y'\Sigma_b^{-1}y\right)\prod_{i=1}^{k'/2}dy_i=\frac{P^*+1}{2}. \qquad (2.2.9)$$

Then, if n^* is the smallest integer satisfying

$$n^* \geq \frac{8\sigma^2 b^2}{(ra)^2} \qquad (2.2.10)$$

then the probability requirement (2.2.6) is satisfied. This completes the proof of the Theorem 1.

■

Remark 1 *Note that the constant r satisfying $0 < r < \frac{1}{2}$ determines the potential size of the "Medium or Indifferent" set and the constant a equals the distance between the "Good" and the "Bad" populations ($a = \delta_4 - \delta_1$). As r approaches $\frac{1}{2}$, the proposed generalized rule will approach the partition rule proposed in Tong (1969).*

The solution $b = b\,(k, P^*)$ of (2.2.9) is the equi-coordinate percentage point of a *k'*-dimensional multivariate normal distribution with mean vector 0 and the covariance matrix Σ_b described above. k is the total number of populations, and $k' = \left[\frac{k}{2}\right]$ ($[x]$ equals $\frac{x}{2}$ if x is even and $\frac{x+1}{2}$ if x is odd). The values of b as a function of P^* and k have been tabulated in the Table (2.1). The values of the constant b satisfying the equation (2.2.9) were calculated by Monte Carlo integration and Bisection method.

The single-stage procedure, which assumes that σ^2 is known, starts with the computation of the value of n^* as defined in (2.2.10). The value of the design constant b comes from Table (2.1) for the given value of k and the target value of the probability of correct decision P^*. The design parameters δ_1 and δ_4 are provided by the experimenter based on the definition of the "Good" and "Bad" populations. Note that $a = \delta_4 - \delta_1$. The experimenter tolerance for the misclassification of the populations in-between the "Good" and "Bad" populations determines the value of the design constant r. The smaller value of r, $0 < r < \frac{1}{2}$, the larger the value of n^* would be and smaller the size of the "Medium" classification set would be. After, the value of n^* is computed, a sample of

Table 2.1: Equi-coordinate percentage points of b of a multivariate normal distribution with mean vector 0 and covariance matrix $\Sigma_{k' \times k'}$

k	P^*						
	0.50	0.75	0.80	0.90	0.95	0.98	0.99
1	1.1504	1.5341	1.6448	1.9600	2.2414	2.5753	2.8052
2	1.1504	1.5341	1.6448	1.9600	2.2414	2.5753	2.8052
3	1.4538	1.8128	1.9165	2.2122	2.4783	2.7957	3.0202
4	1.4538	1.8128	1.9165	2.2122	2.4783	2.7957	3.0202
5	1.6146	1.9611	2.0625	2.3496	2.6082	2.9100	3.1298
6	1.6146	1.9611	2.0625	2.3496	2.6082	2.9100	3.1298
7	1.7218	2.0618	2.1603	2.4420	2.6959	2.9964	3.2030
8	1.7218	2.0618	2.1603	2.4420	2.6959	2.9964	3.2030
9	1.8019	2.1377	2.2348	2.5146	2.7703	3.0428	3.2910
10	1.8019	2.1377	2.2348	2.5146	2.7703	3.0428	3.2910
11	1.8644	2.1954	2.2922	2.5666	2.8110	3.1233	3.3031
12	1.8644	2.1954	2.2922	2.5666	2.8110	3.1233	3.3031
13	1.9160	2.2452	2.3401	2.6089	2.8531	3.1648	3.3634
14	1.9160	2.2452	2.3401	2.6089	2.8531	3.1648	3.3634
15	1.9606	2.2875	2.3831	2.6602	2.8980	3.1767	3.3748
16	1.9606	2.2875	2.3831	2.6602	2.8980	3.1767	3.3748
17	1.9979	2.3229	2.4165	2.6806	2.9252	3.1852	3.4117
18	1.9979	2.3229	2.4165	2.6806	2.9252	3.1852	3.4117
19	2.0318	2.3552	2.4513	2.7271	2.9596	3.2745	3.4792
20	2.0318	2.3552	2.4513	2.7271	2.9596	3.2745	3.4792

Table 2.2: Simulation Result for Single Stage

n^*	$\bar{P}\left(std\left(\bar{P}\right)\right)$					
	$(2,2,2,2)$	$(3,1,1,3)$	$(1,3,3,1)$	$(1,6,1)$	$(2,4,2)$	$(4,0,0,4)$
25	0.9728	0.9760	0.9761	0.9920	0.9869	0.9765
	0.0012	0.0011	0.0011	0.0006	0.0008	0.0011
50	0.9702	0.9746	0.9746	0.9923	0.9855	0.9741
	0.0024	0.0011	0.0011	0.0006	0.0008	0.0011
100	0.9751	0.9757	0.9766	0.9924	0.9862	0.9769
	0.0011	0.0011	0.0011	0.0006	0.0008	0.0011
200	0.9740	0.9763	0.9752	0.9927	0.9863	0.9749
	0.0011	0.0011	0.0011	0.0006	0.0008	0.0011
300	0.9743	0.9760	0.9771	0.9929	0.9872	0.9752
	0.0011	0.0011	0.0011	0.0006	0.0008	0.0011
400	0.9744	0.9760	0.9745	0.9937	0.9861	0.9759
	0.0011	0.0011	0.0011	0.0005	0.0008	0.0011

size at least n^* is collected from all the k populations and the control population. After the sample means are computed, the partition rule (2.1.1) is used to partition the populations with respect to the control population. Theorem 1 guarantees that the probability of correct decision would be at least P^*.

In Table (2.2), we have summarized the performance of the single-stage procedure under various parametric configurations for the case when $k = 8$. Note that the configuration (2,2,2,2) is the Least favorable Configuration under which there are all equal number of populations on all the four boundaries as shown in Figure 2.2. And the other parametric configurations summarized in Table (2.2) are (3,1,1,3), (1,3,3,1), (4,0,0,4), (1,6,1) and (2,4,2). Note that in the last two parametric configurations, not all the populations are located on the boundaries as such. For example, in (2,4,2) there are 2 populations each on the left and the right most boundary and there are 4 populations in the midpoint of the two middle boundaries.

The findings in Table (2.2), confirm the theoretical results derived in the Theorem 1 that the generalized partition procedure satisfies the probability requirement (2.2.5) for all the parametric configurations we studied. And one will also note that the estimated value of the probability of the correct decision is least for the parametric configuration (2,2,2,2). That is, this is the configuration for which the $P(CD)$ is least, or, this is the LFC among all the parametric configurations.

Next, for the unknown σ^2 case, we develop a purely-sequential procedure and a two-stage procedure, for the partition problem which will guarantee the probability of correct decision to be at least P^*.

2.3 Purely Sequential Procedure

In this section, we will construct a purely sequential procedure along the lines of Mukhopadhyay and Solanky (1994). One may also see Robbins et al.(1968), and Robbins (1959) to review a brief history of the purely sequential procedures. Recall that $n^* = 8\sigma^2 b^2/(ra)^2$, which is the optimal fixed sample size required from each population, had σ been known. As before, based on the sample of size n, let X_{ij} denote the jth observation from the population π_i with density function

$$f\left(X_{ij}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{\left(X_{ij}-\mu_i\right)^2}{2\sigma^2}},$$

$i = 0, 1, \cdots, k$; $j = 1, \cdots, n$. Where the parameter μ_i is the mean for population π_i, and σ^2 is the common population variance for all π_i's, $i = 0, 1, \cdots, k$. Note that based on a sample of size n, a natural estimator for σ^2 is the usual pooled estimator of variance as defined below:

$$\begin{aligned} \bar{X}_j &= \frac{\sum_{i=1}^{n} X_i j}{n}, \quad j = 0, 1, \cdots, k. \\ S_j^2 &= \frac{\sum_{i=1}^{n} \left(X_{ij} - \bar{X}_j\right)^2}{n-1} \\ S_n^2 &= \frac{\sum_{j=0}^{k} S_j^2}{k+1}. \end{aligned} \tag{2.3.11}$$

We start the purely sequential procedure with $m\,(\geq 2)$ observations from each of the k populations and the control population. Then, keep taking one additional sample at a time from each of the k populations and the control population according to the following stopping rule:

$$N = N\left(a\right) = \inf\left\{n \geq m : n \geq \frac{8b^2 S_n^2}{\left(ra\right)^2}\right\}. \tag{2.3.12}$$

For fixed $\tilde{\mu}$, σ^2, m, r and a, we first prove that the purely sequential procedure as defined about terminates with probability 1. Note that $P\left(N < \infty\right) = 1 - \lim_{n\to\infty} P\left(N > n\right) \geqslant 1 - \lim_{n\to\infty} P\left\{n < 8b^2 S_n^2 \left(ra\right)^{-2}\right\} = 1$, since $S_n^2 \to \sigma^2$ w.p. 1 as $n \to \infty$.That is, one has $P\left(N < \infty\right) = 1$, in other words, the purely sequential procedure (2.3.12) terminates with probability one. Based on the totality of all samples, that is having $X_{i1}, \cdots, X_{iN}$ from π_i, $i = 0, 1, \cdots, k$, one would next implement the partition rule $\wp_N$ given by (2.2.3) to obtain the generalized partition of the k populations with respect to the control population. Next, we derive some theoretical properties of the purely sequential procedure (2.3.12).

Theorem 2 *For the purely sequential procedure* (2.3.12)*, we have as* $a \to 0$*:*

(i) $N/n^* \to 1$ w.p. *1;*

(ii) $E\left(N/n^*\right) \to 1$*;*

(iii) $n^{*\frac{1}{2}}\left(N - n^*\right) \to N\left(0, 2\big/\left(k+1\right)\right)$*;*

(iv) $\liminf P\left(CD\right) > P^*$ under the LFC*;*

where $n^* = 8b^2\sigma^2\big/(ra)^2$ *and the constant b comes from Table* (2.1).

Proof. Utilizing Lemma 1 of Chow and Robbins (1965), it follows that as $a \to 0$, we have $N \to \infty$, w.p. 1, $S_N^2 \to \sigma^2$ w.p. 1, and $S_{N-1}^2 \to \sigma^2$ w.p. 1. The above sequential procedure agree with $N = N_\nu = \inf\left\{n \geq m : n > \Psi_\nu T_n\right\}$ (equation 2.4.1 in Mukhopadhyay and Solanky (1994)), where $\Psi_\nu = \frac{8b^2}{(ra)^2}$, $T_n = S_n^2$. The basic inequality (equation 2.4.3 in Mukhopadhyay and Solanky (1994)) simplifies to

$$\frac{8b^2S_N^2}{(ra)^2} \leq N \leq m + \frac{8b^2S_{N-1}^2}{(ra)^2}. \tag{2.3.13}$$

Now divide throughout (2.3.13) by n^* and take limits as $a \to 0$. This leads to part (i). Next, consider the equation (2.3.11), invoke the Helmert's orthogonal transformation to construct W_1', W_2', $\cdots$ which are i.i.d. $(k+1)^{-1}\sigma^2\chi_{k+1}^2$ so that we can express

$$S_n^2 = (n-1)^{-1}\sum\nolimits_{i=1}^{n-1} W_i'. \tag{2.3.14}$$

Let $W^* = \sup\limits_{n\geq 2}\left\{(n-1)^{-1}\sum\limits_{i=1}^{n-1} W_i'\right\}$. From the right hand side of the basic inequality in (2.3.14), we can write $N \leq m + \frac{8b^2S_{n-1}^2}{(ra)^2}$ as $N \leq m + \frac{8b^2W^*}{(ra)^2}$, that is $Nn^{*-1} \leq m + \sigma^2W^*$ for sufficiently small values of a such that n^{*-1} becomes smaller than unity. By Wiener's (1939) dominated ergodic theorem one concludes that $E\left(W^*\right) < \infty$. Now, the dominated convergence theorem and part (i) together imply part (ii), that is $E\left(N/n^*\right) \to 1$.

Next, along the lines of Theorem 2.4.1 in Mukhopadhyay and Solanky (1994), we can obtain $\frac{a'^{\frac{1}{2}}(N_\nu - a'\Psi_\nu)}{b'\Psi^{\frac{1}{2}}} \to N(0,1)$, with $\Psi_\nu = \frac{8b^2}{(ra)^2}$. Let $a' = \sigma^2$, $b' = \left(\frac{k+1}{2}\right)^{\frac{1}{2}}\sigma^2$, then part (iii) follows from the Theorem 2.4.1 of Mukhopadhyay and Solanky (1994).

Next, to prove the part (iv), note that following the steps from the Theorem 1, the $P(CD)$ expression based on the sample of size N can be simplified as

$$\begin{aligned}
\frac{P(CD+1)}{2} \geq\ & P\left\{d_1 < \bar{X}_j - \bar{X}_0 < d_2, r_1+1 \leq j \leq r_1+r_2;\right.\\
& \left. d_1 < \bar{X}_m - \bar{X}_0 < d_2, r_1+r_2+1 \leq m \leq r_1+r_2+r_3\right\}\\
=\ & P\left\{d_1 - \delta_2 < \left(\bar{X}_j - \mu_j\right) - \left(\bar{X}_0 - \mu_0\right) < d_2 - \delta_2, r_1+1 \leq j \leq r_1+r_2;\right.\\
& \left. d_1 - \delta_3 < \left(\bar{X}_m - \mu_m\right) - \left(\bar{X}_0 - \mu_0\right) < d_2 - \delta_3, r_1+r_2+1 \leq m \leq r_1+r_2+r_3\right\}\\
=\ & P\left\{\frac{d_1-\delta_2}{\sqrt{\sigma^2/N}} < \frac{\left(\bar{X}_j - \mu_j\right)}{\sqrt{\sigma^2/N}} - \frac{\left(\bar{X}_0 - \mu_0\right)}{\sqrt{\sigma^2/N}} < \frac{d_2-\delta_2}{\sqrt{\sigma^2/N}}, r_1+1 \leq j \leq r_1+r_2;\right.\\
& \left.\frac{ra}{\sqrt{\sigma^2/N}} < \frac{\left(\bar{X}_m - \mu_m\right)}{\sqrt{\sigma^2/N}} - \frac{\left(\bar{X}_0 - \mu_0\right)}{\sqrt{\sigma^2/N}} < \frac{d_2-\delta_3}{\sqrt{\sigma^2/N}}, r_1+r_2+1 \leq m \leq r_1+r_2+r_3\right\}\\
=\ & P(-b < Z_j - Z_0 < c, \quad r_1+1 \leq j \leq r_1+r_2, \quad -c < Z_m - Z_0 < b,\\
& r_1+r_2+1 \leq m \leq r_1+r_2+r_3)\\
=\ & P(-b < Z_j - Z_0 < c, \quad r_1+1 \leq j \leq r_1+r_2, \quad -c < Z_m - Z_0 < b,\\
& r_1+r_2+1 \leq m \leq r_1+r_2+r_3),
\end{aligned}$$

where,

$$\bar{X}_j \sim N\left(\mu_j, \frac{\sigma^2}{N}\right), \qquad \mu_j = \mu_0 + \delta_2,$$

$$\bar{X}_m \sim N\left(\mu_m, \frac{\sigma^2}{N}\right), \qquad \mu_m = \mu_0 + \delta_3.$$

$$\delta_2 - \delta_1 = \delta_4 - \delta_3 = ra, \qquad d_1 = \frac{\delta_1+\delta_2}{2}, \qquad d_2 = \frac{\delta_3+\delta_4}{2}.$$

$$b = \frac{ra}{\sqrt{\sigma^2/N}}, \qquad c = \frac{(2a-3ra)/2}{\sqrt{\sigma^2/N}}.$$

Next, as shown in Theorem 1, one can easily verify that $-b > -c$ and $b < c$, and using these we can further write

$$\begin{aligned}
\frac{P\left(CD+1\right)}{2} &\geq P(-b < Z_j - Z_0 < c, \quad r_1+1 \leq j \leq r_1+r_2, \quad -c < Z_m - Z_0 < b, \\
&\quad r_1+r_2+1 \leq m \leq r_1+r_2+r_3) \\
&> P(-b < Z_j - Z_0 < b, \quad r_1+1 \leq j \leq r_1+r_2, \quad -b < Z_m - Z_0 < b, \\
&\quad r_1+r_2+1 \leq m \leq r_1+r_2+r_3) \\
&= P(-b + Z_0 < Z_i < b + Z_0, \quad r_1+1 \leq i \leq r_1+r_2+r_3) \\
&= E[\int_{-\infty}^{+\infty} \left\{\Phi(b+z) - \Phi(-b+z)\right\}^{r_2+r_3} \phi(z) dz | Z_0 = z]. \qquad (2.3.15)
\end{aligned}$$

Also, from part (i), one gets $N^{\frac{1}{2}}\left(ra\left(2\sqrt{2}\sigma\right)^{-1}\right) \to b$ w.p. 1 as $a \to 0$, and hence (2.3.15) together with the dominated convergence theorem will lead to part (iv).

