mathematics

MDPI

Article

A Non-Parametric Sequential Procedure for the Generalized Partition Problem

Tumulesh K. S. Solanky * and Jie Zhou

Department of Mathematics, University of New Orleans, New Orleans, LA 70148, USA; jzhouchen2012@gmail.com
* Correspondence: tsolanky@uno.edu

Abstract: In selection and ranking, the partitioning of treatments by comparing them to a control treatment is an important statistical problem. For over eighty years, this problem has been investigated by a number of researchers via various statistical designs to specify the partitioning criteria and optimal strategies for data collection. Many researchers have proposed designs in order to generalize formulations known at that time. One such generalization adopted the indifference-zone formulation to designate the region between the boundaries for "good" and "bad" treatments as the indifference zone. Since then, this formulation has been adopted by a number of researchers to study various aspects of the partition problem. In this paper, a non-parametric purely sequential procedure is formulated for the partition problem. The "first-order" asymptotic properties of the proposed non-parametric procedure are derived. The performance of the proposed non-parametric procedure for small and moderate sample sizes is studied via Monte Carlo simulations. An example is provided to illustrate the proposed procedure.

Keywords: purely sequential procedure; control population; indifference zone; probability of correct decision; non-parametric distribution; simulations; example

MSC: 62F07; 62L10

Citation: Solanky, T.K.S.; Zhou, J. A Non-Parametric Sequential Procedure for the Generalized Partition Problem. *Mathematics* **2024**, *12*, 591. https://doi.org/10.3390/math12040591

Academic Editor: Stefano Bonnini

Received: 29 December 2023
Revised: 29 January 2024
Accepted: 13 February 2024
Published: 17 February 2024

1. Introduction

The statistical problem of comparing treatments with a control population has been an active area of research for nearly eighty years. One of the earlier research studies that had proposed a formal statistical design to compare treatments with a control is reported in [1]. Soon after this, Ref. [2] investigated this problem for normal means and binomial proportions with an idea of spacing between treatments. Ref. [3] extended this further by exploring the idea of multiple comparisons and formulated a procedure to carry out comparisons with a control population. The idea of spacing was further refined in [4] which formally conceptualized the "indifference zone" formulation for selecting the best normal population from a group of several normally distributed populations in the preference zone with the predetermined probability. In statistical literature, the region outside the indifference zone is referred to as the preference zone. Also in the 1950s, another formulation was proposed for the problem of selecting or isolating the best population in [5], which had the property that it did not restrict the selection from the preference zone but rather the selection was carried out from the entire parameter space. This formulation of the problem, known as the "subset-selection formulation", selects a subset of the populations of random size which includes the best treatment with the prespecified probability. A number of researchers have studied this problem by formulating it under various requirements and goals and while adopting various sampling methodologies. Once such formulation that has been extensively studied in the literature is in which the experimenter wants the selected population to be some "specified amount better" than other treatments, which is referred to as a control or standard. This area of research is typically known as the problem

of "comparisons with a control" or the "partition problem" in statistical literature. For the partition problem formulation, one formulation that has been used by a number of practitioners and researchers is the one introduced in [6] for the populations that follow a normal distribution.

In Section 2, we have summarized the [6] formulation and provided a summary of the current research in the area. In Section 3, we have proposed a distribution-free version of the [6] formulation and proposed a purely sequential methodology and derived its first-order asymptotic properties. In Section 4, we have studied the performance of the proposed non-parametric procedure by picking different values of design constants to study how the asymptotic expansions provided in Theorem 1 compare with the observed values when the procedure is simulated for small and moderate sample sizes. In Section 5, we have provided an example to illustrate an application of the proposed non-parametric purely sequential procedure.

2. Normal Populations Case

Assume that we have $(k+1)$ independently distributed normal populations to be donated as $\pi_0, \pi_1, \cdots, \pi_k$, with respective means $\mu_0, \mu_1, \cdots, \mu_k$ and a common variance σ^2. We will assume that all the parameters are unknown. The population π_0 is referred to as the control or standard population. The formulation presented in [6] starts by mathematically defining the "good" and "bad" populations based on the input from practitioners or experts in the area of the application.

