AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

ISSN: 1082-989X

2025, Vol. 30, No. 2, 425–439
https://doi.org/10.1037/met0000570

A Sequential Approach for Noninferiority or Equivalence of a Linear Contrast Under Cost Constraints

Bhargab Chattopadhyay[1], Tathagata Bandyopadhyay[2], Ken Kelley[3], and Jishnu J. Padalunkal[4]
[1] School of Management & Entrepreneurship, Indian Institute of Technology Jodhpur
[2] Dhirubhai Ambani Institute of Information and Communication Technology
[3] Mendoza College of Business, University of Notre Dame
[4] Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas

Abstract

Planning an appropriate sample size for a study involves considering several issues. Two important considerations are cost constraints and variability inherent in the population from which data will be sampled. Methodologists have developed sample size planning methods for two or more populations when testing for equivalence or noninferiority/superiority for a linear contrast of population means. Additionally, cost constraints and variance heterogeneity among populations have also been considered. We extend these methods by developing a theory for sequential procedures for testing the equivalence or noninferiority/superiority for a linear contrast of population means under cost constraints, which we prove to effectively utilize the allocated resources. Our method, due to the sequential framework, does not require prespecified values of unknown population variance(s), something that is historically an impediment to designing studies. Importantly, our method does not require an assumption of a specific type of distribution of the data in the relevant population from which the observations are sampled, as we make our developments in a data distribution-free context. We provide an illustrative example to show how the implementation of the proposed approach can be useful in applied research.

Translational Abstract

Two important considerations when designing a study are cost constraints and variability inherent in the population from which data will be sampled. Sample size planning methods for two or more populations when testing for equivalence or noninferiority/superiority for a linear contrast of population means have developed. Additionally, cost constraints and variance heterogeneity among populations have also been considered. We extend these methods by developing a theory for sequential procedures for testing the equivalence or noninferiority/superiority for a linear contrast of population means under cost constraints. Our method, due to the sequential framework, does not require prespecified values of unknown population variance(s), something that is historically an impediment to designing studies. Importantly, our method does not require an assumption of a specific type of distribution of the data in the relevant population from which the observations are sampled, as we make our developments in a data distribution-free context. We provide an illustrative example to show how the implementation of the proposed approach can be useful in applied research.

Keywords: sample size planning, study design, research design, power, sequential analysis, stopping rule

Sample size planning is a key aspect of designing a study. It should therefore be seriously considered before starting any study. Suppose there is some ideal sample size in order to accomplish one or more goals of interest with a specified probability for a particular design in a particular context. Although the ideal sample size is generally unknowable in practice, choosing a smaller sample size might be economically beneficial (e.g., the study can be done more quickly and be less expensive to conduct), the results will have a smaller probability of accomplishing the desired goal(s). For example, consider power analysis, in which one plans the sample size in order to have some specified probability of rejecting a false null hypothesis. When using a smaller sample size than the ideal sample size for some desired value of statistical power, the statistical test will have less statistical power than the desired value. One problem with statistical power being less than the desired level, such as using a smaller sample size than the power analysis suggests based on the specifications provided, is having a smaller probability of rejecting the null hypothesis than is desired. An implication of this is that, in such situations, underpowered studies can contribute to the replication problems of a discipline (e.g., Anderson & Kelley, 2022; Button et al., 2013; Dumas-Mallet et al., 2017). For attaining higher power, larger sample sizes are needed, which generally increases study costs and often lengthens the time to complete data collection. Additionally, a larger

This article was published Online First May 11, 2023.
Ken Kelley https://orcid.org/0000-0002-4756-8360
Jishnu J. Padalunkal contributed toward simulation coding, tables and Figure 1 creation. Bhargab Chattopadhyay, Tathagata Bandyopadhyay, and Ken Kelley contributed toward conceptualization, methodology development, and paper writing.
Correspondence concerning this article should be addressed to Ken Kelley, 234 Mendoza College of Business, University of Notre Dame, Notre Dame, IN 46556, United States. Email: kkelley@nd.edu

sample size is required to achieve greater power if the population variability is larger, holding everything else constant. Additionally, for a research study with two or more populations, heterogeneous variances are a common phenomenon (e.g., Keselman et al., 1998). Therefore, the required sample size for sampling from the populations will depend on the variability and also on the sampling cost.

The idea of a cost constraint is to make clear that the sample size is not unconstrained, and it does this by factoring in how much each participant is included in the sample costs. This is a useful approach because most real situations do involve a cost constraint, where studies cannot be unconstrained in how much they cost to conduct in practice. Luh & Guo (2016) developed a procedure to plan the optimal sample size required to obtain the maximum power under cost constraints. Their procedure was based on knowing the population values of the population variances under the assumption of normality, which will not always be applicable. In order to address the issue, Luh & Guo (2016) suggested using a pilot study in order to find the optimal sample sizes. However, to do the pilot study, some amount of money needs to be allocated separately, an issue that was not considered. This article addresses these issues and extends their procedure by framing the problem in terms of a sequential procedure. The study provides a sample size plan for the tests of equivalence or noninferiority/superiority of a linear contrast under cost constraints without assuming a particular distribution of the data and values of the population variances. As Chattopadhyay & Kelley (2016, 2017) discussed in the context of psychological, educational, managerial, and social research, there are many advantages to a sequential framework, as compared to using supposed population parameter values, in order to plan the sample size of a study prior to the start of the study (and before reasonable parameter estimates may be known). In fact, sequential procedures essentially solve the long-frustrating problem—by avoiding it—of prespecifying generally unknowable population parameters.