■

Next, for the purely sequential procedure (2.3.12) we will derive a second-order expansion to determine the amount of over-sampling the procedure does asymptotically. The amount of over-sampling β is defined below and also tabulated for the practitioners. We will also compare the validity of the asymptotic expression β for the small and moderate sample sizes.

Theorem 3 *For the purely sequential procedure* (2.3.12)*, we have as* $a \to 0$:

(i) $E\left(N\right) = n^* + \beta + o(1)$ for all $\mu \in \Omega\left(a\right)$ if $m \geq 2$ when $k \geq 2$*;*

(ii) $P[CD|\boldsymbol{\mu}^0(r_1, r_2, r_3, r_4), \ \sigma^2] > P^* + \frac{2}{n^*}\left\{\frac{\nu(k)-2}{k+1}g'(1) + \frac{1}{2(k+1)}g''(1)\right\} - 1 + o(\frac{1}{n^*})$ under the LFC if (a) $m_1 \geq 3$ when $k = 2, 3$, (b) $m_1 \geq 2$ when $k \geq 4$ *;*

where $n^* = \frac{2b^2\sigma^2}{(ra/2)^2}$, $P^* = 2g(1)$, $g(x)$ *is defined in* (2.3.19)*,* $g'(x)$ and $g''(x)$ are defined in (2.3.20), $\beta = (k+1)^{-1}\left\{\nu\left(k\right) - 2\right\}$ and $\nu\left(k\right)$ is defined in equation (2.3.18). *The values of the constant* $\beta = \beta(k)$ *are provided in Table* (2.3).

Proof. First note that using (2.3.14), we can rewrite $S_n^2 = (n-1)^{-1}\sum_{i=1}^{n-1} W_i'$, where W_1', W_2', $\cdots$ are i.i.d. $(k+1)^{-1}\sigma^2\chi_{k+1}^2$ random variables. Let's write $W_i = (k+1)\,\sigma^{-2}W_i'$, with $W_i's$ being i.i.d. χ_{k+1}^2. Using this the purely sequential procedure could be rewritten as

$$N = \inf\left\{n \geq m : (k+1)\,n\,(n-1)\,n^{*-1} \geq \sum\nolimits_{i=1}^{n-1} W_i\right\}. \tag{2.3.16}$$

Note that $N = Q+1$ where

$$Q = \inf\left\{n \geq m-1 : (k+1)\,n^2\left(n^{-1}+1\right)n^{*-1} \geq \sum\nolimits_{i=1}^{n} W_i\right\}. \tag{2.3.17}$$

The stopping variable Q is of the form of Mukhopadhyay and Solanky (1994)'s equation (2.4.7) with $\delta = 2$, $L_0 = 1$, $h^* = \frac{k+1}{n^*}$, $\theta = E(W_1) = k+1$, $r^2 = E(W_1^2) - \theta^2 = 2(k+1)$, $\beta^* = (\delta-1)^{-1} = 1$, $n_0^* = (\theta/h^*)^{\beta^*} = n^*$, $P = \beta^{*2}r^2\theta^{-2} = 2(k+1)^{-1}$, $b = (k+1)\big/2$, and

$$\nu = \nu(k) = \frac{1}{2}(k+3) - \sum\nolimits_{n=1}^{\infty} n^{-1}E\left[\max\left\{0, \chi_{n(k+1)}^2 - 2n(k+1)\right\}\right]. \tag{2.3.18}$$

Next, note that the constant η as defined in the equation (2.4.10) in Mukhopadhyay and Solanky (1994) simplifies to

$$\begin{aligned}
\eta &= \beta^*\theta^{-1}\nu - \beta^* L_0 - \frac{1}{2}\delta\beta^{*2}r^2\theta^{-2} \\
&= \nu(k+1)^{-1} - 1 - 2(k+1)^{-1} \\
&= (\nu-2)\big/(k+1) - 1.
\end{aligned}$$

Using the Theorem 2.4.8(v) of Mukhopadhyay and Solanky (1994) with $w = 1$ leads to

$$\begin{aligned}
E(N) &= E(Q) + 1 \\
&= 1 + n^* + \eta + \circ(1)
\end{aligned}$$

$$= \quad n^{*} + (\nu - 2) \big/ (k+1) + \circ (1) \, ,$$

if $m - 1 > 2\,(k+1)^{-1}$, that is, if $m > 1 + 2\,(k+1)^{-1}$. This is part (i).

For part (ii), we have the following from (2.3.15)

$$\frac{P(CD)+1}{2} \quad > \quad E[\int \left\{ \Phi(b+z) - \Phi(-b+z) \right\}^{r_2+r_3} \phi(z) dz | Z_0 = z]$$

Let $b = \sqrt{2}x$ and

$$\beta(x) = \int \Big\{ \Phi(\sqrt{2}x + z) - \Phi(-\sqrt{2}x + z) \Big\}^{r_2+r_3} \phi(z) dz$$

then

$$\beta'(x) = \int \sqrt{2}(r_2+r_3) \left\{ \Phi(b+z) - \Phi(-b+z) \right\}^{r_2+r_3-1} (\phi(\sqrt{2}x+z) + \phi(-\sqrt{2}x+z))\phi(z)dz$$

$$\begin{aligned} \beta''(x) \quad &= \quad \int 2(r_2+r_3)(r_2+r_3-1) \left\{ \Phi(b+z) - \Phi(-b+z) \right\}^{r_2+r_3-2} (\phi(\sqrt{2}x+z) + \phi(-\sqrt{2}x+z))^2 \\ &- \quad 2(r_2+r_3) \left\{ \Phi(b+z) - \Phi(-b+z) \right\}^{r_2+r_3-1} \Big((\sqrt{2}x+z)\phi(\sqrt{2}x+z) \\ &+ \quad (\sqrt{2}x-z)\phi(-\sqrt{2}x+z)) \Big) \phi(z)dz. \end{aligned}$$

Then, define

$$g(x) = \beta(bx^{\frac{1}{2}}), \quad x > 0. \tag{2.3.19}$$

It is easy to verify that

$$\begin{aligned} g'(x) &= \frac{1}{2} bx^{\frac{1}{2}} \beta'(bx^{\frac{1}{2}}) \\ g''(x) &= \frac{1}{4} bx^{-1} \beta''(bx^{\frac{1}{2}}) - \frac{1}{4} bx^{-\frac{3}{2}} \beta'(bx^{\frac{1}{2}}) \end{aligned} \tag{2.3.20}$$

and

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

$a_1, \ \ a_2 \ \ a_3$ being positive constants.

By using Theorem 3.2.1 of Mukhopadhyay and Solanky (1994), we have

$$\frac{(\inf P[CD|\boldsymbol{\mu}^0(r_1, r_2, r_3, r_4), \ \ \sigma^2] + 1)}{2} > E(g(N/n^*)).$$

Expanding $g(x)$ at $x = 1$ gives us

$$g(x) = g(1) + g'(1)(x-1) + g''(Z(x))(x-1)^2 \Big/ 2,$$

where Z is positive random variable such that

$$\min(1, N/n^*) \leq Z \leq \max(1, N/n^*).$$

Since $|g''(x)| \leq \sum_{i=1}^{3} \alpha_i / x^{\alpha_i}$, by Lemma 3.5.1 of Mukhopadhyay and Solanky (1994), for $m > \frac{5}{k+1} + 1$, one will obtain

$$\begin{aligned} E(g(N/n^*)) &= g(1) + g'(1)E(N/n^* - 1) + E(g''(Z(x))(N-n^*)^2/(2n^{*2})) \\ &= g(1) + \frac{1}{n^*}g'(1)E(N-n^*) + \frac{1}{k+1}\frac{1}{2n^*}g''(1) + o(\frac{1}{n^*}) \end{aligned}$$

Using the Theorem 3, part (i), we have

$$E(N) = n^* + (\nu(k) - 2)(k+1)^{-1} + o(1)$$

and

$$E(g(N/n^*)) = g(1) + \frac{1}{n^*}\left\{\frac{\nu(k)-2}{k+1}g'(1) + \frac{1}{2(k+1)}g''(1)\right\} + o(\frac{1}{n^*}).$$

That is,

$$\frac{(\inf P[CD|\boldsymbol{\mu}^0(r_1,r_2,r_3,r_4),\ \ \sigma^2]+1)}{2} > g(1)+\frac{1}{n^*}\left\{\frac{\nu(k)-2}{k+1}g'(1)+\frac{1}{2(k+1)}g''(1)\right\}+o(\frac{1}{n^*}).$$

Hence, we have

$$P[CD|\boldsymbol{\mu}^0(r_1,r_2,r_3,r_4),\ \ \sigma^2] > 2g(1)+\frac{2}{n^*}\left\{\frac{\nu(k)-2}{k+1}g'(1)+\frac{1}{2(k+1)}g''(1)\right\}-1+o(\frac{1}{n^*}).$$

This completes the proof of the theorem.

■

Table 2.3: The values of the constant β as defined in Theorem 3

(The value on top is k and the value underneath it is β)

2	3	4	5	6	7	8
0.0348	0.1716	0.2495	0.2991	0.3331	0.3577	0.3762
9	10	11	12	13	14	15
0.3905	0.4019	0.4111	0.4188	0.4252	0.4307	0.4354
16	17	18	19	20	21	22
0.4395	0.4431	0.4463	0.4492	0.4517	0.4540	0.4561
23	24	25	26	27	28	29
0.4580	0.4597	0.4613	0.4628	0.4641	0.4654	0.4666
30	31	32	33	34	35	36
0.4677	0.4687	0.4696	0.4705	0.4714	0.4722	0.4729
37	38	39	40	41	42	43
0.4737	0.4743	0.4750	0.4756	0.4762	0.4767	0.4773
44	45	46	47	48	49	50
0.4778	0.4783	0.4787	0.4792	0.4796	0.4800	0.4804
51	52	53	54	55	56	57
0.4808	0.4811	0.4815	0.4818	0.4821	0.4825	0.4828
58	59					
0.4831	0.4833					

In the next section, the purely sequential partition procedure (2.3.12) is simulated under several parametric conditions changing the values of the location of the populations, the values of the variance σ^2, the values of constant r, and the values of the δ_1 and δ_4. The goal is to confirm the

LFC and to verify the derived theoretical properties from Theorems 2 and 3 via simulations.

2.4 Monte Carlo Simulation Study of the Purely Sequential Procedure

The purely sequential procedure (2.3.12), starts with $m(\geq 2)$ observations from each of the k populations and the control population. The procedure takes one additional sample at a time from each all the k populations and the control population according to the following stopping rule (2.3.12). For all the simulations reported in this section, we took the value of $m = 5$, $k = 8$, and $P^* = 0.95$. The value of the design constant b was obtained from (2.1) for the given value of k and the target value of the probability of correct decision P^*. Recall that the design parameters δ_1 and δ_4 are provided by the experimenter based on the definition of the "Good" and "Bad" populations. Note that $a = \delta_4 - \delta_1$. And the experimenter tolerance for the misclassification of the populations in-between the "Good" and "Bad" populations determines the value of the design constant r. The smaller value of r, $0 < r < \frac{1}{2}$, the larger the value of n^* would be and smaller the size of the "Medium" classification set would be. After the value of n^* is computed, a sample of size at least n^* is collected from all the k populations and the control population.

In Table (2.4), we have chosen $\delta_4 - \delta_3 = \delta_3 - \delta_2 = \delta_2 - \delta_1 = c$, $c = \frac{2\sqrt{2}b\sigma}{\sqrt{n^*}}$, $\sigma = 9$, for several values of the optimal sample size n^*.

Table 2.4: Simulation Result under LFC for $(2, 2, 2, 2)$

n^*	$\bar{n}$	$std\left(\bar{n}\right)$	$\bar{P}$	$std\left(\bar{P}\right)$	c
25	25.335	0.0341	0.9718	0.0023	1.5250
50	50.350	0.0484	0.9728	0.0023	1.0784
100	100.343	0.0679	0.9746	0.0022	0.7625
200	200.364	0.0946	0.9764	0.0021	0.5392
300	300.359	0.1140	0.9770	0.0021	0.4402
400	400.593	0.1328	0.9734	0.0023	0.3813

In Table (2.4), we have summarized the performance of the purely sequential procedure under parametric configuration given by the LFC for $k = 8$ giving 2 populations on each of the four boundaries. Note that this configuration, denoted as (2,2,2,2), is the Least favorable Configuration under which there are an equal number of populations on all the four boundaries as shown in (2.2). Note that the average sample size $\bar{n}$ is fairly close to the unknown optimal sample size n^* for all the cases which we considered. Also, note that from Theorem 3, that the second-order expansion provides that asymptotically the difference between the value of $\bar{n}$ and n^* should be β. From Table (2.3) one obtains that asymptotically this difference should be 0.3762. That is, the purely sequential procedure (2.3.12) over-samples by a third of a sample asymptotically. The simulated values in Table (2.4) confirm this asymptotic difference between the $\bar{n}$ and n^*. Also, note that the average value of the probability of correct decision $\bar{P}$ matches the target value of 0.95 in all the cases considered. The findings in Table (2.4), confirm the theoretical results derived in Theorem 2 and 3 for the purely sequential procedure.

Table 2.5: simulation results for different number of groups at each point with $(\delta_1, \delta_2, \delta_3, \delta_4) = (5, 15, 25, 35)$

(k_1, k_2, k_3, k_4)	n^*	$\bar{n}$	$std(\bar{n})$	$\bar{P}$	$std(\bar{P})$	b
$(2, 2, 2, 2)$	47	47.5224	0.0471	0.9732	0.0023	2.6959
$(1, 3, 3, 1)$	47	47.4224	0.0468	0.9750	0.0022	2.6959
$(3, 1, 1, 3)$	47	47.4960	0.0473	0.9744	0.0022	2.6959
$(4, 0, 0, 4)$	47	47.4316	0.0463	0.9772	0.0021	2.6959
$(1, 6, 1)$	47	47.4408	0.0479	0.9928	0.0012	2.6959
$(2, 4, 2)$	47	47.4634	0.0471	0.9856	0.0017	2.6959

In Table (2.5) the purely sequential procedure (2.3.12) for the generalized partition procedure is simulated under several parametric configurations to verify the LFC and to confirm that the probability requirement (2.2.5) holds for all the parametric configurations. Note that for the parametric configuration (2,2,2,2), the $P(CD)$ is least, among all the parametric configurations satisfying (2.1.1). In this table we fixed the value of $\sigma = 9$. Let, a_1 be the number of populations with mean $\mu_0 + \delta_1$, a_2 is the number with mean $\mu_0 + \delta_2$, a_3 is the number with mean $\mu_0 + \delta_3$, and a_4 is number of populations with mean $\mu_0 + \delta_4$. We express this parametric configuration as (a_1, a_2, a_3, a_4). Let

e denote the mid point between $\mu_0 + \delta_2$ and $\mu_0 + \delta_3$ and the three-tuple (a_1, e, a_4) denotes the parametric configuration in which the populations are located on the three locations only. Note that in the last two parametric configurations in the Table (2.5), all the populations are not located on the boundaries as such. For example, in (2,4,2) there are 2 populations each on the left and the right most boundary and there are 4 populations in the midpoint of the two middle boundaries. Using this notation, we have reported the performance of several parametric configurations in Table (2.5) for the values of $(\delta_1, \delta_2, \delta_3, \delta_4) = (5, 15, 25, 35)$. Note that n^* is fixed as 47 in this table and we only vary the location of the populations. The $\bar{P}$ values show that the configuration (2,2,2,2) is clearly the LFC and for all the parametric configurations the $\bar{P}$ value is above the target value of $P^* = 0.95$.