Next, for fixed but arbitrary constants, δ_1 and δ_2, with $\delta_2 > \delta_1$, ref. [6] defined the "good" and "bad" populations via three sets by adopting the [4] indifference zone formulation, as defined below

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

The set Ω_G is termed to as the set of "good" populations while the set Ω_B is termed as the set of "bad" populations. Note that the two constants δ_1 and δ_2 are determined based on the input of experts in the area specifying how much better or worse a population has to be compared to the control to be termed as a good population or a bad population. The goal in [6] was to partition the populations that belong to Ω_G or Ω_B correctly with the prespecified probability. On the other hand, the set Ω_I is termed as the indifference-zone set, and the experimenter is indifferent to the correct partition of the populations that fall in the set Ω_I. The parition problem is designed to partition the set $\Omega = \{\pi_i, i = 1, \cdots, k\}$ into two mutually disjoint sets S_B and S_G, with high accuracy, so that all populations in Ω_B fall inside S_B and all populations in Ω_G fall inside S_G. That is, when all the populations in Ω_B or Ω_G are partitioned correctly, then such a partition is defined as a correct decision (CD). Mathematically, let us denote by P^* the probability of correct decision that the experimenter wants to achieve. Note that $\frac{1}{2^k} < P^* < 1$, as the probability of selecting correctly randomly is $\frac{1}{2}$ for each of the k populations.

Next, using a sampling design, determine N as the sample size from each of the k populations and the control population and the sample mean $\bar{X}_{iN}$ from $\pi_i, i = 0, 1, \cdots, k$. Define $d = (\delta_1 + \delta_2)/2$; then, the decision rule proposed by [6] to partition all the populations in Ω took the following form:

$$\begin{array}{rcl} S_B & = & \{\pi_i : \bar{X}_{iN} - \bar{X}_{0N} \leq\ d,\ i = 1, \cdots, k\}, \\ S_G & = & \{\pi_i : \bar{X}_{iN} - \bar{X}_{0N} \geq\ d,\ i = 1, \cdots, k\}. \end{array} \tag{2}$$

Ref. [6] has shown that if the sample size N satisfies $N \geqslant \frac{2\sigma^2}{a^2 b^2}$, and we partition the k populations according to the partition rule (2), then

$$P[CD] \geqslant P^*, \qquad \forall\ \boldsymbol{\mu} \in \boldsymbol{R}^{k+1},\ \sigma \in \boldsymbol{R}^+. \tag{3}$$

Note that $a = (\delta_2 - \delta_1)/2$, $l = k/2$ when k is even and $l = (k+1)/2$ when k is odd, and the $k \times k$ matrix covariance matrix $\Sigma = (\sigma_{ij})$ is a given by

$$\sigma_{ij} = \begin{cases} 1 & \text{when } i = j, \\ 1/2 & \text{when } i \neq j \text{ and } 0 < i,\ j \leq l \text{ or } l < i,\ j \leq k, \\ -1/2 & \text{when } 0 < i \leq l \text{ and } l < j \leq k, \end{cases}$$

and b is a constant satisfying the integral equation given by

$$P^* = \int_{-\infty}^{b} \cdots \int_{-\infty}^{b} \frac{|\Sigma|^{\frac{1}{2}}}{(2\pi)^{\frac{k}{2}}} \exp\left(-y'\Sigma^{-1}y/2\right) dy_1 \cdots dy_k. \tag{4}$$

Ref. [6] has tabulated the values of design constant b for various choices of k and P^*. For the unknown σ^2 case, ref. [6] also constructed a two-stage and a purely sequential procedure.

For the normal distributions case, ref. [7] constructed several multistage methodologies focusing on the second-order asymptotic expansions. For references on the partition problem for binomial treatments, the reader is referred to [8]. In [9], a generalization of the "Tongs formulation" was introduced so that the treatments that fall between the "good" and "bad" treatments can be partitioned as a separately identifiable group by introducing two indifference zones. Ref. [10] extended this generalization by constructing an asymptotically unbiased fine-tuned purely sequential procedure to guarantee the probability requirement.