This article discusses the sample size estimation procedure in a sequential analysis framework. In sequential analysis, sample sizes are not specified in advance, rather, they are estimated in stages. At each stage, a decision is made as to whether to stop or continue sampling using a predetermined stopping rule. The stopping rule is based on the estimates of the parameters of interest obtained at each stage. At the first stage (often called the pilot stage), a small sample is selected and obtained, and then the decision to stop or continue sampling is made using the particular stopping rule. If the stopping rule is not met, an additional sample of a particular size (as dictated by the method, and using estimates from the first stage) is selected in the second stage. This process continues until the stage is reached when the stopping rule is met for the first time. It is literally a "learn as you go" approach and is known to prevent under- or oversampling to accomplish the prespecified goals. Consequently, unlike the fixed sample size procedures, this procedure does not have a set sample size before the sampling begins, which is a positive (in that sampling too few observations does not occur) but also a potential negative (in that the required sample size may be larger than what is feasible). This limitation means that the sample size that needs to be collected may be more than the researcher can obtain (e.g., due to cost considerations, time, and limited population from which to sample). However, this method avoids using unreasonable supposed values of the unknowable (e.g., when the population variance must be specified) that are required in many situations. Nevertheless, the researcher collects additional observation(s) from the population of interest while sampling sequentially and updates the parameter estimates until the stopping rule is satisfied. We view sequential sampling in a similar way as computerized adaptive testing (CAT), in which items are sampled until the standard error of measurement is sufficiently small. This linking of sequential methods for designing studies and CAT helps to illustrate that the method is not so foreign in psychology and related fields.

For details about how sequential methods can be implemented in psychology and related areas, see Chattopadhyay & Kelley (2016, 2017) and Kelley et al. (2018, 2019), and the references contained therein. For the theory of sequential estimation procedures in more general contexts, see Chattopadhyay & Banerjee (2021), Chattopadhyay (2020), and others.

Tests for Noninferiority and Equivalence: An Application

Applied psychologists, human resource professionals, and managers in work settings often consider issues of training and onboarding, which are very important in many organizations. Consider the standard training program at a large organization in which new employees travel to a central (e.g., corporate) site for face-to-face training in a one-on-one fashion with a trainer. However, face-to-face interaction cannot always occur and an alternative approach proposed is to use one-on-one video calls (e.g., Zoom, Microsoft Teams, and BlueJeans). In such an approach, each trainee can remain in his or her home and the trainer works through the materials in an analogous way to what would be done in face-to-face training. Yet another alternative approach of interest is the use of a prerecorded training video, along with workbooks and reference cards, in which approximately 75% of what was historically in-person face-to-face material is now asynchronous training via the video workbooks so that trainees can work at their own pace and go back-and-forth as needed. After successfully completing the material asynchronously, only then is there a face-to-face video call for the remaining part of training (e.g., role-playing various scenarios).

In essence, the scenario described here is similar to what one could imagine in a wide variety of organizations. More importantly, though, this scenario can be abstracted to many situations. For example, it is applicable to situations where there are often individual conversations, such as individualized instructional methods in schools, therapy sessions, consultation with physicians, academic advising, supervisory coaching, etc. Indeed, during the COVID-19 pandemic, universities grappled with the effectiveness of quickly prepared online instruction versus traditional face-to-face instruction, and the implications for how prompt the return to in-person instruction should be. Similarly, questions arose involving the differences—if any—of taking high-stakes tests (e.g., GRE, SAT, GMAT, and MCAT) in a testing center versus at home using the internet with monitoring from a third party vendor (i.e., with the aim to have more test integrity). Another important consideration in this context is face-to-face therapy sessions with a counselor in person versus online only (i.e., telehealth and teletherapy) versus hybrid/mixed face-to-face sessions interspersed with online teletherapy sessions. All of these examples illustrate alternative approaches to the delivery platform, and we seek to develop a method to evaluate the effectiveness of these different approaches and to answer the question, "Does the approach matter?"