Table 2.6: Simulation Result for $(2, 2, 2, 2)$ with $(\delta_1, \delta_2, d) = (10, 15, 15)$

σ	n^*	$\bar{n}$	$std(\bar{n})$	$\bar{P}$	$std(\bar{P})$
1	3(2.33)	5.000	0.0000	0.9998	0.0002
2	10(9.30)	9.633	0.0228	0.9740	0.0023
3	21(20.93)	21.263	0.0321	0.9742	0.0022
4	38(37.21)	37.600	0.0415	0.9742	0.0022
5	58(58.14)	58.526	0.0520	0.9752	0.0022
6	84(83.73)	84.115	0.0615	0.9768	0.0021
7	114(113.96)	114.368	0.0727	0.9780	0.0021
8	149(148.85)	148.256	0.0804	0.9754	0.0013
9	189(188.38)	188.795	0.0925	0.9758	0.0022
10	233(232.57)	233.011	0.1020	0.9750	0.0022
12	335(334.90)	335.285	0.1236	0.9764	0.0021
14	456(455.84)	455.924	0.1414	0.9752	0.0022
16	596(595.38)	595.903	0.1615	0.9764	0.0021
20	931(930.29)	930.706	0.2028	0.9766	0.0021

Next, in Table (2.6), we have further explored the LFC configuration (2,2,2,2). Let us define d as the distance from point $\mu_0 + \delta_1$ to point $\mu_0 + \delta_3$. In this Table we have fixed the values of $(\delta_1, \delta_2, d) = (10, 15, 15)$, and we have varied the values of σ. Recall that theoretically from the Theorem 3, the purely sequential procedure (2.3.12) over-samples by a third of a sample asymptotically. The simulated values in Table (2.6) confirm this asymptotic difference between the $\bar{n}$

and n^* in every instance. Also, note that the average value of the probability of correct decision $\bar{P}$ matches the target value of 0.95 in all the cases considered. The findings in Table (2.6) also confirm the theoretical results derived in Theorem 2 and 3 for the purely sequential procedure.

Table 2.7: Simulation Result for $(2, 2, 2, 2)$ with $(\delta, \delta + r * a, \delta + (1 - r) * a, \delta + a)$

a	r	n^*	$\bar{n}$	$std(\bar{n})$	$\bar{P}$	$std(\bar{P})$
10.7836	0.100	50	50.380	0.0486	0.9770	0.0021
	0.075	89	89.259	0.0630	0.9758	0.0022
	0.050	200	200.458	0.0937	0.9748	0.0022
	0.025	800	800.438	0.1877	0.9722	0.0023
7.6252	0.200	25	25.346	0.0344	0.9770	0.0021
	0.100	100	100.370	0.0685	0.9744	0.0022
	0.075	178	178.154	0.0895	0.975	0.0022
	0.050	400	400.486	0.1342	0.9750	0.0022
5.3918	0.200	50	50.344	0.0484	0.9750	0.0022
	0.100	200	200.345	0.0941	0.9764	0.0021
	0.075	356	355.699	0.1259	0.9758	0.0022
	0.050	800	800.613	0.1932	0.9746	0.0022
3.8126	0.400	25	25.342	0.0343	0.9710	0.0024
	0.300	45	44.857	0.0456	0.9742	0.0022
	0.200	100	100.408	0.0667	0.9732	0.0023
	0.100	400	400.209	0.1362	0.9760	0.0022
	0.075	711	711.278	0.1809	0.9762	0.0022

In Table (2.7) we consider another parametric configuration under which the leftmost and the rightmost have fixed locations but we vary the location of the middle two positions. Say, the length of indifference zone $\delta_2 - \delta_1 = \delta_4 - \delta_3 = ra$. We took $\mu_0 = 0$, $\sigma = 1$, and let $\delta = 10$ to create the parametric configuration $(\mu_0 + \delta_1, \mu_0 + \delta_2, \mu_0 + \delta 3, \mu_0 + \delta_4) = (\delta, \delta + ra, \delta + (1 - r)a, \delta + a)$. The simulated values in Table (2.7) confirm this asymptotic difference between the $\bar{n}$ and n^* in every instance. Also, note that the average value of the probability of correct decision $\bar{P}$ matches the target value of 0.95 in all the cases considered. The findings in Table (2.7) also confirm the theoretical results derived in the Theorem 2 and 3 for the purely sequential procedure.

To conclude, the purely sequential procedure (2.3.12) for the generalized partition procedure is able to partition the populations as “Good populations”, “Bad populations”, also as a separate identifiable group “Medium or Indifferent populations” with a high degree of precision matching the pre-specified target probability. The simulations confirm the nature of the LFC and the theoretical properties of the purely sequential procedure (2.3.12).

Chapter 3

Two-Stage Procedure

3.1 Two Stage Procedure

In this Chapter we will propose a two-stage procedure for the generalized partition problem introduced in the chapter 2 for the unknown σ^2 case. The two-stage procedures are operationally more convenient to implement than the purely sequential procedures. This is so because unlike for the purely sequential procedure, in which the experimenter has to decide whether or not to continue sampling after each sample, in the two-stage procedure the sample size is determined only once. Meaning, the experimenter would select a small pilot sample and then based on that pilot sample it is determined how many additional samples need to be collected. This feature of the two-stage procedure makes it more user friendly and operationally convenient. For more literature on the two-stage procedures, the reader is recommended to Solanky (2006).

Next, we describe the a two-stage procedure to obtain the generalized partition problem presented in the chapter 2.

Stage I. Let $m\ (\geq 2)$ denote the common starting sample size from k treatments and the control group. The procedure begins by taking a sample X_{ij} from π_i; $i = 0, 1, \cdots, k$; $j = 1, \cdots, m$. Let

$$\bar{X}_i = \sum\nolimits_{j=1}^{m} X_{ij}/m,$$

$i = 0, 1, \cdots, k$, denote the sample means based on the stage I sampling. Also, let U_m be the usual pooled estimator of σ^2, where

$$U_m = \sum\nolimits_{i=0}^{k} S_{im}^2 \big/ (k+1), \qquad S_{im}^2 = \sum\nolimits_{j=1}^{m} \left(X_{ij} - \bar{X}_i\right)^2 \big/ (m-1).$$

Note that the pooled estimator U_m has $f = (k+1)(m-1)$ degree of freedom and fU_m/σ^2 has the χ_f^2 distribution.

Stage II. In the second stage, $N - m$ additional samples are taken from π_0 and π_i, $i = 1, \cdots, k$, where N is defined as

$$N = \max\left\{m, \left[2\tau^2 U_m \left(ra/2\right)^{-2}\right]\right\}. \tag{3.1.1}$$

The constant $\tau = \tau(k, m, P^*)$ is a positive constant defined in (3.2.8), and $[x]$ denotes the smallest integer greater than or equal to x. Note that, if $N = m$, we do not take any samples from any population in the stage II. However, if $N > m$ then we sample the difference from each π and the control population in the second stage.

3.2 Asymptotic Properties of The Two-stage Procedure

In this section, we will derive some asymptotic theoretical properties of the proposed two-stage procedure (3.1.1).

Theorem 4 *For the two-stage procedure* (3.1.1)*, with τ defined in* (3.2.8)*, we have as $a \to 0$:*

(i) $P(CD) \geq P^*$ for all $\mu \in \Omega(a)$*;*

(ii) $2\tau^2\sigma^2\left(ra/2\right)^{-2} \leq E(N) \leq m + 2\tau^2\sigma^2\left(ra/2\right)^{-2}$*;*

(iii) $E\left(N/n^*\right) \to \tau^2/b^2\,(>1)$ *as $a \to 0$;*

(iv) $\liminf P(CD) \geq P^*$*,* under LFC *as $a \to 0$;*

where $n^* = 2\tau^2\sigma^2 \Big/ \left(ra/2\right)^2$ *and τ comes from the Tables* (3.1) *and* (3.2).

Proof. We start by noting that the basic inequality based on the definition of the sample size N from (3.1.1) is

$$2\tau^2 U_m^2 \big(ra/2\big)^{-2} \le N \le m + 2\tau^2 U_m^2 \big(ra/2\big)^{-2} \qquad (3.2.2)$$

and now taking expectations throughout leads to part (ii), since $E(U_m) = \sigma^2$. Next, dividing throughout the expressions in part (ii) by n^* and the taking limit as $a \to 0$ leads to part (iii).

For this procedure, the probability of *correct decision* can be expressed as

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

Next, under the LFC the above expression simplifies to:

$$\begin{aligned}
&P\Big[CD|\boldsymbol{\mu}^0(r_1, r_2, r_3, r_4),\ \ \sigma^2\Big] = P\Big[\,\big((\bar{X}_i - \mu_i) - (\bar{X}_0 - \mu_0)\big) < (d_1 - \delta_1)\Big/\sqrt{\frac{2\sigma^2}{n}},\\
&(d_1 - \delta_2)\Big/\sqrt{\frac{2\sigma^2}{n}} < \big((\bar{X}_j - \mu_j) - (\bar{X}_0 - \mu_0)\big)\Big/\sqrt{\frac{2\sigma^2}{n}} < (d_2 - \delta_2)\Big/\sqrt{\frac{2\sigma^2}{n}},\\
&(d_1 - \delta_3)\Big/\sqrt{\frac{2\sigma^2}{n}} < \big((\bar{X}_m - \mu_m) - (\bar{X}_0 - \mu_0)\big)\Big/\sqrt{\frac{2\sigma^2}{n}} < (d_2 - \delta_3)\Big/\sqrt{\frac{2\sigma^2}{n}},\\
&\big((\bar{X}_l - \mu_l) - (\bar{X}_0 - \mu_0)\big)\Big/\sqrt{\frac{2\sigma^2}{n}} > (d_2 - \delta_4)\Big/\sqrt{\frac{2\sigma^2}{n}},\\
&1 \le i \le r_1, r_1 + 1 \le j \le r_1 + r_2, , r_1 + r_2 + 1 \le m \le r_1 + r_2 + r_3, r_1 + r_2 + r_3 + 1 \le l \le k\Big]\\
=\ \ &P\Big[\big((\bar{X}_i - \mu_i) - (\bar{X}_0 - \mu_0)\big)\Big/2\sqrt{\frac{2\sigma^2}{n}} < ra\Big/2\sqrt{\frac{2\sigma^2}{n}},\\
&-ra\Big/2\sqrt{\frac{2\sigma^2}{n}} < \big((\bar{X}_j - \mu_j) - (\bar{X}_0 - \mu_0)\big)\Big/2\sqrt{\frac{2\sigma^2}{n}} < (2a - 3ra)\Big/2\sqrt{\frac{2\sigma^2}{n}},\\
&-(2a - 3ra)\Big/2\sqrt{\frac{2\sigma^2}{n}} < \big((\bar{X}_m - \mu_m) - (\bar{X}_0 - \mu_0)\big)\Big/2\sqrt{\frac{2\sigma^2}{n}} < ra\Big/2\sqrt{\frac{2\sigma^2}{n}},\\
&\big((\bar{X}_l - \mu_l) - (\bar{X}_0 - \mu_0)\big)\Big/2\sqrt{\frac{2\sigma^2}{n}} > -ra\Big/2\sqrt{\frac{2\sigma^2}{n}},
\end{aligned}$$

$$
\begin{aligned}
& 1 \le i \le r_1, r_1+1 \le j \le r_1+r_2, r_1+r_2+1 \le m \le r_1+r_2+r_3, r_1+r_2+r_3+1 \le l \le k\big] \\
= \; & P\big[Y_i < ra\Big/2\sqrt{\frac{2\sigma^2}{n}}, 1 \le i \le r_1, r_1+r_2+r_3+1 \le i \le k, \\
& -(2a-3ra)\Big/2\sqrt{\frac{2\sigma^2}{n}} < Y_j < ra\Big/2\sqrt{\frac{2\sigma^2}{n}}, r_1+1 \le j \le r_1+r_2+r_3\big], \qquad (3.2.3)
\end{aligned}
$$

where,

$$Y_i = ((\bar{X}_i - \mu_i) - (\bar{X}_0 - \mu_0))\Big/\sqrt{\frac{2\sigma^2}{n}},$$

for $0 < i \le r_1$, and $r_1 + r_2 < i \le r_1 + r_2 + r_3$,

$$Y_i = -((\bar{X}_i - \mu_i) - (\bar{X}_0 - \mu_0))\Big/\sqrt{\frac{2\sigma^2}{n}},$$

for $r_1 < i \le r_1 + r_2$, and $r_1 + r_2 + r_3 < i \le k$. Note that under the parameter configuration $\boldsymbol{\mu}^0(r_1, r_2, r_3, r_4)$, Y_i has the standard normal distribution, $i = 1, \cdots, k$. Let us define the $(k \times k)$ covariance matrix $\Sigma_Y = (\sigma_{ij})$ as

$$
\begin{aligned}
\sigma_{ij} &= 1 \text{ for } i = j \\
&= 1/2 \text{ for } i \ne j, \text{ and } i,\, j \in [1, r_1] \cup [r_1+r_2+1, r_1+r_2+r_3]\,, \\
&\quad \text{or } i,\, j \in [r_1+1, r_1+r_2] \cup [r_1+r_2+r_3+1, k] \\
&= -1/2 \text{ for } i \in [1, r_1] \cup [r_1+r_2+1, r_1+r_2+r_3]\,,\; j \in [r_1+1, r_1+r_2] \cup [r_1+r_2+r_3+1, k]\,.
\end{aligned}
$$

Since $0 < r < \frac{1}{2}$, we have

$$-a\Big/\sqrt{\frac{2\sigma^2}{n}} < (3r-2)\,a/2\sqrt{\frac{2\sigma^2}{n}} < -a\Big/4\sqrt{\frac{2\sigma^2}{n}},$$

and $0 < ra\Big/2\sqrt{\frac{2\sigma^2}{n}} < a\Big/4\sqrt{\frac{2\sigma^2}{n}}$. Combining these two, one can obtain

$$-ra\Big/2\sqrt{\frac{2\sigma^2}{n}} > -a/4\sqrt{\frac{2\sigma^2}{n}} > (3r-2)\,a\Big/2\sqrt{\frac{2\sigma^2}{n}}.$$

Hence, we can claim that

$$\begin{aligned} & P\Big[CD|\boldsymbol{\mu}^0(r_1, r_2, r_3, r_4), \;\; \sigma^2\Big] \\ \geq \;\; & E\Big[P\left(Y_i \leq \frac{ra/2}{\sqrt{2}\sigma/\sqrt{n}}, i = 1, \cdots, r_1, r_1 + r_2 + r_3 + 1, \cdots, k\right) \\ & + P\left(-\frac{ra/2}{\sqrt{2}\sigma/\sqrt{n}} \leq Y_j \leq \frac{ra/2}{\sqrt{2}\sigma/\sqrt{n}}, j = r_1 + 1, \cdots, r_1 + r_2 + r_3 + 1\right) - 1\Big] \end{aligned} \tag{3.2.4}$$

under LFC. Since $N \geq \tau^2 U_m \big/ (ra/2)^2$ w.p.1, which follows from the left hand side of $(3.2.2)$, then

$$\begin{aligned} \inf_{\mu \in \Omega(a)} P\,(CS) \;\; = \;\; & E\Big[P\left(Y_i \leq \frac{\tau U_{m_1}^{\frac{1}{2}}}{\sqrt{2}\sigma}, i = 1, \cdots, r_1, r_1 + r_2 + r_3 + 1, \cdots, k\right) \\ & + P\left(-\frac{\tau U_{m_1}^{\frac{1}{2}}}{\sqrt{2}\sigma} \leq Y_j \leq \frac{\tau U_{m_1}^{\frac{1}{2}}}{\sqrt{2}\sigma}, j = r_1 + 1, \cdots, r_1 + r_2 + r_3 + 1\right) \\ & - 1|U_{m_1}\Big]. \end{aligned} \tag{3.2.5}$$

Let

$$T_i = \frac{Y_i}{U_m^{\frac{1}{2}} \sigma^{-1}},$$

$i = 1, \cdots, k$, then $(T_1, T_2, \cdots, T_k)$ is distributed as k dimensional multivariate t with equicorrelation $= \frac{1}{2}$, and the degree of freedom $= (k + 1)\,(m - 1)$, so the $(3.2.5)$ can be written as

$$\begin{aligned} \inf_{\mu \in \Omega(a)} P\,(CS) \;\; \geq \;\; & P\left(T_i \leq \frac{\tau}{\sqrt{2}}, i = 1, \cdots, r_1, r_1 + r_2 + r_3 + 1, \cdots, k\right) \\ + \;\; & P\left(-\frac{\tau}{\sqrt{2}} \leq T_j \leq \frac{\tau}{\sqrt{2}}, j = r_1 + 1, \cdots, r_1 + r_2 + r_3 + 1\right) - 1. \end{aligned} \tag{3.2.6}$$

In other words, we determine $\tau = \tau\,(m, k, P^*)$ in such a way that

$$P\left(T_i \leq \frac{\tau}{\sqrt{2}}\right) + P\left(-\frac{\tau}{\sqrt{2}} \leq T_j \leq \frac{\tau}{\sqrt{2}}\right) - 1 \geq 2P\left(-\frac{\tau}{\sqrt{2}} \leq T_j \leq \frac{\tau}{\sqrt{2}}\right) - 1 = P^*, \tag{3.2.7}$$

where $i = 1, \cdots, r_1, r_1 + r_2 + r_3 + 1, \cdots, k$, $j = r_1 + 1, \cdots, r_1 + r_2 + r_3 + 1$.