Next, we have constructed a non-parametric procedure to partition the k populations compared to a control population that does not require the populations to be normally distributed. However, we have assumed that the unknown distributions are symmetric. Next, in Section 3, we have proposed a distribution-free version of the [6] formulation, proposed a purely sequential methodology and derived its first-order asymptotic properties.

3. Non-Parametric Partition Problem

Assume that we are given $(k+1)$ independent populations $\pi_0, \pi_1, \pi_2, \cdots, \pi_k$, where the control population is denoted as π_0 . Assume that the cumulative distribution function (cdf) of π_i is $F(x - \Delta_i)$ for $i = 0, 1, \cdots, k$. We will assume the cdf $F(.)$ is continuous and symmetric. Note that the function $F(.)$ and all the centers of symmetries, namely, Δ_0, $\Delta_1, \cdots, \Delta_k$ are assumed to be unknown. Following [6], we have defined below what an experimenter may define as "good" and "bad" populations compared to a control based on the input from experts in the area of application. As in Section 2 for the normal populations, we will partition all k populations by comparing the centers of symmetry Δ_i, $i = 1, \cdots, k$ with the control population's center of symmetry Δ_0 to define the set of "good" and "bad" populations which has the probability of correct decision (CD) of at least P^*. As before, $\frac{1}{2^k} < P^* < 1$.

Based on the input from experts in the area, the statistical design would start by selecting two arbitrary but fixed design constants, δ_1 and δ_2, with $\delta_2 > \delta_1$. Next, as in [6], we define three subsets for $\Omega = \{\pi_1, \cdots, \pi_k\}$ following the idea of spacing from [4] the indifference-zone formulation as follows:

$$\begin{aligned} \Omega_L &= \{\pi_i : \Delta_i \leq \Delta_0 + \delta_1,\ i = 1, \cdots, k\}, \\ \Omega_R &= \{\pi_i : \Delta_i \geq \Delta_0 + \delta_2,\ i = 1, \cdots, k\}, \\ \Omega_I &= \{\pi_i : \Delta_0 + \delta_1 < \Delta_i < \mu_0 + \delta_2,\ i = 1, \cdots, k\}. \end{aligned} \tag{5}$$

Note that Ω_R and Ω_L are the sets of "good" populations and "bad" populations, respectively, whereas Ω_I is the set of populations the experimenter would be indifferent to. We define two constants based on δ_1 and δ_2 as $d = (\delta_1 + \delta_2)/2$ and $\delta^* = (\delta_2 - \delta_1)/2$. Let Λ denote a class of symmetric and continuous distributions which satisfy some regularity conditions to be specified in Section 4. Next, we propose a purely sequential procedure for the partition problem described in (5). The procedure starts with an initial sample size of

$m(\geq 2)$ observations from all the $(k+1)$ populations. Next, implementing the "vector-at-a-time" sampling procedure, we will sample one observation from all the $(k+1)$ populations according to the stopping rule defined below in (7). Having recorded an independent sample $X_{i1}, X_{i2}, \cdots, X_{in}$, a sample of size n from π_i, $i = 0, 1, \cdots, k$, a statistic $L_i(n)$, to be defined below, is proposed to estimate the center of symmetry Δ_i, $i = 0, 1, \cdots, k$. The estimator $L_i(n)$ has an asymptotic normal distribution. That is, $N\left(\Delta_i, 1/\left(nA^2\right)\right)$, as $n \to \infty$ for $i = 1, \cdots, k$, $F(.) \in \Lambda$. Note that the unknown constant A is a finite and positive function of F. For the literature of non-parametric procedures in the area of selecting the best population, the reader is refereed to [11]. One may also refer to [12] who had constructed a non-parametric accelerated sequential procedure to select the population with the largest center of symmetry.