In other words, suppose that the mean score of the face-to-face method is truly higher than the mean score of the online and hybrid/mixed approach. However, the question that we would like to address here is whether the effectiveness of these alternative methods may be considered reasonably close to the face-to-face method, as measured in terms of the group mean, such that there is little to no practical or managerial significance. Suppose μ_1 represents the population mean score from the face-to-face training session, while μ_2 and μ_3, respectively, represent the population means from the online and hybrid approaches, respectively. Let μ_1 be greater than or equal to both μ_2 and μ_3. Thus the effect size of interest, termed ψ, is defined as

$$\psi = -1 \times \mu_1 + (.5) \times \mu_2 + (.5) \times \mu_3. \tag{1}$$

It would be useful to determine if the value of μ_1 is close enough to the mean of μ_2 and μ_3, that is, ψ is reasonably close to 0. This implies that the mean of the alternative methods can be considered about as effective as the face-to-face method, though not literally equal to or better. Therefore, in such situations, we are not evaluating a theoretical argument about which approach is the best but rather a practical argument asking if the methods are close enough that the acknowledged effectiveness of one method over the others does not matter so much that the decision-maker necessarily needs to shift methods to benefit in a substantively meaningful way. Also, if the size of the contrast is more than some prespecified (and thus fixed) negative value, we could say that none of the alternative methods are inferior to the face-to-face method, which, as it turns out is the most expensive approach. This negative value is to be prespecified by the researcher depending on what is considered a tolerable negative size of ψ. If it is less than the prespecified negative value, then at least one of the alternative methods is inferior to the face-to-face method. In that case, one needs to consider a pairwise comparison of each of these alternative methods with the face-to-face method.

The contrast of interest can be estimated by

$$\hat{\psi}_n = -1 \times \bar{X}_1 + (.5) \times \bar{X}_2 + (.5) \times \bar{X}_3. \tag{2}$$

This contrast may be considered from multiple perspectives. One could ask, for example, "What sample size is necessary for a sufficiently high probability of correctly rejecting a false null hypothesis?" (i.e., the statistical power approach to sample size planning) or "What sample size is necessary for a sufficiently narrow confidence interval?" (i.e., the accuracy in parameter approach to sample size planning). However, other questions can be of interest in this context, such as "Is the size of the contrast small enough that the effectiveness of the treatments can be regarded as equivalent" or "Is neither of the two alternative treatments any worse than the standard treatment (i.e., business as usual)?"

To demonstrate how these two tests would be used in practice, the next two subsections work through the process. We emphasize that one or the other method is used for a particular study, not both. But, for illustrative purposes, we show an example of each of the two methods.

Test for Equivalence

Our question is to assess if alternative training protocols are indistinguishable from a practical perspective (i.e., that the effect is no greater than $|\psi|$). The contrast of primary interest here is the population mean score corresponding to the face-to-face format as compared to the average of the mean scores corresponding to the virtual and mixed formats. If the contrast of interest is not distinguishable at the specified level with an effect size of ψ, then the effectiveness of the standard training protocol can be regarded as practically *equivalent* to the mean of the other two treatment options. Thus, the organization could shift to a more virtual training and video followed by a one-on-one video call. If evidence exists that there is little to no difference in these approaches, an organization can save money by not requiring new hires to travel for onsite one-on-one, face-to-face training (and all the expenses it incurs).

The question we consider here, and which our method solves, is "Is there an effect no larger than $|\psi|$ for the contrast of interest?" In particular, for our example, is the absolute difference between the population average score corresponding to the face-to-face format and the average of the population mean scores corresponding to the virtual and the mixed-formats no larger than 1.25 (say)? Note that the 1.25 value used here is based on situational factors that would be context-specific, and we use this value here for illustrative purposes.

Test for Noninferiority

Another way of conceptualizing a similar question that can be considered is about noninferiority. This question is not asking if the contrast of primary interest is *indistinguishable* from a practical perspective, but rather, if the mean of the alternative methods is *not worse than* (i.e., not inferior to) the "business as usual" approach. It may seem that "equivalence" versus "not inferior" is asking the same question, but that is not the case upon careful consideration. The noninferiority framework allows the alternative methods to be *better* (and thus not evaluating if the groups have equal effectiveness), not only addressing if the methods are the same (i.e., equivalence, as discussed above). For example, a noninferior treatment can, in fact, show that the alternative methods can be better. This approach allows a treatment that is cheaper to administer that is not worse, but also allows for the treatment to be better. A better treatment satisfies a test of noninferiority.

It is these two tests that this article addresses. Goeman et al. (2010) discuss testing for equivalence, superiority, and inferiority in the context of clinical trials. Guo et al. (2011) discuss the extension of these approaches to research in psychology, policymaking, and so on. Although the use of such techniques is not very common in social sciences at present, the framework offers opportunities for better addressing certain questions than does null hypothesis significance testing (NHST) alone. In the next section, we formulate the two hypothesis tests.

Formulation: Tests for Equivalence and Noninferiority

Suppose that there are K populations, and $\bar{X}_i$ represents the sample mean corresponding to the ith population where, $X_{i1}, \ldots, X_{in_i}$ represent a random sample of size n_i from the ith population with unknown mean μ_i and unknown variance σ_i^2 ($i = 1, 2, \ldots, K$). Here, we consider a single-factor between-subjects unbalanced design (for more details, we refer to example 6 in Kelley et al., 2018). It is important to notice that we are not making any assumption about the form of the population distributions except that its mean and variance exist.