This completes the proof of part (i). Actually, (3.2.7) is equivalent to determine $\tau = \tau\left(m, k, P^*\right)$ by

$$P\left(-\frac{\tau}{\sqrt{2}} \leq T_j \leq \frac{\tau}{\sqrt{2}}\right) = \frac{P^* + 1}{2},$$

which can be simplified as:

$$\int_{-\frac{\tau}{\sqrt{2}}}^{\frac{\tau}{\sqrt{2}}} \cdots \int_{-\frac{\tau}{\sqrt{2}}}^{\frac{\tau}{\sqrt{2}}} \frac{\Gamma\left(\frac{1}{2}\left(k' + \left(k' + 1\right)\left(m - 1\right)\right)\right)\left(1 + \frac{y'\Sigma_{k'}^{-1}y}{\left(k'+1\right)\left(m-1\right)}\right)^{-\frac{1}{2}\left(k' + \left(k'+1\right)\left(m-1\right)\right)}}{\left(\pi\left(k' + 1\right)\left(m - 1\right)\right)^{\frac{1}{2}k'}\Gamma\left(\frac{1}{2}\left(k' + 1\right)\left(m - 1\right)\right)\left|\Sigma_{k'}\right|^{\frac{1}{2}}} \prod_{i=1}^{k'} dy_i \quad (3.2.8)$$

for $j = r_1 + 1, \cdots, r_1 + r_2 + r_3 + 1$. The Σ_k' is the $k' \times k'$ matrix defined by

$$\Sigma_k' = \begin{pmatrix} \begin{pmatrix} 1 & \cdots & \frac{1}{2} \\ \vdots & \ddots & \vdots \\ \frac{1}{2} & \cdots & 1 \end{pmatrix}_{r_2 \times r_2} & \begin{pmatrix} -\frac{1}{2} & \cdots & -\frac{1}{2} \\ \vdots & \ddots & \vdots \\ -\frac{1}{2} & \cdots & -\frac{1}{2} \end{pmatrix}_{r_2 \times r_3} \\ \begin{pmatrix} -\frac{1}{2} & \cdots & -\frac{1}{2} \\ \vdots & \ddots & \vdots \\ -\frac{1}{2} & \cdots & -\frac{1}{2} \end{pmatrix}_{r_3 \times r_2} & \begin{pmatrix} 1 & \cdots & \frac{1}{2} \\ \vdots & \ddots & \vdots \\ \frac{1}{2} & \cdots & 1 \end{pmatrix}_{r_3 \times r_3} \end{pmatrix}_{k' \times k'}.$$

Here, $k' = \left[\frac{k}{2}\right]$, $r_2 = \left[\frac{k'}{2}\right]$, and $r_3 = k' - r_2$, where where $[x]$ equals $\frac{x}{2}$ if x is even and $\frac{x+1}{2}$ if x is odd.

For part (iv), since $N^{\frac{1}{2}}\left(ra/2\right) \rightarrow \tau U_m^{\frac{1}{2}}$, w.p.1 as $a \rightarrow 0$, from (3.2.4) and the dominated convergency theorem, one obtains

$$\lim_{\mu \in \Omega(a)} \inf P\left(CS\right) \geq E\left[P\left(Y_i \leq \frac{\tau U_m^{\frac{1}{2}}}{\sqrt{2}\sigma}\right) + P\left(-\frac{\tau U_m^{\frac{1}{2}}}{\sqrt{2}\sigma} \leq Y_j \leq \frac{\tau U^{\frac{1}{2}}m}{\sqrt{2}\sigma}\right) - 1\right] \qquad (3.2.9)$$

$i = 1, \cdots, r_1, r_1 + r_2 + r_3 + 1, \cdots, k$, $j = r_1 + 1, \cdots, r_1 + r_2 + r_3 + 1$, for all $\tilde{\mu} \in \Omega(a)$. Hence, part (iv) follows from the (3.2.5)-(3.2.7).

To get the design constant $\tau = \tau(m, k, P^*)$ as defined in (3.2.8), we have next provided the values of $h_v = \tau/\sqrt{2}$ for $P^* = 0.5, 0.75, 0.80, 0.90, 0.95, 0.99$ when starting sample size $m_1 = 5, 10$ in the Table (3.1) and (3.2). That is $h_v = h_v(m, k, P^*)$.

Table 3.1: Values of $h_v = h_v\,(m, k, P^*)$ as define in $(3.2.8)$ with $m = 5$

k	P					
	0.50	0.75	0.80	0.90	0.95	0.99
1	1.240299	1.713294	1.859482	2.305932	2.751325	3.828689
2	1.240299	1.713294	1.859482	2.305932	2.751325	3.828689
3	1.404046	1.975402	2.106914	2.502561	2.885677	3.774655
4	1.404046	1.975402	2.106914	2.502561	2.885677	3.774655
5	1.693772	2.102810	2.226410	2.591934	2.942414	3.726125
6	1.693772	2.102810	2.226410	2.591934	2.942414	3.726125
7	1.792774	2.186051	2.303701	2.651128	2.979896	3.713427
8	1.792774	2.186051	2.303701	2.651128	2.979896	3.713427
9	1.865686	2.247297	2.361476	2.696991	3.014658	3.749320
10	1.865686	2.247297	2.361476	2.696991	3.014658	3.749320
11	1.923676	2.295958	2.153059	2.732687	3.040290	3.730431
12	1.923676	2.295958	2.153059	2.732687	3.040290	3.730431
13	1.970673	2.336380	2.445635	2.767066	3.064614	3.740755
14	1.970673	2.336380	2.445635	2.767066	3.064614	3.740755
15	2.011149	2.370211	2.482101	2.799736	3.092243	3.745842
16	2.011149	2.370211	2.482101	2.799736	3.092243	3.745842
17	2.045481	2.400764	2.506694	2.810381	3.094848	3.747565
18	2.045481	2.400764	2.506694	2.810381	3.094848	3.747565
19	2.075348	2.429809	2.530986	2.831065	3.095379	3.7490286
20	2.075348	2.429809	2.530986	2.831065	3.095379	3.7490286

3.3 Monte Carlo Simulation Study of the Two-Stage Procedure

In the Tables (3.3) and (3.4), the two-stage procedure (3.1.1) for the generalized partition procedure is simulated under several configurations with starting sample size $m = 5, 10$ to verify the LFC and to confirm that the probability requirement (3.2.9) holds for all parametric configurations. Note that the parametric configuration $(2, 2, 2, 2)$ is the LFC, among all the parametric configuration satisfying (2.1.1). In these Tables, we have fixed the value of $\sigma = 9$, $(\delta_1, \delta_2, \delta_3, \delta_4) = (5, 15, 25, 35)$. Note that n^* is fixed as 58 for $m = 5$ and 52 for $m = 10$, and we only vary the location of the

Table 3.2: Values of $h_v = h_v\left(m, k, P^*\right)$ as define in $(3.2.8)$ with $m = 10$

k	P					
	0.50	0.75	0.80	0.90	0.95	0.99
1	1.188698	1.609074	1.734192	2.100932	2.445194	3.196301
2	1.188698	1.609074	1.734192	2.100932	2.445194	3.196301
3	1.491817	1.881897	1.997228	2.333527	2.645600	3.320333
4	1.491817	1.881897	1.997228	2.333527	2.645600	3.320333
5	1.649170	2.022076	2.131634	2.452495	2.746304	3.383865
6	1.649170	2.022076	2.131634	2.452495	2.746304	3.383865
7	1.752886	2.115749	2.222135	2.530575	2.814163	3.414983
8	1.752886	2.115749	2.222135	2.530575	2.814163	3.414983
9	1.829937	2.185211	2.288666	2.593141	2.863535	3.438544
10	1.829937	2.185211	2.288666	2.593141	2.863535	3.438544
11	1.890257	2.239930	2.342548	2.406209	2.906977	3.461548
12	1.890257	2.239930	2.342548	2.406209	2.906977	3.461548
13	1.940306	2.285121	2.385681	2.678166	2.942322	3.437621
14	1.940306	2.285121	2.385681	2.678166	2.942322	3.437621
15	1.981928	2.3261077	2.429161	2.701895	2.967613	3.461993
16	1.981928	2.3261077	2.429161	2.701895	2.967613	3.461993
17	2.019229	2.356830	2.455798	2.739006	3.002371	3.703326
18	2.019229	2.356830	2.455798	2.739006	3.002371	3.703326
19	2.051009	2.387428	2.483978	2.764590	3.017216	3.705109
20	2.051009	2.387428	2.483978	2.764590	3.017216	3.705109

populations to confirm the LFC. The $\bar{P}$ values shows that the configuration $(2,2,2,2)$ is clearly the LFC and for all parametric configurations considered the $\bar{P}$ value is above the target value of $P^* = 0.95$. Also, the $\bar{P}$ under $m = 5$ is over the value of $\bar{P}$ under $m = 10$.

Table 3.3: Simulation results for different number of groups at each point with $(\delta_1, \delta_2, \delta_3, \delta_4) = (5, 15, 25, 35)$ with $m = 5$

(k_1, k_2, k_3, k_4)	n^*	$\bar{n}$	$std\,(\bar{n})$	$\bar{P}$	$std\left(\bar{P}\right)$	b
$(2,2,2,2)$	58(57.54)	58.17	0.1912	0.9806	0.0020	2.979896
$(1,3,3,1)$	58(57.54)	58.03	0.1912	0.9834	0.0018	2.979896
$(3,1,1,3)$	58(57.54)	58.27	0.1914	0.9828	0.0018	2.979896
$(1,6,1)$	58(57.54)	58.28	0.1938	0.9952	0.0010	2.979896
$(2,4,2)$	58(57.54)	57.67	0.1887	0.9894	0.0014	2.979896
$(4,0,0,4)$	58(57.54)	58.30	0.1955	0.9826	0.0018	2.979896

Table 3.4: Simulation results for different number of groups at each point with $(\delta_1, \delta_2, \delta_3, \delta_4) = (5, 15, 25, 35)$ with $m = 10$

(k_1, k_2, k_3, k_4)	n^*	$\bar{n}$	$std\,(\bar{n})$	$\bar{P}$	$std\left(\bar{P}\right)$	b
$(2,2,2,2)$	52(51.32)	51.88	0.1142	0.9772	0.0021	2.814163
$(1,3,3,1)$	52(51.32)	51.79	0.1130	0.9812	0.0019	2.814163
$(3,1,1,3)$	52(51.32)	51.86	0.1118	0.9816	0.0019	2.814163
$(1,6,1)$	52(51.32)	51.93	0.1161	0.9934	0.0011	2.814163
$(2,4,2)$	52(51.32)	52.01	0.1146	0.9860	0.0017	2.814163
$(4,0,0,4)$	52(51.32)	51.86	0.1126	0.9808	0.0019	2.814163

Next, in the Tables (3.5) - (3.10), we have further explored the configurations $(2,2,2,2)$, $(3,1,1,3)$, $(1,3,3,1)$, $(1,6,1)$, $(2,4,2)$, $(4,0,0,4)$ with starting sample size $m = 5$, and in the Tables (3.11) - (3.16) for the starting sample size $m = 10$. Define d as the distance from point $\mu_0 + \delta_1$ to the point $\mu_0 + \delta_3$. In this tables, we have fixed the values of $(\delta_1, \delta_2, d) = (10, 15, 15)$, and we have varied the values of σ. Note that the parametric configuration $(2,2,2,2)$ is least favorable again, among all the parametric configurations considered satisfying (2.1.1). Also the average value of probability of correct decision $\bar{P}$ is above the target value of $P^* = 0.95$.

In the last two Tables (3.17) and (3.18), the LFC configuration $(2,2,2,2)$ is further explored under $(\delta_1, \delta_2, d) = (10, 12, 18)$ and $(\delta_1, \delta_2, d) = (10, 18, 12)$ by varying the σ. The results of the

Table 3.5: Simulation Result for $(2, 2, 2, 2)$ with $(\delta_1, \delta_2, d) = (10, 15, 15)$ with $m = 5$

std	n^*	$\bar{n}$	$std\,(\bar{n})$	$\bar{P}$	$std\left(\bar{P}\right)$
1	3(2.84)	5.00	0.0008	0.9996	0.0003
2	12(11.37)	11.85	0.0379	0.9840	0.0018
3	26(25.57)	26.13	0.0849	0.9814	0.0019
4	46(45.46)	46.14	0.1502	0.9834	0.0018
5	71(70.03)	71.72	0.2413	0.9800	0.0020
6	103(102.30)	102.29	0.3397	0.9806	0.0020
7	140(139.23)	140.16	0.4677	0.9816	0.0019
8	182(181.85)	182.63	0.6017	0.9802	0.0020
9	230(230.16)	230.45	0.7649	0.9762	0.0022
10	284(284.15)	284.36	0.9524	0.9802	0.0020
12	409(409.18)	410.79	1.3531	0.9814	0.0019
14	557(556.94)	557.95	1.8787	0.9804	0.0020
16	728(727.43)	726.11	2.4407	0.9820	0.0019
20	1137(1136.61)	1138.32	3.7712	0.9808	0.0019

Table 3.6: Simulation Result for $(3, 1, 1, 3)$ with $(\delta_1, \delta_2, d) = (10, 15, 15)$ with $m = 5$

std	n^*	$\bar{n}$	$std\,(\bar{n})$	$\bar{P}$	$std\left(\bar{P}\right)$
1	3(2.84)	5.00	0.0009	1.0000	0.0000
2	12(11.37)	11.86	0.0384	0.9850	0.0017
3	26(25.57)	26.18	0.0859	0.9848	0.0017
4	46(45.46)	46.25	0.1559	0.9864	0.0016
5	71(70.03)	71.47	0.2371	0.9818	0.0019
6	103(102.30)	103.22	0.3453	0.9830	0.0018
7	140(139.23)	139.47	0.4609	0.9822	0.0019
8	182(181.85)	182.94	0.6171	0.9846	0.0017
9	230(230.16)	229.98	0.7545	0.9824	0.0019
10	284(284.15)	283.40	0.9435	0.9848	0.0019
12	409(409.18)	408.03	1.3578	0.9830	0.0018
14	557(556.94)	561.21	1.8940	0.9834	0.0018
16	728(727.43)	556.17	1.8362	0.9822	0.0019
20	1137(1136.61)	1139.60	3.8720	0.9818	0.0019

Table 3.7: Simulation Result for $(1, 3, 3, 1)$ with $(\delta_1, \delta_2, d) = (10, 15, 15)$ with $m = 5$

std	n^*	$\bar{n}$	$std\,(\bar{n})$	$\bar{P}$	$std\left(\bar{P}\right)$
1	3(2.84)	5.00	0.0008	0.9998	0.0002
2	12(11.37)	11.90	0.0383	0.9852	0.0017
3	26(25.57)	26.02	0.0841	0.9842	0.0018
4	46(45.46)	46.03	0.1516	0.9848	0.0017
5	71(70.03)	71.65	0.2397	0.9834	0.0018
6	103(102.30)	103.07	0.3368	0.9840	0.0018
7	140(139.23)	140.16	0.4685	0.9828	0.0018
8	182(181.85)	183.23	0.6000	0.9842	0.0018
9	230(230.16)	230.17	0.7670	0.9846	0.0017
10	284(284.15)	284.79	0.9290	0.9838	0.0018
12	409(409.18)	409.91	1.3501	0.9826	0.0018
14	557(556.94)	558.02	1.8356	0.9818	0.0019
16	728(727.43)	726.59	2.4633	0.9836	0.0018
20	1137(1136.61)	1136.39	3.7768	0.9838	0.0018

Table 3.8: Simulation Result for $(1, 6, 1)$ with $(\delta_1, \delta_2, d) = (10, 15, 15)$ with $m = 5$

std	n^*	$\bar{n}$	$std\,(\bar{n})$	$\bar{P}$	$std\left(\bar{P}\right)$
1	3(2.84)	5.01	0.0010	1.0000	0.0000
2	12(11.37)	11.81	0.0011	1.000	0.0000
3	26(25.57)	26.13	0.0374	0.9950	0.0010
4	46(45.46)	45.94	0.1502	0.9946	0.0010
5	71(70.03)	71.78	0.2377	0.9948	0.0010
6	103(102.30)	103.27	0.3287	0.9960	0.0009
7	140(139.23)	140.44	0.4635	0.9956	0.0009
8	182(181.85)	181.39	0.5935	0.9944	0.0010
9	230(230.16)	231.83	0.7702	0.9956	0.0009
10	284(284.15)	283.04	0.9599	0.9954	0.0010
12	409(409.18)	409.90	3.3596	0.9936	0.0011
14	557(556.94)	561.28	1.8437	0.9946	0.0010
16	728(727.43)	722.88	2.3701	0.9960	0.0009
20	1137(1136.61)	1143.46	2.8168	0.9952	0.0010