Based on a sample of size n, the decision rule is to compare each $L_i(n)$ with $L_0(n)$, $i = 1, \cdots, k$, and then partition the k populations following the partition rule given by:

$$\begin{aligned} P_L &= \{\pi_i : L_i(n) - L_0(n) < d, i = 1, \cdots, k\} \\ P_R &= \{\pi_i : L_i(n) - L_0(n) \geq d, i = 1, \cdots, k\}, \end{aligned} \tag{6}$$

Next, as in [11], we will assume that the following regularity conditions are satisfied by the unknown distribution $F(.)$ and the purely sequential stopping rule, which is implemented to obtain the sample size N:

Regularity Conditions: We will assume the following three conditions hold 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 a finite second moment and $0 < A = A(F) < \infty$.
2. For an estimator S_n^2 of A, as $n \to \infty$, we have $\lim S_n^2 = A^{-2}$ a.s.
3. The set $\left\{\delta^2 N(\delta) : \delta > 0\right\}$ is uniformly integral.

Next, following [7], one can obtain that $P(CD)$ is asymptotically at least P^* if the sample size n is at least $\frac{2b^2}{(A\delta^*)^2}$. Here, "$b$" is a constant, as reported earlier, which is a function of k and P^*. Let us denote $n^* = \frac{2b^2}{(A\delta^*)^2}$. The expression n^* is known as the optimal sample size. However, it is unknown as A is unknown. Next, to estimate A, a purely sequential procedure is constructed which satisfies the correct decision probability requirement and has $\liminf P(CD) \geq P^*$ whenever $\theta \in \omega(\delta^*)$ and the unknown cdf $F(.) \in \Lambda$, as $\delta^* \to 0$. The purely sequential procedure starts with m observations from each population, and it samples one observation from all $(k+1)$ according to the stopping rule:

$$N = \inf\{n \geq m : n \geq \tfrac{2b^2 S_n^2}{\delta^{*2}}\} \tag{7}$$

where S_n^2, an estimator of A, is computed using the control and all k populations. Also, S_n^2 depends on the estimator of the center of symmetry $\Delta_i, i = 0, 1, \cdots, k$. Next, we present a theorem to the first-order properties of the proposed purely sequential procedure (7).

Theorem 1. *The purely sequential procedure defined in* (7)*, under the assumptions as outlined above, satisfies the following properties for all* $F(.) \in \Lambda$ *and* $\omega(\delta^*) \in \Omega$*:*

(i) $N(\delta^*) \to \infty$ *monotonically as* $\delta^* \to 0$ *a.s.*
(ii) $E(N(\delta^*)) \to \infty$ *as* $\delta^* \to 0$.
(iii) $\lim \delta^{*2} N(\delta^*) = 2b^2/A^2$ *a.s.*
(iv) $\liminf P(CD) \geq P^*$ *as* $\delta^* \to 0$.

Proof. We start with an estimator S_n^2 for the center of symmetry. Based on a sample of size n, let $L_i(n)$ denote the Hodges–Lehmann estimator for the center of symmetry Δ_i of the ith population $i = 0, 1, \cdots, k$. That is, the sample median of the $n(n+1)/2$ quantiles

$\left(X_{ij}+X_{il}\right)/2$ for $j \le l$, $j, l = 1, \cdots, n$; $i = 0, 1, \cdots, k$. Then, we consider the estimator of A^{-2} is given by

$$S_n^2 = \frac{n\left((k+1)K_\alpha^2\right)^{-1}}{4} \sum_{i=0}^{k} \left(W_{n,a(n)}(i) - W_{n,b(n)}(i)\right)^2, \tag{8}$$

where $W_{n,1}(i) \le W_{n,2}(i) \le \cdots \le W_{n,n(n+1)/2}(i)$ are the ordered $\left(X_{ij}+X_{il}\right)/2$ for $1 \le j \le l \le n$ and for $i = 0, 1, \cdots, k$. The sequence $\{a(n)\}$ and $\{b(n)\}$ are specified as

$$\begin{aligned} b(n) &= \max\left\{1, \left[n(n+1)/4 - K_\alpha\left(n(n+1)(2n+1)/24\right)^{\frac{1}{2}}\right]\right\} \\ a(n) &= n(n+1)/2 - b(n) + 1. \end{aligned} \tag{9}$$

where $[x]$ is defined as the largest integer less than or equal to x. K_α is defined by $\phi(K_\alpha) = 1-\alpha$ for some $1/2 < \alpha < 1$. The Hodges–Lehmann estimator has been used extensively in statistical literature, and it is well known that $L_i(n)$ is a consistent estimator of the center of symmetry. The reader is referred to [13] for details.