Suppose we are interested in drawing inference about a contrast, say ψ, of the K population means μ_i with properly chosen coefficients c_i that are given by

$$\psi = \sum_{i=1}^{K} c_i \mu_i, \tag{3}$$

such that the sum of the coefficients is zero (i.e., $\sum_{i=1}^{K} c_i$) is 0). We estimate ψ using

$$\hat{\psi}_n = \sum_{i=1}^{K} c_i \bar{X}_i, \tag{4}$$

with variance

$$\mathrm{Var}(\hat{\psi}_n) = \sum_{i=1}^{K} c_i^2 \sigma_i^2 / n_i. \tag{5}$$

Testing for Noninferiority/Superiority

In testing for the noninferiority of a new treatment compared to several existing treatments, the aim is to show that the new treatment is no worse (inferior) than similar existing treatments. For example, if a new treatment is compared to two similar existing treatments, then the effect size (ψ) is defined as in Equation 1.

For testing superiority, on the other hand, the aim is to show that the new treatment is better (superior) than the similar reference treatments. Testing for noninferiority/superiority is formulated as a one-sided hypotheses testing problem. Without any loss of generality, we assume that better treatment leads to a higher mean response. With this assumption, testing for noninferiority of the new treatment is formulated as testing the null hypothesis

$$\text{(Inferiority) } H_0: \psi \leq -|\psi_0|, \tag{6}$$

against the alternative hypothesis

$$\text{(Noninferiority) } H_A: \psi > -|\psi_0|, \tag{7}$$

where the value of $|\psi_0|$ represents the least relevant value of the contrast. The least relevant value is a crucial input for formulating the problem of testing noninferiority. It is fixed in advance and represents the threshold for minimum improvement or clinical significance. This is in clear contrast with the standard null hypothesis significance testing (NHST) approach, where ψ_0 is taken as zero.

Similarly, testing for superiority[1] can be formulated as a one-sided hypothesis testing for the null hypotheses

$$\text{(Non-Superiority) } H_0: \psi \leq |\psi_0|, \tag{8}$$

against the alternative hypotheses

$$\text{(Superiority) } H_A: \psi > |\psi_0|, \tag{9}$$

where ψ_0 is to be interpreted similarly as above. As before in the case of equivalence, the value of ψ_0 must be based on situational factors that are context-specific. The power function for the test for noninferiority is deduced and provided in Appendix A.

Testing for Equivalence

In various human-centered fields, such as psychology, education, management, marketing, human–computer interaction, drug development, psychiatry, and so on, two or more treatments are often compared. In such comparisons, researchers are often interested in testing whether the difference between the treatment effects is small enough to be considered essentially equivalent or not large enough to invoke change. This is in sharp contrast with the NHST framework of testing the null hypothesis of (exactly) no difference. For example, a null hypothesis of

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_K \tag{10}$$

for testing a one-way fixed-effect analysis of variance (ANOVA), commonly used in many human-centered fields, is omnibus in the sense that it is testing a general hypothesis (e.g., Maxwell et al., 2018).

Instead of testing only an omnibus effect, if researchers are interested in the similarity or comparability of specific effects among groups, rather than all of the groups being equal, such a question is formulated as a linear contrast. There, one can ask questions related to equivalence. Also, it is to be noted that, unlike ANOVA, in testing for equivalence, the hypothesis of equivalence is considered as an alternative hypothesis, rather than a null hypothesis.

Suppose we are interested in comparing the new treatment with several similar treatments. The new treatment will be considered to have equivalent effectiveness within specified margins ψ_1 and ψ_2 ($\psi_1 < 0 < \psi_2$) if, and only if, the effect size (ψ) lies between ψ_1 and ψ_2 (see Guo et al., 2011; Lehmann & Romano, 2006, p. 81; Luh & Guo, 2016) otherwise, the new treatment will be considered to be nonequivalent. The equivalence limits ψ_1 and ψ_2 are specified in advance to indicate the maximum clinically acceptable difference. Thus, the null and alternative hypotheses for testing equivalence are, respectively,

$$H_0: \psi \leq \psi_1, \quad \text{or} \quad \psi \geq \psi_2 \quad \text{and} \quad H_A: \psi_1 < \psi < \psi_2. \tag{11}$$

The power function for this test is deduced and provided in Appendix A.

Now, in order to plan a study with either an equivalence question or a noninferiority question, more information is needed, particularly about costs. For our example of comparing online training with face-to-face and hybrid training, there are costs per trainee for each of the training methods. Suppose, for face-to-face training, on average, the cost is \$500 per trainee, then this cost is based on (a) travel to the training site, (b) hotel stay, and (c) per diem expenses.[2] Similarly, the cost per trainee can be determined for the other two training methods as well.

For this example, suppose that the organization is willing to allocate up to \$50,000 to run the study in order to address one of the two questions.[3] Under this budget constraint, the study should be conducted to address either the equivalence question or the noninferiority question with maximum statistical power.