Table 3.9: Simulation Result for $(2, 4, 2)$ with $(\delta_1, \delta_2, d) = (10, 15, 15)$ with $m = 5$

std	n^*	$\bar{n}$	$std\,(\bar{n})$	$\bar{P}$	$std\left(\bar{P}\right)$
1	3(2.84)	5.00	0.0009	1.0000	0.0000
2	12(11.37)	11.88	0.0388	0.992	0.0013
3	26(25.57)	26.12	0.0861	0.9932	0.0012
4	46(45.46)	46.07	0.1508	0.9910	0.0013
5	71(70.03)	71.49	0.2377	0.9884	0.0015
6	103(102.30)	103.08	0.3445	0.9920	0.0013
7	140(139.23)	139.46	0.4684	0.9898	0.0014
8	182(181.85)	183.01	0.6075	0.9908	0.0014
9	230(230.16)	230.26	0.7629	0.9910	0.0013
10	284(284.15)	284.65	0.9701	0.9906	0.0013
12	409(409.18)	406.96	1.3591	0.9914	0.0013
14	557(556.94)	557.08	1.8404	0.9866	0.0016
16	728(727.43)	726.89	2.3692	0.9894	0.0014
20	1137(1136.61)	1141.74	3.8265	0.9902	0.0014

Table 3.10: Simulation Result for $(4, 0, 0, 4)$ with $(\delta_1, \delta_2, d) = (10, 15, 15)$ with $m = 5$

std	n^*	$\bar{n}$	$std\,(\bar{n})$	$\bar{P}$	$std\left(\bar{P}\right)$
1	3(2.84)	5.00	0.0008	1.0000	0.0000
2	12(11.37)	11.87	0.0383	0.9858	0.0017
3	26(25.57)	25.94	0.0853	0.9840	0.0018
4	46(45.46)	46.04	0.1524	0.9858	0.0017
5	71(70.03)	71.44	0.2343	0.9832	0.0018
6	103(102.30)	102.60	0.3357	0.9816	0.0019
7	140(139.23)	139.75	0.4641	0.9820	0.0019
8	182(181.85)	182.40	0.6068	0.9850	0.0017
9	230(230.16)	230.82	0.7635	0.9838	0.0018
10	284(284.15)	284.43	0.9384	0.9836	0.0018
12	409(409.18)	409.14	1.3440	0.9840	0.0018
14	557(556.94)	555.58	1.8293	0.9830	0.0018
16	728(727.43)	727.75	2.4531	0.9842	0.0018
20	1137(1136.61)	1139.80	3.7612	0.9826	0.0018

Table 3.11: Simulation Result for $(2, 2, 2, 2)$ with $(\delta_1, \delta_2, d) = (10, 15, 15)$ with $m = 10$

std	n^*	$\bar{n}$	$std\left(\bar{n}\right)$	$\bar{P}$	$std\left(\bar{P}\right)$
1	3(2.53)	10.00	0.0000	1.0000	0.0000
2	10(10.14)	10.99	0.0175	0.9854	0.0017
3	23(22.80)	23.32	0.0513	0.9772	0.0021
4	41(40.55)	41.15	0.0899	0.9770	0.0021
5	64(63.36)	63.99	0.1388	0.9764	0.0021
6	92(91.23)	91.75	0.1999	0.9788	0.0020
7	124(124.18)	124.78	0.2760	0.9766	0.0021
8	162(162.19)	162.66	0.3630	0.9784	0.0021
9	206(205.27)	205.38	0.4520	0.9790	0.0020
10	254(253.42)	253.92	0.5526	0.9758	0.0022
12	365(364.93)	365.44	0.8250	0.9776	0.0021
14	497(496.71)	497.13	1.1158	0.9758	0.0022
16	649(648.77)	647.30	1.4403	0.9780	0.0021
20	1014(1013.70)	1015.56	2.2501	0.9728	0.0023

Table 3.12: Simulation Result for $(3, 1, 1, 3)$ with $(\delta_1, \delta_2, d) = (10, 15, 15)$ with $m = 10$

std	n^*	$\bar{n}$	$std\left(\bar{n}\right)$	$\bar{P}$	$std\left(\bar{P}\right)$
1	3(2.53)	10.00	0.0000	1.0000	0.0000
2	10(10.14)	10.97	0.0170	0.9890	0.0015
3	23(22.80)	23.37	0.0512	0.9830	0.0018
4	41(40.55)	40.97	0.0895	0.9818	0.0019
5	64(63.36)	63.91	0.1410	0.9790	0.0020
6	92(91.23)	91.67	0.2034	0.9810	0.0019
7	124(124.18)	124.66	0.2762	0.9810	0.0019
8	162(162.19)	162.97	0.3548	0.9818	0.0019
9	206(205.27)	206.45	0.4622	0.9812	0.0019
10	254(253.42)	253.01	0.5452	0.9812	0.0019
12	365(364.93)	365.79	0.8084	0.9812	0.0019
14	497(496.71)	496.01	1.0894	0.9814	0.0019
16	649(648.77)	646.59	1.4315	0.9800	0.0020
20	1014(1013.70)	1012.11	2.2667	0.9772	0.0021

Table 3.13: Simulation Result for $(1, 3, 3, 1)$ with $(\delta_1, \delta_2, d) = (10, 15, 15)$ with $m = 10$

std	n^*	$\bar{n}$	$std\,(\bar{n})$	$\bar{P}$	$std\left(\bar{P}\right)$
1	3(2.53)	10.00	0.0000	1.0000	0.0000
2	10(10.14)	10.97	0.0171	0.9858	0.0017
3	23(22.80)	23.22	0.0503	0.9820	0.0019
4	41(40.55)	41.06	0.0900	0.9810	0.0019
5	64(63.36)	63.89	0.1401	0.9790	0.0020
6	92(91.23)	91.87	0.2043	0.9798	0.0020
7	124(124.18)	124.72	0.2746	0.9814	0.0019
8	162(162.19)	162.87	0.3592	0.9808	0.0019
9	206(205.27)	205.89	0.4491	0.9808	0.0019
10	254(253.42)	253.65	0.5667	9.9820	0.0019
12	365(364.93)	366.21	0.8269	0.9788	0.0020
14	497(496.71)	498.82	1.1090	0.9792	0.0020
16	649(648.77)	650.21	1.4233	0.9802	0.0020
20	1014(1013.70)	1012.39	2.2312	0.9764	0.0021

Table 3.14: Simulation Result for $(1, 6, 1)$ with $(\delta_1, \delta_2, d) = (10, 15, 15)$ with $m = 10$

std	n^*	$\bar{n}$	$std\,(\bar{n})$	$\bar{P}$	$std\left(\bar{P}\right)$
1	3(2.53)	10.00	0.0000	1.0000	0.0000
2	10(10.14)	10.99	0.0174	0.9950	0.0010
3	23(22.80)	23.34	0.0503	0.9938	0.0011
4	41(40.55)	41.08	0.0900	0.9956	0.0009
5	64(63.36)	63.87	0.1409	0.9946	0.0010
6	92(91.23)	91.81	0.2067	0.9944	0.0010
7	124(124.18)	124.54	0.2736	0.9946	0.0010
8	162(162.19)	162.18	0.3634	0.9942	0.0011
9	206(205.27)	205.59	0.4622	0.9936	0.0011
10	254(253.42)	254.19	0.5515	0.9932	0.0012
12	365(364.93)	365.51	0.8178	0.9952	0.0009
14	497(496.71)	497.25	1.0977	0.9946	0.0010
16	649(648.77)	650.10	1.4291	0.9942	0.0011
20	1014(1013.70)	1019.13	2.2676	0.9942	0.0011

Table 3.15: Simulation Result for $(2, 4, 2)$ with $(\delta_1, \delta_2, d) = (10, 15, 15)$ with $m = 10$

std	n^*	$\bar{n}$	$std\,(\bar{n})$	$\bar{P}$	$std\,(\bar{P})$
1	3(2.53)	10.00	0.0000	1.0000	0.0000
2	10(10.14)	10.99	0.0173	0.9918	0.0013
3	23(22.80)	23.30	0.0508	0.9910	0.0013
4	41(40.55)	41.07	0.0908	0.9876	0.0016
5	64(63.36)	63.73	0.1397	0.9888	0.0015
6	92(91.23)	91.65	0.1990	0.9864	0.0016
7	124(124.18)	124.84	0.2754	0.9876	0.0016
8	162(162.19)	162.88	0.3605	0.9922	0.0012
9	206(205.27)	206.25	0.4636	0.9876	0.0016
10	254(253.42)	253.00	0.5559	0.9904	0.0014
12	365(364.93)	365.94	0.8143	0.9862	0.0016
14	497(496.71)	495.78	1.1093	0.9870	0.0016
16	649(648.77)	650.68	1.4442	0.9872	0.0016
20	1014(1013.70)	1013.69	2.2622	0.9864	0.0016

Table 3.16: Simulation Result for $(4, 0, 0, 4)$ with $(\delta_1, \delta_2, d) = (10, 15, 15)$ with $m = 10$

std	n^*	$\bar{n}$	$std\,(\bar{n})$	$\bar{P}$	$std\,(\bar{P})$
1	3(2.53)	10.00	0.0000	1.0000	0.0000
2	10(10.14)	10.98	0.0175	0.9868	0.0016
3	23(22.80)	23.36	0.0509	0.9846	0.0017
4	41(40.55)	41.10	0.0898	0.9792	0.0020
5	64(63.36)	63.87	0.1428	0.9800	0.0020
6	92(91.23)	91.82	0.2017	0.9804	0.0020
7	124(124.18)	124.70	0.2789	0.9798	0.0020
8	162(162.19)	163.05	0.3606	0.9810	0.0019
9	206(205.27)	205.53	0.4605	0.9800	0.0020
10	254(253.42)	253.21	0.5618	0.9792	0.0020
12	365(364.93)	366.53	0.8129	0.9834	0.0018
14	497(496.71)	496.86	1.1096	0.9772	0.0021
16	649(648.77)	649.87	1.4270	0.9800	0.0020
20	1014(1013.70)	1014.06	2.2349	0.9826	0.0018

simulations are similar to the results obtained in the earlier Tables.

Table 3.17: Simulation Result for $(2, 2, 2, 2)$ with $(\delta_1, \delta_2, d) = (10, 12, 18)$

starting sample	std	n^*	$\bar{n}$	$std\,(\bar{n})$	$\bar{P}$	$std\left(\bar{P}\right)$
m=5	1	18(17.76)	18.19	0.0592	0.9830	0.0018
	2	71(71.03)	71.95	0.2405	0.9838	0.0018
	3	160(159.36)	160.22	0.5351	0.9828	0.0018
	4	284(284.15)	284.27	0.9608	0.9804	0.0020
	5	444(443.99)	445.57	1.5052	0.9802	0.0020
	6	640(639.34)	641.26	2.1365	0.9820	0.0019
	7	870(870.21)	870.16	2.9484	0.9820	0.0019
	8	1137(1136.61)	1132.11	3.7346	0.9816	0.0019
	9	1439(1438.52)	1447.64	4.9001	0.9836	0.0018
	10	1776(1775.96)	1770.72	5.9065	0.9824	0.0019
	12	2558(2557.38)	2558.28	8.4012	0.9830	0.0018
	14	3481(3480.87)	3500.72	11.5976	0.9820	0.0019
	16	4547(4546.45)	4541.79	15.3050	0.9880	0.0015
	20	7104(7103.82)	7123.40	23.2683	0.9814	0.0019
m=10	1	16(15.84)	16.39	0.0360	0.9802	0.0020
	2	64(63.36)	63.95	0.1422	0.9740	0.0023
	3	143(142.55)	142.88	0.3170	0.9786	0.0020
	4	254(253.42)	254.09	0.5751	0.9772	0.0021
	5	396(395.98)	396.16	0.8719	0.9782	0.0021
	6	571(570.20)	571.90	1.2631	0.9786	0.0020
	7	776(776.11)	777.90	1.7084	0.9772	0.0021
	8	1014(1013.70)	1011.00	2.2409	0.9760	0.0022
	9	1283(1282.96)	1278.94	2.8283	0.9754	0.0022
	10	1583(1583.90)	1579.39	3.4873	0.9768	0.0021
	12	2281(2280.82)	2273.33	4.9799	0.9796	0.0020
	14	3105(3104.45)	3115.03	7.0656	0.9792	0.0020
	16	4055(4054.79)	4063.83	8.8890	0.9768	0.0021
	20	6336(6335.61)	6310.97	14.4452	0.980	0.0020

Table 3.18: Simulation Result for $(2,2,2,2)$ with $(\delta_1, \delta_2, d) = (10, 18, 12)$

starting sample	std	n^*	$\bar{n}$	$std\,(\bar{n})$	$\bar{P}$	$std\left(\bar{P}\right)$
m=5	1	1(1.11)	5.00	0.0000	1.0000	0.0000
	2	5(4.44)	5.40	0.0099	0.9954	0.0010
	3	10(9.99)	10.51	0.0339	0.9884	0.0015
	4	18(17.76)	18.26	0.0595	0.9838	0.0018
	5	28(27.75)	28.06	0.0919	0.9844	0.0018
	6	40(39.96)	40.54	0.1326	0.9828	0.0018
	7	55(54.39)	54.83	0.1810	0.9854	0.0017
	8	71(71.04)	71.73	0.2399	0.9840	0.0018
	9	90(89.91)	90.22	0.3010	0.9874	0.0016
	10	111(110.99)	111.88	0.3710	0.9844	0.0018
	12	160(159.84)	160.35	0.5279	0.9824	0.0019
	14	218(217.55)	217.41	0.7335	0.9830	0.0018
	16	284(284.15)	285.05	0.9482	0.9828	0.0018
	20	444(443.99)	445.56	1.4894	0.9844	0.0018
m=10	1	1(0.99)	5.00	0.0000	1.0000	0.0000
	2	4(3.96)	5.38	0.0098	0.9936	0.0011
	3	9(8.91)	10.32	0.0101	0.9898	0.0014
	4	16(15.84)	16.35	0.0360	0.9792	0.0020
	5	25(24.75)	25.24	0.0548	0.9796	0.0020
	6	36(35.64)	36.15	0.0791	0.9778	0.0021
	7	48(48.51)	49.10	0.1081	0.9744	0.0022
	8	64(63.36)	63.84	0.1413	0.9776	0.0021
	9	80(80.19)	80.75	0.1797	0.9780	0.0021
	10	99(98.99)	99.28	0.2195	0.9776	0.0021
	12	143(142.55)	143.31	0.3177	0.9750	0.0022
	14	194(194.03)	194.85	0.4367	0.9788	0.0020
	16	254(253.42)	254.53	0.5576	0.9770	0.0021
	20	396(395.97)	395.11	0.8894	0.9780	0.0021

Chapter 4

A Nonparametric Procedure

4.1 Introduction

In the Chapters 1-3, we have focused on partitioning a set of normally distributed populations with respect to a control. In this Chapter, we will generalize the assumption of the normality of the distribution by not assuming that the populations are from a known distribution. In order for the partition problem to be well defined, we will assume that the distributions are symmetric.

Let $\pi_0, \pi_1, \pi_2, \cdots, \pi_k$, (with π_0 as the standard or the control population) denote $k+1$ independent populations with cumulative distribution function $F(x-\Delta_0)$, $F(x-\Delta_1)$, $\cdots$, $F(x-\Delta_k)$ respectively. Note that the cumulative distribution function $F\,(x)$ is assumed to be continuous and symmetric with Δ_0, Δ_1, $\cdots$, Δ_k are the centers of symmetry of the respective distribution. Here, the function $F(x)$ and the parameters Δ_0, Δ_1, $\cdots$, Δ_k are all assumed to be unknown. The problem considered in this chapter is to partition all k populations as “Good” and “Bad” populations by comparing the comparing the centers of symmetry Δ_i, $i = 1, \cdots, k$ with the control population Δ_0. We will require that the partition satisfies a pre specified probability of correct decision (CD), P^*, $2^{-k} < P^* < 1$. Denote Λ as a subclass of continuous and symmetric distribution satisfying certain regularity conditions to be specified later, and Ω as the set of all populations with $\omega = (\Delta_1, \Delta_2, \cdots, \Delta_k)$. In this chapter, we will construct a purely sequential procedure which satisfies the requirement $\liminf P(CD) > P^*$ for all $\omega \in \Omega$ and all $F \in \Lambda$. Given two arbitrary but fixed numbers δ_1, δ_2, and $\delta_1 < \delta_2$, as in Tong (1969), let us denote the three subsets for Ω along the lines of Bechhofer's (1954) indifference-zone formulation as:

$$\begin{cases} \Omega_L = \{\pi_i : \Delta_i \leq \Delta_0 + \delta_1\} \\ \Omega_M = \{\pi_i : \Delta_0 + \delta_1 < \Delta_i < \Delta_0 + \delta_2\} \\ \Omega_R = \{\pi_i : \Delta_i \geq \Delta_0 + \delta_2 .\} \end{cases} \tag{4.1.1}$$

Note that, as in chapter 1, Ω_L is the set of "Bad" populations, Ω_R is the set of "Good" populations, and Ω_M is set of "Indifferent" populations.