Next, note that $N(\delta_1^*) \ge N(\delta_2^*)$ w.p. 1 if $0 < \delta_1^* < \delta_2^*$, that is $N(\delta^*)$ is non-decreasing in δ^*. Now, the assumption 1.1 [13] in regularity conditions will lead to part (i). Part (ii) follows by applying the monotone convergence theorem. Since the stopping rule is

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

then the basic inequality simplifies to

$$2b^2 S_n^2 \Big/ \delta^{*2} \le N \le m_0 + 2b^2 S_{n-1}^2 \Big/ \delta^{*2}. \tag{10}$$

Now, multiply δ^{*2} throughout (10) 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

$$\begin{aligned} & P(CD|\theta \in \Omega(\delta^*)) \\ = \; & P\left\{L_i(N) - L_0(N) < d, 0 < i \le r; L_j(N) - L_0(N) \ge d, r < j \le k\right\} \\ = \; & P\Bigg\{\left((L_i(N)-\Delta_i) - (L_0(N)-\Delta_0)\right)\frac{\sqrt{n^*}A}{\sqrt{2}} < \left(d - (\Delta_i - \Delta_0)\right)\frac{\sqrt{n^*}A}{\sqrt{2}}, 0 < i \le r; \\ & \left(\left(L_j(N)-\Delta_j\right) - (L_0(N)-\Delta_0)\right)\frac{\sqrt{n^*}A}{\sqrt{2}} \ge \left(d - \left(\Delta_j - \Delta_0\right)\right)\frac{\sqrt{n^*}A}{\sqrt{2}}, r < j \le k\Bigg\} \\ = \; & P\left\{\frac{Z_i - Z_0}{\sqrt{2}} < \frac{\sqrt{n^*}A\delta^*}{\sqrt{2}}, 0 < i \le r; \frac{Z_j - Z_0}{\sqrt{2}} \ge -\frac{\sqrt{n^*}A\delta^*}{\sqrt{2}}, r < j \le k\right\} \\ = \; & P\left\{Y_i(N) \le \frac{\sqrt{n^*}A\delta^*}{\sqrt{2}}, i = 1, \cdots, k\right\}. \end{aligned} \tag{11}$$

where

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

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

$$P(CD|\theta \in \Omega(\delta^*)) = \int_{-\infty}^{\frac{\sqrt{n^*}A\delta^*}{\sqrt{2}}} \cdots \int_{-\infty}^{\frac{\sqrt{n^*}A\delta^*}{\sqrt{2}}} (2\pi)^{-\frac{k}{2}} |\Sigma_r|^{-\frac{k}{2}} \exp\left(-\frac{1}{2} y' \Sigma_r^{-1} y\right) \prod_{i=1}^{k} dy_i. \quad (12)$$

Equation (12) gives the infimum of the $P(CD)$ 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 right side of (12) achieves a minimum over all $r(0 < r \leq k)$ under the LFC. Let $b = b(P, k)$ be the solution of the equation

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

Also, for any real number c and q, let

$$P_q(c) = \int_{-\infty}^{c} \int_{-\infty}^{c} \cdots \int_{-\infty}^{c} (2\pi)^{-\frac{q}{2}} |\Sigma_q|^{-\frac{q}{2}} \exp\left(-\frac{1}{2} y' \Sigma_q^{-1} y\right) \prod_{i=1}^{q} dy_i \quad (13)$$

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

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

then

$$P_r(b) + P_{k-r}(b) = 1 + P^*$$

which leads to

$$P(A \cap B) = P\{Y_i(N) \leq b, i = 1, \cdots, k\} = P(CD|\theta \in \Omega(\delta^*)) \geq P^*$$