[1] If better treatment leads to lower response on the average, the testing for superiority can be formulated as, $H_0: \psi \geq -|\psi_0|$ against the alternatives $H_A: \psi < -|\psi_0|$.

[2] Without loss of generality, we use dollars, denoted \$, to denote cost.

[3] The idea of allocation costs may seem like an unusual value to specify given the context of the study design. However, this is done all the time while requesting grants to carry out the survey or conduct an experiment. Indeed, going into a study without consideration of how much one is willing to pay to address the question is shortsighted, we believe. All studies have an implicit budget allocation cost but here we seek to make it explicit.

In the next section, we develop a procedure to attain the maximum statistical power of the tests for noninferiority and equivalence subject to the budget constraint.

Sample Size Planning Under Cost Constraints

In practice, a researcher usually has a fixed budget, say $\$A_0$, for data collection. Prior to the actual collection of the data, the experimenter decides on the sample sizes obtained from the K populations subject in part to the budget constraint. The choice of sample sizes is especially important when the costs of collecting data per unit for different populations vary, and the populations are heterogeneous. Intuitively, it makes sense to allocate a larger sample size to a population in which the per unit cost is less compared to a population in which it is higher. Similarly, for a more heterogeneous population, the sample size should be higher compared to a less heterogeneous population. We now formulate the problem of finding the optimal sample sizes, where we use optimal in the sense of maximizing the statistical power of the tests for inferiority/superiority (e.g., see Equation C6), and equivalence (e.g., see Equation C16) subject to the cost constraint. Here, maximizing the statistical power of the tests for inferiority/superiority is equivalent to minimizing the variance (i.e., $\text{Var}(\hat{\psi}_n)$) subject to the budget constraint, holding all else equal.

Suppose that the cost of sampling per unit from the ith population is a_i, $i = 1, 2, \ldots, K$. Therefore, $A_0 = \sum_{i=1}^{K} a_i n_i$. Under the proportional allocation ratio, Theorem 1 provides the expression of optimal sample sizes for the K populations under cost constraints to get the minimized variance.

Theorem 1. *The optimal sample size required to maximize the statistical power of the tests for inferiority/superiority, which is equivalent to minimizing the variance subject to the available budget* ($\$A_0$), *is given by*

$$n_{i0} = \frac{A_0|c_i|\sigma_i}{\sum_{l=1}^{K} |c_l|\sigma_l\sqrt{a_l a_i}}, \quad \text{for } i = 1, \ldots, K \tag{12}$$

and the minimized variance of $\hat{\psi}_n$ *is*

$$V_o = \frac{1}{A_0}\left(\sum_{i=1}^{K} |c_i|\sigma_i\sqrt{a_i}\right)^2. \tag{13}$$

Then, the minimized variance can be used to have maximum statistical power under cost constraints.

Thus, the optimal sample sizes given in Theorem 1, with proof in Appendix B, maximize the power of the test subject to the constraint that the cost of sampling cannot exceed $\$ A_0$. Notice that to determine the optimal sample sizes, given in Theorem 1, the population variances (σ_i^2, $i = 1, 2, \ldots, K$) must be known. In practice, however, these are unknown quantities, and further, they may be unequal. To circumvent this difficulty, Luh & Guo (2016) propose an approximate solution for calculating optimal sample sizes. Their steps are as follows:

Step 1. In the expression for the sample sizes, replace the unknown population variances (σ_i^2) by the corresponding sample variances (s_i^2) based on an initial sample from the ith population, $i = 1, 2, \ldots, K$.

Step 2. Approximate the distribution of $\hat{\psi}_n$ by a t-distribution using the approximation given by Welch (1947) for the degrees of freedom ν, and then replace normal percentiles in the sample size formula by the corresponding t percentiles.

While proposing this approach, Luh & Guo (2016) assume that the initial samples are obtained either from a previous study or a pilot study undertaken specifically for the purpose of estimation of population variances. However, this assumption raises a number of serious theoretical, as well as practical, issues which Luh & Guo (2016) have ignored. First, Luh & Guo (2016) do not consider the cost of collecting samples in the pilot survey and, hence, do not include it in the cost constraint while determining the optimal sample sizes. Second, even if we decide to include the cost of collecting samples in the pilot study, the question that immediately arises is, how much of the *total sampling cost* (i.e., $\$ A_0$) should one spend on the pilot survey and how much should one set aside for the rest of the data collection? Third, Luh & Guo (2016) do not consider the use of the information in the initial samples for testing inferiority, superiority, or equivalence. Although the work of Luh & Guo (2016) is quite useful, in the next section, we develop a sequential approach to estimate the unknown optimal sample sizes, which circumvents all the issues stated above in a *learn-as-you-go* sequential framework.

Sequential Approach to Sample Size Determination

The optimal sample sizes n_{i0} ($i = 1, 2, \ldots, K$) given in Theorem 1 depend on the unknown variances σ_i^2, $i = 1, 2, \ldots, K$ of the K populations. Thus, in order to estimate the optimal sample sizes, estimators of the population variances are required; henceforth for each population, we will use the sample variance as an estimator of the corresponding unknown population variance. We now propose the following sequential estimation procedure for the estimation of the optimal sample sizes for the K populations.