Having recorded an independent sample of size n, $X_{i1}, X_{i2}, \cdots, X_{in}$ from π_i, an appropriate statistic $L_i\,(n)$ is proposed to estimate the center of symmetry Δ_i, where it is assumed that $L_i\,(n)$ has a $N\,(\Delta_i, 1/(nA^2))$, as $n \to \infty$ for $i = 1, \cdots, k$, $F \in \Lambda$. Here A is a finite and positive function of F. For the literature of nonparametric procedures in the area of selecting the best population the reader is refereed to Geertsema (1972) and Mukhopadhay and Solanky (1993). In Mukhopadhay and Solanky (1993) the authors constructed a nonparametric accelerated sequential procedure to select the population with the largest center of symmetry.

A natural decision rule will be to compare each of $L_i(n), i = 1, \cdots, k$, with $L_0(n)$ and partition according to the rule:

$$\begin{cases} P_L = \{\pi_i : L_i(n) - L_0(n) < d, i = 1, \cdots, k\} \\ P_R = \{\pi_i : L_i(n) - L_0(n) \geq d, i = 1, \cdots, k\}, \end{cases} \tag{4.1.2}$$

where $d = (\delta_1 + \delta_2)/2$. we write $\delta^* = (\delta_2 - \delta_1)/2$. Along the lines of Geertsema (1972) and Mukhopadhyay anf Solanky (1993) we will assume that the following regularity conditions are satisfied by the distribution and the stopping rule which determines the sample size N: noindent **Assumptions:** For all $\omega(\delta^*) \in \Omega$ and $F \in \Lambda$

1. $n^{1/2}\,(L_i\,(n) - \Delta_i) = A^{-1} Z_i\,(n) + o\,(1)$ a.s. as $n \to \infty$ where $Z_i\,(n)$ is a standardized average of independent and identically distributed random variables having finite second

moment and $0 < A = A(F) < \infty$.

2. $\lim S_n^2 = A^{-2}$ a.s. as $n \to \infty$.

3. The set $\{\delta^2 N(\delta) : \delta > 0\}$ is uniformly integral.

4.2 Nonparametric Purely Sequential Procedure

First we will construct a purely sequential procedure which has the desired property that $\liminf P(CD) \geq P^*$ whenever $\theta \in \Omega(\delta^*)$ and $F \in \Lambda$, as $\delta^* \to 0$. Next, following the steps from Chapter 2, one can derive that $P(CD)$ is asymptotically (as $\delta^* \to \infty$) at least P^* if $n \geq n^* = 2b^2(A\delta^*)^{-2}$, where "$b$" is a constant which depends on the values of k and P^*. The values of constant $b = b(k, P^*)$ have been tabulated in Tong (1969) and also in Solanky and Wu (2004). However, not that n^* defined above is unknown since A is unknown. In order to overcome this, we have constructed a purely sequential procedure. The purely sequential procedure starts with m (a suitable positive integer) observations from each population π_i, $i = 0, 1, \cdots, k$. And, we continue sampling one observation at a time according to the taking stopping rule:

$$N = N(\delta^*) = \inf\left\{n \geq m : n \geq 2b^2 S_n^2/\delta^{*2}\right\} \tag{4.2.3}$$

where S_n^2 is an appropriately defined estimator of A based on the control and all k populations. Note that the estimator S_n^2 depends on the choice of the nonparametric estimator being used to estimate the center of symmetry $\Delta_i, i = 0, 1, \cdots, k$. Next, we provide a therem to summarize the basic properties of the purely sequential procedure.

Theorem 5 *Under the Assumptions described above, the purely sequential procedure satisfies the following properties for all $\omega(\delta^*) \in \Omega$ and $F \in \Lambda$:*

1. *$N(\delta^*) \to \infty$ monotonically as $\delta^* \to 0$ a.s..*

2. *$E(N(\delta^*)) \to \infty$ as $\delta^* \to 0$.*

3. $\lim \delta^{*2} N(\delta^*) = 2b^2/A^2$ *a.s.*.

4. $\liminf P(CD) \geq P^*$ *as* $\delta^* \to 0$.

Proof. First we introduce an estimator for the center of symmetry. Let $L_i(n)$ be the Hodges-Lehmann estimator for the center of symmetry of the ith population based on n observation, i.e., the sample median of the $n(n+1)/2$ quantiles $(X_{ij} + X_{ij'})/2$ for $j \leq j'$, j, $j' = 1, \cdots, n$; $i = 0, 1, \cdots, k$. Choose the following estimator of A^{-2}

$$S_n^2 = \frac{n\left((k+1)K_\alpha^2\right)^{-1}}{4} \sum_{i=0}^{k} \left(W_{n,a(n)}(i) - W_{n,b(n)}(i)\right)^2. \tag{4.2.4}$$

Where $W_{n,1}(i) \leq W_{n,2}(i) \leq \cdots \leq W_{n,n(n+1)/2}(i)$ are the ordered $(X_{ij} + X_{ij'})/2$ for $1 \leq j \leq j' \leq n$ and for $i = 0, 1, \cdots, k$. The sequence $\{a(n)\}$ and $\{b(n)\}$ are now given by

$$\begin{aligned} b(n) &= \max\left\{1, \left[n(n+1)/4 - K_\alpha (n(n+1)(2n+1)/24)^{\frac{1}{2}}\right]\right\} \\ a(n) &= n(n+1)/2 - b(n) + 1 \end{aligned} \tag{4.2.5}$$

here $[x]$ is the largest integer less than or equal to x. K_α is defined by $\phi(K_\alpha) = 1 - \alpha$ for some $1/2 < \alpha < 1$. It is well known in the statistical literature that $L_i(n)$, the Hodges-Lehmann estimator, is a consistent estimator of the center of symmetry.

Next, note that $N(\delta_1^*) \geq N(\delta_2^*)$ w.p. 1 if $0 < \delta_1^* < \delta_2^*$, that is $N(\delta^*)$ is nondecreasing in δ^*. Now the *assumption 1.1* will lead to part (1). Part (2) follows by applying the monotone convergence theorem. Since the stopping rule is

$$N(\delta^*) = \inf\left\{n \geq m_0 : n \geq 2b^2 S_n^2/\delta^{*2}\right\},$$

then the basic inequality (2.4.3) for proof of *theorem 2.4.1* in Mukhopadhyay and Solanky (1994) simplifies to

$$2b^2S_n^2/\delta^{*2} \leq N \leq m_0 + 2b^2S_{n-1}^2/\delta^{*2}. \tag{4.2.6}$$

Now multiply δ^{*2} through out (4.2.6) and take limits as $\delta^* \to 0$, this leads to part (3). For the population π_i, statistic $L_i(N)$ is proposed to estimate Δ_i. For $\theta \in \Omega(\delta^*)$, we have

$$\begin{aligned}
& P\left(CD|\,\mu^0\left(r\right), \sigma^2, R\right) \\
= \;& P\left\{L_i\left(N\right) - L_0\left(N\right) < d, 0 < i \leq r; L_j\left(N\right) - L_0\left(N\right) \geq d, r < j \leq k\right\} \\
= \;& P\Bigg\{\left(\left(L_i\left(N\right) - \Delta_i\right) - \left(L_0\left(N\right) - \Delta_0\right)\right)\frac{\sqrt{n^*}A}{\sqrt{2}} < \left(d - \left(\Delta_i - \Delta_0\right)\right)\frac{\sqrt{n^*}A}{\sqrt{2}}, 0 < i \leq r; \\
& \left(\left(L_j\left(N\right) - \Delta_j\right) - \left(L_0\left(N\right) - \Delta_0\right)\right)\frac{\sqrt{n^*}A}{\sqrt{2}} \geq \left(d - \left(\Delta_j - \Delta_0\right)\right)\frac{\sqrt{n^*}A}{\sqrt{2}}, r < j \leq k\Bigg\} \\
= \;& P\left\{\frac{Z_i - Z_0}{\sqrt{2}} < \frac{\sqrt{n^*}A\delta^*}{\sqrt{2}}, 0 < i \leq r; \frac{Z_j - Z_0}{\sqrt{2}} \geq -\frac{\sqrt{n^*}A\delta^*}{\sqrt{2}}, r < j \leq k\right\} \\
= \;& P\left\{Y_i\left(N\right) \leq \frac{\sqrt{n^*}A\delta^*}{\sqrt{2}}, i = 1, \cdots, k\right\}
\end{aligned} \tag{4.2.7}$$

Where

$$Z_i(N) = \sqrt{n^*}A\left(L_i(N) - \Delta_i\right)$$

for $i = 1, \ldots, k$,

$$Y_i(N) = \frac{Z_i(N) - Z_0(N)}{2}, \qquad Y_j(N) = \frac{Z_0(N) - Z_j(N)}{2}$$

for $0 < i \leq r, \quad r < j \leq k$. If we define the $(k \times k)$ covariance matrix $\Sigma_r = (\sigma_{ij})$ by

$$\begin{aligned}
\sigma_{ij} &= 1, \;\; for \;\; i = j; \\
&= \frac{1}{2}, \;\; for \;\; 0 < i, j \leq r \;\; or \;\; r < i, j \leq k; \\
&= -\frac{1}{2}, \;\; for \;\; 0 < i \leq r \;\; and \;\; r < j \leq k.
\end{aligned}$$

then

$$P\left(CD|\,\mu^0\left(r\right),\sigma^2,R\right)=\int_{-\infty}^{\frac{\sqrt{n^*}A\delta^*}{\sqrt{2}}}\cdots\int_{-\infty}^{\frac{\sqrt{n^*}A\delta^*}{\sqrt{2}}}\left(2\pi\right)^{-\frac{k}{2}}\left|\Sigma_r\right|^{-\frac{k}{2}}\exp\left(-\frac{1}{2}y'\Sigma_r^{-1}y\right)\prod_{i=1}^{k}dy_i \tag{4.2.8}$$

Equation (4.2.8) gives the infimum of the $P\left(CD\right)$ under R for the set of all configurations such that there are r populations from Ω_L (Bad populations) and $(k-r)$ populations from Ω_R (Good populations). The rhs of (4.2.8) achieves a minimum over all $r\ (0<r\leq k)$ under the LFC.

Let $b=b(P,k)$ be the solution of the equation

$$P=\int_{-\infty}^{b}\int_{-\infty}^{b}\cdots\int_{-\infty}^{b}\left(2\pi\right)^{-\frac{k}{2}}\left|\Sigma_k\right|^{-\frac{k}{2}}\exp\left(-\frac{1}{2}y'\Sigma_k^{-1}y\right)\prod_{i=1}^{k}dy_i$$

Also, for any real number c and q, let

$$P_q\left(c\right)=\int_{-\infty}^{c}\int_{-\infty}^{c}\cdots\int_{-\infty}^{c}\left(2\pi\right)^{-\frac{q}{2}}\left|\Sigma_q\right|^{-\frac{q}{2}}\exp\left(-\frac{1}{2}y'\Sigma_q^{-1}y\right)\prod_{i=1}^{q}dy_i \tag{4.2.9}$$

where the $(q\times q)$ covariance matrix $\Sigma_q=(\sigma_{ij})$ is such that

$$\begin{aligned}\sigma_{ij}&=1,\ \ for\ \ i=j;\\&=\frac{1}{2},\ \ for\ \ i\neq j.\end{aligned}$$

Define

$$A=\left[Y_i\leq b,i=1,\cdots,r\right]$$
$$B=\left[Y_i\leq b,i=r+1,\cdots,k\right]$$

then

$$P_r\left(b\right)+P_{k-r}\left(b\right)=1+P^*\Rightarrow P\left(A\cap B\right)=P\left\{Y_i\left(N\right)\leq b,i=1,\cdots,k\right\}=P\left(CD|\,\mu^0\left(r\right),\sigma^2,R\right)\geq P^*$$

i.e. $\liminf P\left(CD\right) \geq P^{*}$, which is the part (4). This completes the proof of the theorem.

■

4.3 Mote Carlo Simulation Results

In our simulation study, we considered $k = 8$ independent populations and one control population. To construct the LFC, we generated 4 populations with the center of symmetry equal to $\mu_0 - \delta$, and remaining 4 populations are generated to have the center of symmetry as $\mu_0 + \delta$. The control population is generated to have the center of symmetry as μ_0. Without loss of generality, we set $\mu_0 = 0$. For $k = 8$ and $P^{*} = .95$, the value of the constant b equals 2.44177 from Solanky and Wu (2004). Next, we considered the following symmetric distributions: normal distribution, Laplace distribution, t-distribution, uniform distribution, and mixture of two normal distributions. For these distributions, the parameter A^2 is given by

$$A^2 = 12\bigg(\int f^2\left(x\right)dx\bigg)^2$$

$f\left(x\right)$ is the density function for normal distribution, Laplace distribution, t-distribution, uniform distribution and mixture of two normal distributions,respectably. In our simulations, $Normal\left(0,1\right)$, the Laplace distribution with $\mu = 0, b = \sqrt{2}/2$, t-distribution with $df = 5$, $U\left(-1,1\right)$, and two mixed normal distribution: $0.35N\left(x_1;0,1\right)+0.65N\left(x_2;0,2\right)$ and $0.8N\left(x_1;0,1\right)+0.2N\left(x_2;0,5\right)$ were used here.

$$A^2_{Normal} = 12\left(\int_{-\infty}^{+\infty}\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\right)^2 dx\right)^2 = 12\left(\int_{-\infty}^{+\infty}\frac{1}{2\pi}e^{-x^2}dx\right)^2 = 0.9549$$

$$A^2_{Laplace} = 12\left(\int_{-\infty}^{+\infty}\left(\frac{1}{\sqrt{2}}e^{-\sqrt{2}|x|}\right)^2 dx\right)^2 = 12\left(\int_{-\infty}^{+\infty}\frac{1}{2}e^{-2\sqrt{2}|x|}dx\right)^2 = 1.5$$

$$A^2_{Uniform} = 12\left(\int_{-1}^{1}\left(\frac{1}{b-a}\right)^2 dx\right)^2 = 12\left(\int_{-1}^{1}\left(\frac{1}{2}\right)^2 dx\right)^2 = 3$$

$$A_t^2 = 12\left(\int_{-\infty}^{+\infty}\left(\frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{x^2}{v}\right)^{-\frac{v+1}{2}}\right)^2 dx\right)^2\Bigg|_{v=5} = 0.7447$$

$$A_{Mixed1}^2 = 12\left(\int_{-\infty}^{+\infty}\left(0.35\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} + 0.65\frac{1}{2\sqrt{2\pi}}e^{-\frac{x^2}{2\cdot 2^2}}\right)^2 dx\right)^2 = 0.3689$$

$$A_{Mixed2}^2 = 12\left(\int_{-\infty}^{+\infty}\left(0.80\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} + 0.20\frac{1}{5\sqrt{2\pi}}e^{-\frac{x^2}{2\cdot 5^2}}\right)^2 dx\right)^2 = 0.5183$$

After, we obtained the value of the A^2 for each distribution, the value of δ was determined by $\delta = \sqrt{\frac{2b^2}{n^* A^2}}$. The values of n^* which we selected were 50, 100, 200, 400, and 800. For each value of n^*, the corresponding value of δ was obtained and those values have been summarized in the Tables (4.1) to (4.6). As described earlier, the estimator S_n^2 as described in (4.2.4) is used to estimate the unknown parameter A^{-2}. Note that the purely sequential rule does not rely upon the knowledge of A^2.

Next, we generated data from the normal distribution with $\sigma = 1$, Laplace distribution with $\lambda = \sqrt{2}/2$, t-distribution with $df = 5$, uniform distribution, two mixed normal distribution given by $0.35N(x_1; 0, 1) + 0.65N(x_2; 0, 2)$ and $0.8N(x_1; 0, 1) + 0.2N(x_2; 0, 5)$, respectively. Recall that from the section 4.3, the Hodges-Lehmann estimator holds for $1/2 < \alpha < 1$. In the simulations we have considered several possible choices of the α and studied the impact of α on the estimation of A^2. The simulation results are reported in the Tables (4.1) - (4.6).