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

4. Monte Carlo Simulation Results

In this section, using the Monte Carlo simulation study, the "purely sequential procedure" (7) is replicated independently 5000 times by picking different values of design constants to study how the asymptotic expansions provided in Theorem 1 compare with the observed values when the procedure is simulated for small and moderate sample sizes. In our simulation study, we considered $k = 8$ independent populations and one control population. To construct the LFC, we generated $four$ populations with the center of symmetry equal to $\mu_0 - \delta$, and the remaining $four$ 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 $P^* = 0.95$, the value of the constant b equals 2.44177 from [6]. Next, we considered the following symmetric distribu-

tions: normal distribution, Laplace distribution, t-distribution, uniform distribution, and a mixture of two normal distributions. For these distributions, the parameter A^2 is given by

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

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

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

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

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

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

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

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

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 Tables 1–6. As described earlier, the estimator S_n^2 as described in (8) 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 $\lambda = \sqrt{2}\big/2$, t-distribution with $df = 5$, uniform distribution, and two mixed normal distributions 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. Note that 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 Tables 1–6.

From Tables 1 and 2, note that the purely sequential procedure (7) is oversampling by roughly two to three observations 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 0.95 for the normal case. However, for the Laplace distribution, the estimated probability of correct selection matches the target value of 0.95 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 are designed for normally distributed populations, such as the ones described in [7]. However, if the tails are significantly longer

than the normal tails, like for the Laplace distribution, then the non-parametric partition procedures are more appropriate.

Table 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
0.75	0.353	100	102.298	0.189	0.870	0.011
0.75	0.250	200	202.597	0.263	0.870	0.011
0.75	0.177	400	402.507	0.376	0.877	0.010
0.75	0.125	800	803.636	0.492	0.847	0.011
0.85	0.499	50	52.958	0.122	0.865	0.011
0.85	0.353	100	103.046	0.180	0.865	0.011
0.85	0.250	200	203.638	0.255	0.855	0.011
0.85	0.177	400	403.382	0.365	0.857	0.011

Table 2. Simulation results for Laplace distribution with $\lambda = \frac{\sqrt{2}}{2}$.

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

Table 3. Simulation results for T-distribution with $df = 5$.

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

In Table 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 0.95 for smaller values of α. However, as α increases, the estimated probability of correct selection is approaching the target value of 0.95.

Table 4. Simulation results for uniform distribution.

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

Table 5. Simulation results for mixture of two normal distributions: $X = 0.35N(x_1; 0, 1) + 0.65N(x_2; 0, 2)$.

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

Table 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
0.75	0.480	100	104.593	0.259	0.935	0.008
0.75	0.339	200	205.031	0.351	0.932	0.008
0.85	0.678	50	55.826	0.169	0.934	0.008
0.85	0.450	100	106.334	0.254	0.926	0.008
0.85	0.339	200	206.534	0.332	0.933	0.008
0.85	0.240	400	406.497	0.486	0.942	0.007
0.90	0.678	50	56.762	0.177	0.955	0.007
0.90	0.480	100	106.746	0.244	0.924	0.008
0.90	0.339	200	207.888	0.351	0.935	0.008
0.95	0.678	50	58.671	0.173	0.959	0.006
0.95	0.480	100	108.742	0.235	0.947	0.007
0.95	0.339	200	208.019	0.332	0.931	0.008

Next, we have considered the uniform distribution case which has tails even shorter than the normal tails. One will note that the estimated probability of correct selection is well below the target value of 0.95. 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 $0.35N(x_1;0,1)+0.65N(x_2;0,2)$ which is a mixture of two normal populations with somewhat long tails. The first population is the mixture that has a variance of 1, and the second has a variance of 2. In the second mixture of the two normal populations considered, we have $0.8N(x_1;0,1)+0.2N(x_2;0,5)$. 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 the 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.