Stage 1: First, $m_i(\geq 2)$ observations are randomly selected from the ith population ($i = 1, 2, \ldots, K$) such that $\sum_{i=1}^{K} m_i a_i$ is considerably less than A_0. Suppose $s_{im_i}^2$ is the sample variance for the ith population based on m_i observations. If

$$m_i \geq \frac{A_0|c_i|s_{im_i}}{\sum_{l=1}^{K} |c_l|s_{lm_l}\sqrt{a_l a_i}}, \tag{14}$$

stop sampling from the ith population and set the sample size N_{io} for the ith population equal to m_i. If, on the other hand,

$$m_i < \frac{A_0|c_i|s_{im_i}}{\sum_{l=1}^{K} |c_l|s_{lm_l}\sqrt{a_l a_i}}, \tag{15}$$

then proceed to the next stage for additional observations.

Stage 2: Without loss of generality, let us assume, m_i, $i = 1, 2, \ldots, r(\leq K)$, that is, the sample sizes for the first r populations fail to satisfy Equation 14 at stage 1, and $s_{1m_1} \leq s_{2m_2} \leq \cdots \leq s_{rm_r}$.

Draw m_r' additional observations from the rth population, that is, the population having maximum sample variance based on the first stage sample among the first r populations, and if

$$m_r + m'_r \geq \frac{A_0|c_r|s_{r,m_r+m'_r}}{\sum_{l\neq r}|c_l|s_{lm_l}\sqrt{a_l a_r} + |c_r|s_{rm_r+m'_r}a_r}, \tag{16}$$

stop sampling, otherwise, sampling from the rth population is to be continued at the next stage.

Next, we draw m'_{r-1} observations from the $(r-1)$th population, that is, the population having the second maximum sample variance based on the first stage sample among the first r populations, and if

$$m_{r-1} + m'_{r-1} \geq \frac{A_0|c_{r-1}|s_{r-1,m_{r-1}+m'_{r-1}}}{\sum_{l\neq r,r-1}|c_l|s_{lm_l}\sqrt{a_l a_{r-1}} + |c_{r-1}|s_{r-1m_{r-1}+m'_{r-1}}a_{r-1} + |c_r|s_{rm_r+m'_r}\sqrt{a_r a_{r-1}}}, \tag{17}$$

stop sampling, otherwise, continue sampling from the $(r-1)$th population at the next stage.

Continuing in this fashion, finally, we draw m'_1 observations from the first population, the population having minimum standard deviation. However, at every stage of the sampling process, we need to ensure that the total cost of sampling does not exceed the prespecified budget \$ A_0. For example, if it so happens that the total cost of sampling would exceed \$ A_0 if we draw m'_i observations from the i $(1 \leq i \leq r)$th population, but would remain less than \$ A_0 as long as the sample size does not exceed $m^*_i(< m'_i)$, then we stop sampling as soon as we draw m^*_i observations from the ith population. For the implementation of the procedure, we thus suggest that m'_i $(i = 1, 2, \ldots, r)$ be kept at its minimum value 1 from Stage 2 onward unless there is a reason to think otherwise. Figure 1 outlines a flowchart presenting the sequential procedure.

It is evident from the above discussion that the sampling should stop as soon as either of the following conditions is satisfied: (a) the sample sizes for all the K populations meet the stopping rule criterion or (b) at any stage during the sampling process the total sampling cost exceeds the prespecified budget \$ A_0. Let N_i be the final sample size from the ith $(i = 1, 2, \ldots, K)$ population. Clearly, given the cost constraint, some of the N_i values may fail to meet the stopping rule criterion. Also, it is important to note that, at any point of time during the sampling process, the stopping rule uses the updated estimates of the population variances based on the cumulative samples obtained till that point of time from different populations.

The sequential sampling procedure described here, for which we work through the statistical theory, can be summarized as follows: stop sampling from the ith population as soon as $N_i(\geq m_i)$, the cumulative sample size from the ith population, satisfies