Table 4.1: Simulation Results for Normal distribution with $\sigma = 1$

α	δ	n^*	$\bar{n}$	$std(\bar{n})$	$\bar{P}$	$std(\bar{P})$
0.75	0.499	50	52.050	0.143	0.867	0.011
	0.353	100	102.298	0.189	0.870	0.011
	0.250	200	202.597	0.263	0.870	0.011
	0.177	400	402.507	0.376	0.877	0.010
	0.125	800	803.636	0.492	0.847	0.011
0.85	0.499	50	52.958	0.122	0.865	0.011
	0.353	100	103.046	0.180	0.865	0.011
	0.250	200	203.638	0.255	0.855	0.011
	0.177	400	403.382	0.365	0.857	0.011

From the Tables (4.1) and (4.2), note that the purely sequential procedure (4.2.3) is over sampling by roughly 2-3 observation when the population is normally distributed and by just below 10 observations for the Laplace distribution. Also, note that the estimated probability of correct selection is below the target value of $.95$ for the normal case. However, for the Laplace distribution, the estimated probability of correct selection matches the target value of $.95$ quite well. This feature of the statistical estimation should not come as a surprise. The Hodges-Lehmann estimator is more appropriate when the distribution has tails longer than normal distribution tails. That is, when the distribution is close to being normally distributed then the partition procedures designed for normally distributed populations, such as the ones described in the Chapters 1-3 work quite well. However, if the tails are significantly longer than the normal tails, like for the Laplace distribution, then the nonparametric partition procedures are more appropriate.

Table 4.2: Simulation Results for Laplace distribution with $\lambda = \frac{\sqrt{2}}{2}$

α	δ	n^*	$\bar{n}$	$std\left(\bar{n}\right)$	$\bar{P}$	$std\left(\bar{P}\right)$
0.75	0.399	50	55.570	0.183	0.970	0.005
	0.282	100	106.486	0.264	0.978	0.005
	0.199	200	206.231	0.351	0.969	0.005
	0.141	400	408.060	0.514	0.975	0.005
	0.099	800	808.374	0.687	0.975	0.005
0.85	0.399	50	56.872	0.175	0.976	0.005
	0.282	100	107.685	0.244	0.975	0.005
	0.199	200	207.481	0.347	0.978	0.005
	0.141	400	409.598	0.505	0.969	0.006

In the Table (4.3), the underlying distribution is t-distribution with 5 degrees of freedom. The distribution has tails longer than a normal distribution but shorter than the Laplace distribution. Note that the estimated probability of correct selection is somewhat below the target value of $.95$ for smaller values of α. However, as α increases the estimated probability of correct selection is approaching the target value of $.95$.

Next, we have considered the Uniform distribution case which have tails even shorter than the normal tails. One will note that the estimated probability of correct selection is well below the

Table 4.3: Simulation Results for T-distribution with $df = 5$

α	δ	n^*	$\bar{n}$	$std\left(\bar{n}\right)$	$\bar{P}$	$std\left(\bar{P}\right)$
0.75	0.566	50	52.981	0.159	0.896	0.010
	0.400	100	103.358	0.224	0.898	0.010
	0.283	200	202.923	0.269	0.893	0.010
	0.200	400	403.129	0.423	0.901	0.009
0.85	0.566	50	54.494	0.147	0.901	0.009
	0.400	100	104.488	0.209	0.909	0.009
	0.283	200	204.676	0.293	0.913	0.009
	0.200	400	404.660	0.413	0.918	0.009
0.90	0.566	50	54.605	0.144	0.928	0.008
	0.400	100	105.242	0.213	0.893	0.010
	0.283	200	204.816	0.280	0.913	0.009
0.95	0.566	50	55.769	0.135	0.929	0.008
	0.400	100	105.988	0.208	0.912	0.009
	0.283	200	205.799	0.279	0.926	0.008

Table 4.4: Simulation Results for Uniform distribution

α	δ	n^*	$\bar{n}$	$std\left(\bar{n}\right)$	$\bar{P}$	$std\left(\bar{P}\right)$
0.60	0.282	50	42.792	0.564	0.487	0.016
	0.199	100	104.732	0.409	0.599	0.016
	0.141	200	210.747	0.236	0.621	0.015
0.75	0.282	50	56.769	0.117	0.641	0.015
	0.199	100	110.106	0.129	0.64	0.015
	0.141	200	214.045	0.175	0.62	0.015
0.85	0.282	50	58.122	0.094	0.653	0.015
	0.199	100	111.698	0.114	0.610	0.015
	0.141	200	216.071	0.146	0.604	0.015
0.99	0.282	50	63.737	0.070	0.719	0.014
	0.199	100	118.374	0.089	0.648	0.015
	0.141	200	224.796	0.119	0.654	0.015

target value of $.95$. This feature is again along the lines of comments made earlier in this section about the Hodges-Lehmann estimator being more appropriate when the distribution has tails longer than normal distribution tails.

Next, we have considered the mixture of two normal populations. In the first case, we have considered the $0.35N\left(x_1; 0, 1\right) + 0.65N\left(x_2; 0, 2\right)$ which is mixture of two normal populations with

Table 4.5: Simulation Results for Mixture of two normal distributions: $X = 0.35N\left(x_1; 0, 1\right) + 0.65N\left(x_2; 0, 2\right)$

α	δ	n^*	$\bar{n}$	$std\left(\bar{n}\right)$	$\bar{P}$	$std\left(\bar{P}\right)$
0.75	0.804	50	52.859	0.162	0.903	0.009
	0.569	100	103.243	0.213	0.905	0.009
	0.402	200	203.962	0.303	0.911	0.009
0.85	0.804	50	53.685	0.140	0.916	0.007
	0.569	100	104.216	0.216	0.926	0.008
	0.402	200	204.205	0.285	0.912	0.009
0.90	0.804	50	54.817	0.143	0.909	0.009
	0.569	100	104.823	0.203	0.902	0.009
	0.402	200	204.928	0.290	0.900	0.009
0.95	0.804	50	55.676	0.142	0.928	0.008
	0.569	100	105.801	0.202	0.918	0.009
	0.402	200	206.601	0.271	0.913	0.009

somewhat long tails. The first population is the mixture has variance 1 and the second has variance of 2. In the second mixture of the two normal populations considered, we have $0.8N\left(x_1; 0, 1\right) + 0.2N\left(x_2; 0, 5\right)$. This second mixture has two normal populations again, but the two variances being 1 and 5 respectively, are farther apart. Intuitively, these two mixture cases are symmetric but are not unimodal like normal distribution or other distributions considered earlier. The two Tables below again exhibit the same behavior, the longer the tails, the better is the performance of the Hodges-Lehmann estimator.

4.4 An Example

In this section, we study the performance of the nonparametric sequential procedure via an a real-world data set. Zelazo et al. (1972) conducted a pilot investigation to see if active exercise can preserve the walking beyond the 2^nd month. In this experiment, newborn children were randomly placed into one of four treatment groups: (1) Active exercise group; (2) Passive exercise group; (3) No exercise group (these were observed weekly); and (4) Control group (observed once after 8 weeks). Traditional 12 months has been known as the mean time infants take to walk. The statisti-

Table 4.6: Simulation Results for Mixture of two normal distributions: $X = 0.8N(x_1; 0, 1) + 0.2N(x_2; 0, 5)$

α	δ	n^*	$\bar{n}$	$std(\bar{n})$	$\bar{P}$	$std(\bar{P})$
0.75	0.678	50	54.424	0.187	0.952	0.007
	0.480	100	104.593	0.259	0.935	0.008
	0.339	200	205.031	0.351	0.932	0.008
0.85	0.678	50	55.826	0.169	0.934	0.008
	0.450	100	106.334	0.254	0.926	0.008
	0.339	200	206.534	0.332	0.933	0.008
	0.240	400	406.497	0.486	0.942	0.007
0.90	0.678	50	56.762	0.177	0.955	0.007
	0.480	100	106.746	0.244	0.924	0.008
	0.339	200	207.888	0.351	0.935	0.008
0.95	0.678	50	58.671	0.173	0.959	0.006
	0.480	100	108.742	0.235	0.947	0.007
	0.339	200	208.019	0.332	0.931	0.008

cal analysis confirmed that the walking-data is normally distributed with somewhat equal variance. Adopting a 12.5% improvement as significant and anything else than 8% as not significant. We took $\delta_1 = -1.5$ months, $\delta_2 = -1.0$ months, $k = 3$, and the starting sample size $m = 5$.

Table 4.7: Comparison of various Statistical Methodologies

Procedure	simulated sample size
Two-stage	71
Purely Sequential	66
Seq-Elimination-type	43
Two-stage with Elimination	201
Nonparametric Sequential	42 $(\alpha = 0.75)$ 52 $(\alpha = 0.80)$ 53 $(\alpha = 0.85)$ 60 $(\alpha = 0.90)$ 67 $(\alpha = 0.95)$

The data was analyzed via the following five procedures: (1) Two-stage procedure of Tong (1969); (2) Purely sequential procedure of Datta and Mukhopadhyay (1998); (3) Sequential elimination-type procedure of Solanky (2001); (4) Two-stage with elimination procedure of Solanky (2006); (5) Nonparametric sequential procedure proposed in this thesis. Additional samples as needed were generated via SRSWR and saved to have same data for all five procedures. Note that all the five sampling methodologies yielded the same result, that is, the active exercise group was partitioned as better than control, the passive and no exercise group were partitioned as bad compared to the control since the improvement was lower than 8%. The sample size for these five methodologies is reported in the Table above. One will note that for this example:

1. The sample size was somewhat larger for nonparametric sequential procedure. And, it increased further when the parameter α was increased. However, this was quite expected since the data is normally distributed in this case and the procedures based on normal distribution assumption are bound to perform better. Note that from the simulations, the true advantage of the non parametric procedure is when the data is not normal and has long tails.

2. The two-stage procedure with elimination needed much more samples than other methods. Again, this is not unexpected. The elimination type procedure is designed to eliminate non-competing populations early enough during the sampling process. And, if there are no non-competing populations than the two-stage procedure with elimination is unable to eliminate any population early enough to reduce the overall sample size.

To conclude, the partition procedures based on the normal distribution are quite robust and hence are recommended when the populations are normally distributed. However, if the distribution has tails much longer than the normal tails then the nonparametric partition procedures, such as one based on the Hodges-Lehmann estimator, are shown to be more appropriate.

Bibliography

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

[2] Bechhofer, R.E., Santner, T.J., and Goldsman, D.M.. *Design and analysis of experiments for statistical selection, screening, and multiple comparisons*. New York:Wiley.

[3] Chow, Y.S. and Robbins, H.. On the asymptotic theory of fixed width sequential confidence intervals for the means. *Ann. Math. Statist.*, 1965, 36, 457-462.

[4] Datta, S. and Mukhopadhyay, N.. Second-order asymptotics for multistage methodologies in partitioning a set of normal populations having a common unknown variance. *Statistics and Decisions*, 1998, 16, 191-205.

[5] Geertsema, J. C.. Nonparametric Sequential Procedures for Selecting the Best of K Populations. *Journal of the American Statistical Association*, 1972, 67, 614-616.

[6] Gupta, S.S.. On a decision rule for a problem in ranking means. Ph.D. diss., University of North Carolina-Chapel Hill, 1956.

[7] Gupta, S.S.. On some multiple decision (selection and ranking) rules. *Technometrics*, 1965, 7, 225-245.

[8] Lehmann, E.L.. Nonparametric confidence intervals for a shift parameter. *The Ann. of Math. Statist.*, 1963, 34, 1507-1512.

[9] Mukhopadhyay, N. and Solanky, T.K.S., A nonparametric accelerated sequential procedure for selecting the largest center of symmetry. *Nonparametric Statistics*, 1993 3, 155-166.

[10] Mukhopadhyay, N. and Solanky, T.K.S.. *Multistage selection and ranking procedures*. Marcel Dekker: New York, 1994

[11] Norman L. Johnson and Samuel Kotz.. *Distribution in Statistics: Continuous Multivariate Distribution*. Jonh Wiley & Sons, Inc. 2013

[12] Paulson, E.. A sequential procedure for selecting the population with the largest mean from k normal populations. *Ann. Math. Statist.*, 1964, 35, 174-180.

[13] Solanky, T.K.S.. A sequential procedure with elimination for partitioning a set of normal populations having a common unknown variance. *Seq. Anal.*, 2001, 20, 279-292.

[14] Solanky, T.K.S. and Yuefeng, Y.. On unbalanced multistage methodologies for the partition problem. *Proceeding of the International Sri Lankan Statistical Conference: Version of Futuristic Methodologies*, 2004, 447-466.

[15] Solanky, T.K.S.. A Two-stage procedure with elimination for partitioning a set of normal populations with respect to acontrol. *Seq. Anal.*, 2006,25, 297-310.

[16] Tamhane, A.C. and Bechhofer, R.E.. A two-stage minimax procedure with screening for selecting the largest the largest normal mean. *Communications in Statistics, Series A*, 1977, 6, 1003-1033.

[17] Tamhane, A.C. and Bechhofer, R.E.. A two-stage minimax procedure with screening for selecting the largest the largest normal mean (II). *Communications in Statistics, Series A*, 1979, 8, 337-358.

[18] Tong, Y.L.. On partitioning a set of normal populations by their locations with respect to a control. *Ann. Math. Statist.*, 1969, 40, 1300-1324.

[19] Wiener, N.. The ergodic theorem. *Duke Math. J.*, 1939, 5, 1-18.

[20] Woodroofe, M.. Second order approximations for sequential point and interval estimation. *The Anna. of Stat.*, 1977, 5, 984-995.

[21] Zelazo. “Walking” in the Newborm. *Science*, 1972, 176, 314-315.

[22] Ferchoff, L.. http://www.mathworks.com/matlabcentral/fileexchange/12447-mcint

Appendix A

R Source Code for Single-stage Procedure

```
SingleStage<-function(cc){
kk<-8
q1<-4
q2<-0
q3<-0
q4<-4
delta1<-5
std<-1
bb<- 2.6959
csize<-1
itersize<-20000
pvalue<-rep(1.,itersize)
dd <- 2*sqrt(2) * std * bb/sqrt(cc)
cat("q1=",q1,"q2=",q2,"q3=",q3,"q4=",q4,"\n")
cat("bb=",bb,"\n")
cat("standard deviation for each group=",std,"\n")
cat("delta1=",delta1,"dd=",dd,"\n")
cat("No. of Iterations=",itersize,"\n")
cat("kk=",kk,"control size=",csize,"\n")

for(iter in 1:itersize){
m1<-rep(0.,kk)
pzeromean<-rep(0.,csize)
mdat <- matrix(rnorm(kk*cc,0,1), nrow = cc, ncol=kk, byrow=TRUE)
pzero <- matrix(rnorm(csize*cc,0,1), nrow = cc, ncol=csize, byrow=TRUE)

for(j1 in 1:q1){
mdat[,j1]<-std*mdat[,j1]+delta1}
for(j2 in (q1+1):(q1+q2)){
mdat[,j2]<-std*mdat[,j2]+delta1+dd}
for(j3 in (q1+q2+1):(q1+q2+q3)){
```

```
mdat[,j3]<-std*mdat[,j3]+delta1+2*dd}
for(j4 in (q1+q2+q3+1):kk){
mdat[,j4]<-std*mdat[,j4]+delta1+3*dd}

#for(j2 in (q1+1):(q1+q2)){
#mdat[,j2]<-std*mdat[,j2]+delta1+1.5*dd}

for(j in 1:csize){
pzero[,j]<-std*pzero[,j]}
ones <- rep(1., cc)
onescsize <- rep(1.,csize)
#computation of means
for(j in 1:kk){
m1[j]<-(1/cc)*t(ones)%*% mdat[,j]}
for(j in 1:csize){
pzeromean[j]<-(1/cc)*t(ones)%*% pzero[,j]}
pzeromeanallc<-(1/csize)*t(onescsize)%*% pzeromean

for(j1 in 1:q1){
if((m1[j1]-pzeromeanallc)>(delta1+dd/2))     pvalue[iter]<-0}
for(j2 in (q1+1):(q1+q2)){
if((m1[j2]-pzeromeanallc)<(delta1+dd/2) | (m1[j2]-pzeromeanallc)>(delta1+5*dd/2))  pvalue[iter]<-0}
for(j3 in (q1+q2+1):(q1+q2+q3)){
if((m1[j3]-pzeromeanallc)<(delta1+dd/2) | (m1[j3]-pzeromeanallc)>(delta1+5*dd/2))   pvalue[iter]<-0}
for(j4 in (q1+q2+q3+1):kk){
if((m1[j4]-pzeromeanallc)<(delta1+5*dd/2))    pvalue[iter]<-0}
}
pbar <- mean(pvalue)
pvar <- var(pvalue)
pstdev <- pvar^(0.5)/(itersize)^0.5

cat(" cc  ",  "  pbar  ",  "  s(pbar)  ","\n")
cat(cc,  "  ", pbar,"  ", pstdev,"\n")
itersize
}
```