5. An Example

In this section, we study the performance of the non-parametric sequential procedure via a real-world dataset. Ref. [14] conducted a pilot investigation to see if active exercise can preserve walking beyond the 2nd 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). A traditional 12 months has been known as the mean time infants take to walk. The statistical analysis confirmed that the walking data are normally distributed with somewhat equal variance, adopting a 12.5% improvement as significant and anything other than 8% as not significant. We took $\delta_1 = -1.5$ months, $\delta_2 = -1.0$ months, $k = 3$, and the starting sample size $m = 5$. The data were analyzed via the following three procedures: (1) two-stage procedure of [6]; (2) purely sequential procedure of [7]; (3) non-parametric sequential procedure proposed in this manuscript. Additional samples as needed were generated via SRSWR and saved to have the same data for all the procedures. Note that all the three sampling methodologies yielded the same result: that is, the active exercise group was partitioned as better than the control, while the passive and no exercise groups 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 Table 7. One will note that the sample size was somewhat larger for the non-parametric sequential procedure, and it increased further when the parameter α was increased. However, this was quite expected, since the data are 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 are not normal and have long tails.

Table 7. Comparison of various statistical methodologies.

Procedure	Sample Size
Two-stage	71
Purely Sequential	66
Non-Parametric Sequential	42 ($\alpha = 0.75$) 52 ($\alpha = 0.80$) 53 ($\alpha = 0.85$) 60 ($\alpha = 0.90$) 67 ($\alpha = 0.95$)

Author Contributions: Conceptualization, T.K.S.S. and J.Z.; formal analysis, T.K.S.S. and J.Z.; methodology, T.K.S.S. and J.Z.; writing—review and editing, T.K.S.S. and J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article.

Acknowledgments: The authors would like to thank the editor and two referees for their invaluable feedback.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Roessler, E.B. Testing the significance of observations compared with a control. *Proc. Am. Soc. Hortic. Sci.* **1946**, *47*, 249–251.
2. Paulson, E. On the comparison of several experimental categories with a control. *Ann. Math. Stat.* **1952**, *23*, 239–246. [CrossRef]
3. Dunnett, C.W. A multiple comparison procedure for comparing several treatments with a control. *J. Am. Stat. Assoc.* **1955**, *50*, 1096–1121. [CrossRef]
4. Bechhofer, R.E. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Ann. Math. Stat.* **1954**, *25*, 16–39. [CrossRef]
5. Gupta, S.S. On a Decision Rule for a Problem in Ranking Means. Ph.D. Thesis, University of North Carolina, Chapel Hill, NC, USA, 1956.
6. Tong, Y.L. On partitioning a set of normal populations by their locations with respect to a control. *Ann. Math. Stat.* **1969**, *40*, 1300–1324. [CrossRef]
7. Datta, S.; Mukhopadhyay, N. Second-order asymptotics for multistage methodologies in partitioning a set of normal populations having a common unknown variance. *Stat. Decis.* **1998**, *16*, 191–205. [CrossRef]
8. Buzaianu, E.M. Selection among Bernoulli populations in comparison with a standard. *Seq. Anal.* **2019**, *38*, 184–198. [CrossRef]
9. Solanky, T.K.S.; Zhou, J. A generalization of the partition problem. *Seq. Anal.* **2015**, *34*, 483–503. [CrossRef]
10. Solanky, T.K.S. Second Order Asymptotics of a Fine-Tuned Purely Sequential Procedure for the Generalized Partition Procedure. *Stat. Appl.* **2021**, *19*, 401–415.
11. Geertsema, J.C. Nonparametric Sequential Procedures for Selecting the Best of K Populations. *J. Am. Stat. Assoc.* **1972**, *67*, 614–616. [CrossRef]
12. Mukhopadhyay, N.; Solanky, T.K.S. A nonparametric accelerated sequential procedure for selecting the largest center of symmetry. *Nonparametric Stat.* **1993**, *3*, 155–166. [CrossRef]
13. Hodges, J.L.; Lehmann, E.L. Estimation of location based on ranks. *Ann. Math. Stat.* **1963**, *34*, 598–611. [CrossRef]
14. Zelazo, P.R.; Zelazo, N.A.; Kolb, S. "Walking" in the Newborn. *Science* **1972**, *176*, 314–315. [CrossRef]