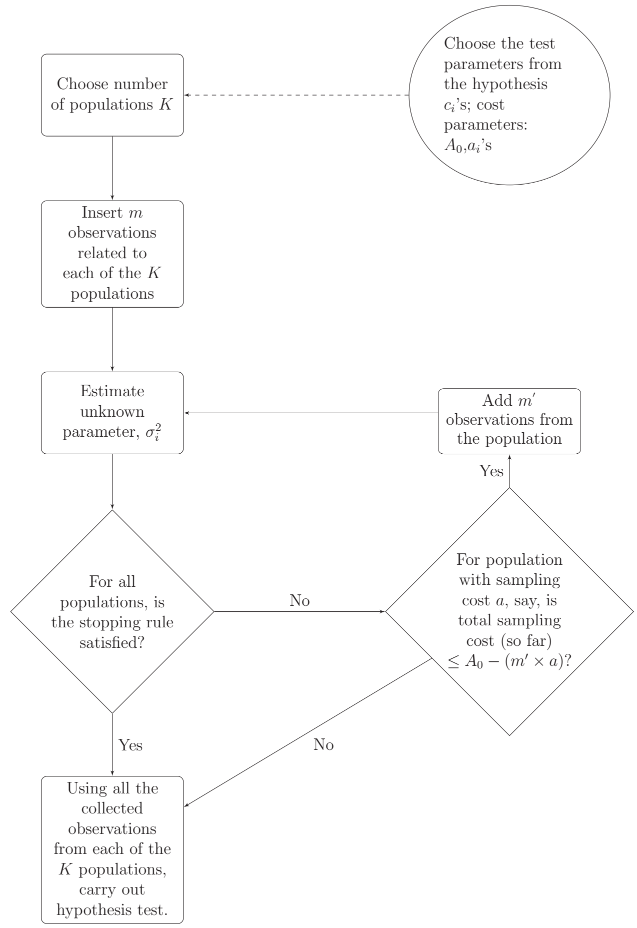
$$N_i \geq \frac{A_0|c_i|s_{iN_i}}{\sum_{l=1}^{K}|c_l|s_{lN_l}\sqrt{a_l a_i}}, \tag{18}$$

such that $\sum_{i=1}^{K} a_i N_i \leq A_0$ and for $l = 1, 2, \ldots, K$, s_l represents an updated estimate of the corresponding population standard deviation (σ_l).

Note that, for large sample sizes, $s_i/(\sum_{l=1}^{K}|c_l|s_l\sqrt{a_l})$ converges to $\sigma_i/(\sum_{l=1}^{K}|c_l|\sigma_l\sqrt{a_l})$. In fact, we prove that for a large enough budget (A_0), on average, the final sample size N_i of the sequential procedure is close to the optimal sample size n_{i0} (see Theorem 1) from the ith population $i = 1, 2, \ldots, K$. The proof of the result is given in Appendix D as Theorem D1.

Figure 1

Flowchart That Describes the Sequential Procedure Developed



With the conclusion of the proposed sequential procedure, there would be N_1 observations from population 1, N_2 observations from population 2, and so forth. Notice that $N_1, N_2, \ldots, N_K$ are estimates of the optimal sample sizes required to achieve a minimal variance that in turn maximizes the statistical power under the budget constraints.

Next, with $N_1, N_2, \ldots, N_K$ observations collected from the corresponding K populations, we carry out the hypothesis test to address either the equivalence question or the inferiority/superiority question.

Consider an example in which there are three populations and the cost of sampling from each population is the same, say \$30 $(= a_i$, $i = 1, 2, 3)$. Suppose that the total amount $A_0 = \$\ 10,000$ is allotted for the sampling process. Provided that the three population variances are the same, it can be deduced that the optimal sample size will be $n_{i0} = 111.11$, which is not an integer. In such a scenario, we round to the nearest integer and thus consider 111 observations per group (i.e., a total sample size of 333), which will make the total cost \$9,990. A ceiling function of the same cannot be taken as that would increase the sampling cost to be more than the budget constraint. So, a realistic and objective assumption is that n_{i0}, the ith group optimal sample size, must be an integer and the total sampling cost in some scenarios may be lesser by some amount than the allocated sampling cost. In such a situation, the proposed sequential procedure will provide estimates (N_{io}, $i = 1, \ldots, 3$) of the optimal sample sizes for which the total sampling cost will be less than the allocated sampling cost \$$A_0$.

Choice of the Initial Sample Sizes

In the absence of any prior information about the relative magnitudes of the K population variances, for implementation of the sequential procedure, the initial sample sizes $m_i(\geq 2)$'s may be taken to be equal, say, m. The minimum number of observations required to get an estimate of the population variance is 2. If $m_i = 2$, then the estimates of the K population variances will be unstable and would require a large number of sampling stages. So the sequential sampling procedure would perform well with a lesser number of sampling stages if the estimates of the K population variances are reasonably good. In order to ensure this, the initial sample sizes should be chosen large enough to yield reasonably good estimates of the K population variances. At the same time, too large initial sample sizes may lead to oversampling (i.e., a larger sample size than necessary). We recommend that the pilot sample be set to what researchers in the field would find reasonable.

Performance of the Sequential Procedure

An attractive property of the proposed sequential procedure is laid out in Theorem D1 in Appendix D. In particular, Theorem D1 ensures that, under appropriate conditions, the average of the ratio of final sample size N_i and the optimal sample size, n_{i0} approaches unity. Since the optimal sample size is required to achieve minimized variance of the contrast and thereby attains maximum power, so from the theorem, it can be expected that the power attained using the final sample size, as provided by the proposed sequential procedure, asymptotically achieves the maximum power in an average sense.