Appendix B

R Source Code for Purely Sequential Procedure

```
Sequential<-function(cc){
#...you need to enter the optimal sample size to run it
kk<-8
q1<-2
q2<-2
q3<-2
q4<-2
delta1<-5
std<-1
bb<- 2.6959
csize<-1
itersize<-5000
nvalue<-rep(0.,itersize)
pvalue<-rep(1.,itersize)
dd <- 2*sqrt(2) * std * bb/sqrt(cc)

cat("q1=",q1,"q2=",q2,"q3=",q3,"q4=",q4,"\n")
cat("bb=",bb,"\n")
cat("standard deviation for each group=",std,"\n")
cat("delta1=",delta1,"dd=",dd,"\n")
cat("No. of Iterations=",itersize,"\n")
cat("kk=",kk,"control size=",csize,"\n")

for(iter in 1:itersize){
m1<-rep(0.,kk)
pzeromean<-rep(0.,csize)
v1<-rep(0.,kk)
vcontrol<-rep(0.,csize)

totalsize<-2*cc
mdat <- matrix(rnorm(kk*totalsize,0,1), nrow = totalsize, ncol=kk, byrow=TRUE)
```

```
pzerofull <- matrix(rnorm(csize*totalsize,0,1), nrow = totalsize, ncol=csize, byrow=TRUE)

for(j1 in 1:q1){
mdat[,j1]<-std*mdat[,j1]+delta1}
for(j2 in (q1+1):(q1+q2)){
mdat[,j2]<-std*mdat[,j2]+delta1+dd}
for(j3 in (q1+q2+1):(q1+q2+q3)){
mdat[,j3]<-std*mdat[,j3]+delta1+2*dd}
for(j4 in (q1+q2+q3+1):kk){
mdat[,j4]<-std*mdat[,j4]+delta1+3*dd }

#for(j2 in (q1+1):(q1+q2)){
#mdat[,j2]<-std*mdat[,j2]+delta1+1.5*dd}

for(j in 1:csize){
pzerofull[,j]<-std*pzerofull[,j]}

for(n1 in 10:totalsize){
xdat <- matrix(0, nrow = n1, ncol=kk, byrow=TRUE)
pzero <- matrix(0, nrow = n1, ncol=csize, byrow=TRUE)

for(i in 1:n1){
   for(j in 1:kk){
       xdat[i,j]<-mdat[i,j]}}

for(i in 1:n1){
   for(j in 1:csize){
       pzero[i,j]<-pzerofull[i,j]}}
ones <- rep(1., n1)
oneskk <- rep(1., kk)
onescsize <- rep(1.,csize)

for(j in 1:kk){
m1[j]<-(1/n1)*t(ones)%*% xdat[,j]}
for(j in 1:csize){
pzeromean[j]<-(1/n1)*t(ones)%*% pzero[,j]}
pzeromeanallc<-(1/csize)*t(onescsize)%*% pzeromean

#for population groups
for(j in 1:kk){
v1[j]<-t(xdat[,j]-m1[j]) %*% (xdat[,j]-m1[j])
}
v1<-v1/(n1-1)

for(j in 1:csize){
vcontrol[j]<- t(pzero[,j]-pzeromean[j]) %*% (pzero[,j]-pzeromean[j])
}
vcontrol<-vcontrol/(n1-1)

#pooled variance
vppopu<-t(oneskk) %*% v1
vpcontrol<-t(onescsize) %*% vcontrol
vp<-(vppopu+vpcontrol)/(kk+csize)
term<-((csize+1)/csize)*(bb^2)*vp/((dd/2)^2)
if(n1 > term)   break
}
nvalue[iter]<-n1
for(j1 in 1:q1){
if((m1[j1]-pzeromeanallc)>(delta1+dd/2))   pvalue[iter]<-0
}
for(j2 in (q1+1):(q1+q2)){
if((m1[j2]-pzeromeanallc)<(delta1+dd/2) | (m1[j2]-pzeromeanallc)>(delta1+(5*dd)/2))   pvalue[iter]<-0
}
for(j3 in (q1+q2+1):(q1+q2+q3)){
if((m1[j3]-pzeromeanallc)<(delta1+dd/2) | (m1[j3]-pzeromeanallc)>(delta1+(5*dd)/2))   pvalue[iter]<-0
}
for(j4 in (q1+q2+q3+1):kk){
if((m1[j4]-pzeromeanallc)<(delta1+(5*dd)/2))    pvalue[iter]<-0
}
}
nbar<-mean(nvalue)
nvar<-var(nvalue)
nstdev <- nvar^(0.5)/(itersize)^0.5
pbar <- mean(pvalue)
pvar <- var(pvalue)
pstdev <- pvar^(0.5)/(itersize)^0.5
cat(" cc   nbar   ", " s(nbar)  ",  " pbar  ",  " s(pbar)  ","  dd  ","\n")
cat(cc, "  ", nbar,"  ", nstdev, "  ", pbar,"  ", pstdev,"  ", dd,"\n")
itersize
}
```

Appendix C

R Source Code for Two-Stage Procedure

```
TwoStage<-function(n1,cc,std){
#here, the starting sample size could be 5 and 10, the corresponding Tau tables are in Two-stage chapter.
#you need to input the optimal sample size---"cc"
#dd is the distance from point 1 to point 3, is actually same as the distance from point 2 to point 4.
if(n1==10) bb<-2.814163
if(n1==5)  bb<-2.979896
delta1<-10
delta2<-18
dd<-12
kk<-8
q1<-2
q2<-2
q3<-2
q4<-2
sigmasq<-std^2
csize<-1
itersize<-5000
totalsize<-4*cc
nvalue<-rep(0.,itersize)
pvalue<-rep(1.,itersize)

cat("Two-Stage procedure","\n")
cat("q1=",q1,"q2=",q2,"q3=",q3,"q4=",q4,"\n")
cat("bb=",bb,"\n")
cat("standard deviation for each group=",std,"\n")
cat("the optimal sample size=",cc,"\n")
cat("the starting sample size=",n1,"\n")
cat("delta1=",delta1,"delta2=",delta2,"dd=",dd,"\n")
cat("No. of Iterations=",itersize,"\n")
cat("kk=",kk,"control size=",csize,"\n")
```

```
for(iter in 1:itersize){
m1<-rep(0.,kk)
pzeromean<-rep(0.,csize)
v1<-rep(0.,kk)
vcontrol<-rep(0.,csize)

mdat <- matrix(rnorm(kk*totalsize,0,1), nrow = totalsize, ncol=kk, byrow=TRUE)
pzerofull <- matrix(rnorm(csize*totalsize,0,1), nrow = totalsize, ncol=csize, byrow=TRUE)

for(j1 in 1:q1){
mdat[,j1]<-std*mdat[,j1]+delta1}
for(j2 in (q1+1):(q1+q2)){
mdat[,j2]<-std*mdat[,j2]+delta2}
for(j3 in (q1+q2+1):(q1+q2+q3)){
mdat[,j3]<-std*mdat[,j3]+delta1+dd}
for(j4 in (q1+q2+q3+1):kk){
mdat[,j4]<-std*mdat[,j4]+delta2+dd }

#the following two is for the populations in the middle point of point 2 and point 3
#for(j2 in (q1+1):(q1+q2)){
#mdat[,j2]<-std*mdat[,j2]+0.5*(delta1+delta2+dd)}

for(j in 1:csize){
pzerofull[,j]<-std*pzerofull[,j]}

#Two-stage procedure starts
xdat <- matrix(0, nrow = n1, ncol=kk, byrow=TRUE)
pzero <- matrix(0, nrow = n1, ncol=csize, byrow=TRUE)

for(i in 1:n1){
   for(j in 1:kk){
       xdat[i,j]<-mdat[i,j]}}
for(i in 1:n1){
   for(j in 1:csize){
       pzero[i,j]<-pzerofull[i,j]}}

ones <- rep(1., n1)
oneskk <- rep(1., kk)
onescsize <- rep(1.,csize)

for(j in 1:kk){
m1[j]<-(1/n1)*t(ones)%*% xdat[,j]}
```

```
for(j in 1:csize){
pzeromean[j]<-(1/n1)*t(ones)%*% pzero[,j]}
pzeromeanallc<-(1/csize)*t(onescsize)%*% pzeromean

for(j in 1:kk){
v1[j]<-t(xdat[,j]-m1[j]) %*% (xdat[,j]-m1[j])
}
v1<-v1/(n1-1)

for(j in 1:csize){
vcontrol[j]<- t(pzero[,j]-pzeromean[j]) %*% (pzero[,j]-pzeromean[j])
}
vcontrol<-vcontrol/(n1-1)

vppopu<-t(oneskk) %*% v1
vpcontrol<-t(onescsize) %*% vcontrol
vp<-(vppopu+vpcontrol)/(kk+csize)
term<-((csize+1)/csize)*(bb^2)*vp/(((delta2-delta1)/2)^2)
nvalue[iter]<-max(as.integer(term)+1,n1)

mm<-nvalue[iter]
m_new<-rep(0.,kk)
pzeromean_new<-rep(0.,csize)

xdat_new <- matrix(0, nrow = nvalue[iter], ncol=kk, byrow=TRUE)
pzero_new <- matrix(0, nrow = nvalue[iter], ncol=csize, byrow=TRUE)

for(i in 1:mm){
   for(j in 1:kk){
       xdat_new[i,j]<-mdat[i,j]}}
for(i in 1:mm){
   for(j in 1:csize){
       pzero_new[i,j]<-pzerofull[i,j]}}

ones <- rep(1., mm)
oneskk <- rep(1., kk)
onescsize <- rep(1.,csize)

for(j in 1:kk){
m_new[j]<-(1/mm)*t(ones)%*% xdat_new[,j]}
```

```
for(j in 1:csize){
pzeromean_new[j]<-(1/mm)*t(ones)%*% pzero_new[,j]}
pzeromeanallc_new<-(1/csize)*t(onescsize)%*% pzeromean_new

for(j1 in 1:q1){
if((m_new[j1]-pzeromeanallc_new)>(delta1+delta2)/2)   pvalue[iter]<-0
}
for(j2 in (q1+1):(q1+q2)){
if((m_new[j2]-pzeromeanallc_new)<(delta1+delta2)/2 | (m_new[j2]-pzeromeanallc_new)>(delta1+delta2)/2+dd)   pvalue[iter]<-0
}
for(j3 in (q1+q2+1):(q1+q2+q3)){
if((m_new[j3]-pzeromeanallc_new)<(delta1+delta2)/2 | (m_new[j3]-pzeromeanallc_new)>(delta1+delta2)/2+dd)   pvalue[iter]<-0
}
for(j4 in (q1+q2+q3+1):kk){
if((m_new[j4]-pzeromeanallc_new)<(delta1+delta2)/2+dd)    pvalue[iter]<-0
}
}

nbar<-mean(nvalue)
nvar<-var(nvalue)
nstdev <- nvar^(0.5)/(itersize)^0.5
pbar <- mean(pvalue)
pvar <- var(pvalue)
pstdev <- pvar^(0.5)/(itersize)^0.5

cat("the starting sample size=",n1,"\n")
cat(" cc    nbar   ", " s(nbar)  ",  "  pbar  ",  " s(pbar)  ","\n")
cat(cc,  "   ", nbar,"  ", nstdev, "  ", pbar,"  ", pstdev,"\n")
itersize
}
```

Appendix D

R Source Code for Nonparametric Procedure By Hodges-Lehmann Method

```
LH_normal<-function(cc){
#...This program does the Sequential Procedure by Lehmann-Hodges method
#...you need to enter the optimal sample size to run it
kk <-8
qq <- 4
csize <- 1
itersize <- 1000
bb <- 2.4417695
alpha<-0.75
A_squ<-0.9549
nvalue <- rep(0., itersize)
pvalue <- rep(1., itersize)
delta <- sqrt((((csize+1)/csize)  * (bb^2.))/(cc*A_squ))
 cat("Normal distribution, Lehmann-Hodges method","\n")
 cat(" bb=", bb,  "c=", cc, " ...delta=", delta,"alpha=",alpha,"A_squ=",A_squ ,"\n")
 cat(" No. of Iterations=", itersize,"\n")
 cat(" kk=", kk,"control size=", csize,"\n")

#..........iteration starts...........
 for (iter in 1:itersize){
totalsize <- cc*2
mdat <- matrix(rnorm(kk*totalsize,0,1), nrow = totalsize, ncol=kk, byrow=TRUE)
control <- matrix(rnorm(csize*totalsize,0,1), nrow = totalsize, ncol=csize, byrow=TRUE)
for (j1 in 1:qq) {
mdat[,j1] <- mdat[,j1]-delta}
for (j1 in (qq+1):kk) {
mdat[,j1] <- mdat[,j1]+delta}

#sequential procedure starts........
```

```
for(n in 10:totalsize){
xdat<-matrix(0,nrow=n,ncol=kk,byrow=TRUE)
condat<-matrix(0,nrow=n,ncol=csize,byrow=TRUE)
for(j in 1:kk){
   for(i in 1:n){
xdat[i,j]<-mdat[i,j]}}
for(j in 1:csize){
    for(i in 1:n){
condat[i,j]<-control[i,j]}}

b<-max(1,floor(n*(n+1)/4-qnorm(1-alpha)*sqrt(n*(n+1)*(2*n+1)/24)))
a<-n*(n+1)/2-b+1
ww<-n*(n-1)/2

newdat<-matrix(0,ncol=kk,nrow=ww,byrow=TRUE)
newcon<-matrix(0,ncol=csize,nrow=ww,byrow=TRUE)
upx<- matrix(0,ncol=kk,nrow=ww)
upc<- matrix(0,ncol=csize,nrow=ww)

for(j in 1:kk){
g<-1
   for(i in 1:(n-1)){
      for(d in 1:(n-i)){
      newdat[g,j]<- (xdat[i,j]+xdat[i+d,j])/2
      g<-g+1
      }}}
for(j in 1:csize){
g<-1
   for(i in 1:(n-1)){
      for(d in 1:(n-i)){
      newcon[g,j]<- (condat[i,j]+condat[i+d,j])/2
      g<-g+1
      }}}
for(j in 1:kk){
upx[,j]<-matrix(sort(newdat[,j]))}
for(j in 1:csize){
upc[,j]<-matrix(sort(newcon[,j]))}

SS<-t(upx[a,]-upx[b,])%*%(upx[a,]-upx[b,])+t(upc[a,]-upc[b,])%*%(upc[a,]-upc[b,])
SS<-n*SS/(4*(kk+csize)*((qnorm(1-alpha))^2))
term <- ((csize+1)/csize)*(bb^2)*SS/(delta^2)
if(n>term) break
}

upxdat<-matrix(0,ncol=kk,nrow=n,byrow=TRUE)
upcon<-matrix(0,ncol=csize,nrow=n,byrow=TRUE)
for(j in 1:kk){
upxdat[,j]<-matrix(sort(xdat[,j]))}
for(j in 1:csize){
upcon[,j]<-matrix(sort(condat[,j]))}

md<-rep(0.,kk)
mdc<-rep(0.,csize)

mm<-as.integer(n/2)+1
for(j in 1:kk){
md[j]<-upxdat[mm,j]}
for(j in 1:csize){
mdc[j]<-upcon[mm,j]}
nvalue[iter] <- n

for (j1 in 1:qq) {
if (md[j1]>mdc[1]) pvalue[iter]=0}
for (j1 in (qq+1):kk) {
if (md[j1]<mdc[1]) pvalue[iter]=0}
}
nbar <- mean(nvalue)
nvar <- var(nvalue)
nstdev <- nvar^(0.5)/(itersize)^0.5
pbar <- mean(pvalue)
pvar <- var(pvalue)
pstdev <- pvar^(0.5)/(itersize)^0.5

cat(" cc  delta   nbar   ", "   s(nbar)  ",  "  pbar  ",  "  s(pbar)  ","\n")
cat(cc, "  ", delta, "  ", nbar,"  ", nstdev, "  ", pbar,"  ", pstdev,"\n")
itersize
}
```

Vita

Miss. J. Zhou was born in Jintan, Jiangsu province, China. She received her primary and high school education in Jintan. And she attended undergraduate and master programs in majors of Biomedical Engineering, at Southeast University (SEU), Nanjin, Jiangsu Province, where she studied the drug delivery systems that could kill tumors without affecting the healthy cells. She joined the Master Program of Statistics and Ph.D program of Engineering and Applied Science at Department of Mathematics, University of New Orleans. During her Ph.D program, she worked on generalization of partition problem in statistics. One of the applications of her research is to help doctors and medical researchers in identifying good drugs and compare the newer drugs with other existing drugs.