We now investigate the performance of the proposed sequential procedure using an extensive simulation study. We fix the total sampling cost (A_0) at \$5,000 and the number (K) of populations at 3. Notice that, the estimate of the optimal sample size n_{i0} depends on the estimate of σ_i besides A_0, a_i and c_i, $i = 1, 2, \ldots, K$. The accuracy of the estimate of σ_i depends on the shape of the population distribution and, in turn, affects the accuracy of the estimate of n_{i0}. Keeping this fact in mind, we consider three kinds of distributions,

Table 1
Summary of Final Sample Sizes

Distributions	ψ V_o	c_i a_i	n_{i0} MAD_R	$\overline{N}_i$ $SD\ (N_i)$	$Av(\frac{N_i}{n_{i0}})$ MAD_{N_i}	$\overline{V}_{oN}$ $\text{MAD}_{V_{oN}}$	$\overline{P}$ $SD\ (P)$	MAD_P
G(shape = 16, n scale = 0.2)		1	41.2035	40.7116	0.9881			
	10	0.158	11.2713	6.547				
G(shape = 50, scale = 0.125)	− 0.2999	− 2	83.1149	83.1656	1.001	0.1849	0.9999	
	0.1885	12	0.0969	17.1615	8.052	0.0206	2.7579×10^{-5}	2.0496×10^{-5}
G(shape = 6, n scale = 1.50001)		1	99.7385	99.3264	0.9959			
		36	0.0346	7.2054	3.407			
L(location = 1, n scale = 2)		1	42.2309	41.6286	0.9857			
	16	0.1629	11.7733	6.917				
L(location = 2, n scale = 1)	1	− 2	45.1468	44.8272	0.9929	2.241	0.9897	
	2.3058	14	0.1567	12.0067	7.087	0.2867	0.0016	0.0012
L(location = 4, n scale = 7)		1	94.6731	94.2162	0.9952			
	39	0.0422	7.1701	3.947				
N(M = 1, n SD = 4.5)		1	96.373	95.868	0.9948			
	38	0.0338	6.7458	3.19				
N(M = 2, n SD = 0.9)	0	− 2	61.3566	61.3886	1.001	0.2819	0.9998	
	0.2869	15	0.1063	13.4775	6.521	0.0285	9.4796×10^{-5}	7.4274×10^{-5}
N(M = 3, n SD = 1)		1	41.7479	41.3042	0.9894			
	10	0.1432	10.8187	6.008				
0.95 N(1,19)+0.05 N(20,15)		1	75.913	76.5274	1.008			
	38	0.055	8.1459	4.13				
0.95 N(2,5)+ 0.05 N(20,8)	− 3.8	− 2	81.5976	80.1144	0.9818	8.228	0.9771	
	0.8486	15	0.1114	16.0867	9.075	0.8933	0.0019	0.0011
0.95 N(3,11)+ 0.05 N(20,11)		1	89.1344	88.0748	0.9881			
	10	0.0888	16.0584	7.784				

Note. ψ is the linear combination of the population means; c_i is the coefficient for the corresponding population mean used for finding ψ; a_i is the sampling cost; V_o and n_{i0} are, respectively, the theoretical minimized variance and the optimal sample size if the procedure is used assuming the population parameters to be known; MAD_R is the mean absolute deviation of the ratio (N_i/n_{i0}) of the final sample size and the optimal sample size; $\overline{N}_i$, $SD(N_i)$, and $\overline{V}_{oN}$ are, respectively, the mean of the final sample sizes, the standard deviation of the final sample sizes, and the average of the estimated minimized variance using the proposed procedure; MAD_{N_i}, and MAD_{V_N} are, respectively, the mean absolute deviation of the final sample sizes, and the estimated variance of the contrast based on the final sample sizes; $\overline{P}$ is the average power for testing for inferiority when $\delta = 1$ and $\alpha = 0.05$, and SD (P) is the standard deviation of power values across replications, and MAD_P is the corresponding mean absolute deviation; tabled values are based on 5,000 replications of a Monte Carlo simulation study from different distributions, namely the Gamma (G), logistic (L), normal (N), and a mixture of normal distributions (95% from one normal distribution and the rest 5% from another normal distribution).

namely, gamma with different shape and scale parameters (can take various shapes), logistics with different location and scale parameters (symmetric and unimodal), and the mixture of two normal distributions with different mean and standard deviation (bimodal). Also, we consider varying choices of the values of c_i and a_i.

We then implement the proposed sequential procedure and report the mean of the final sample sizes, $\bar{N}_i$ (an estimate of the expected final sample size, E[N_i], the standard deviation of N_i's ($s(N_i)$), the average of the ratio of the final sample size and the estimated optimal sample size ($Av(N_i/n_{i0})$) based on 5,000 replications via Monte Carlo simulations. The findings are summarized in Table 1. In each replication, we first draw $m = 10$ observations from all the K populations and then follow the algorithm of the proposed sequential procedure by drawing $m' = 1$ observations from each group at each stage after the pilot stage.

Table 1 presents some performance indicators of the proposed sequential procedure. In the following, we summarize our findings based on it. Column 4 furnishes n_{i0} and the mean absolute deviation of the ratio (N_i/n_{i0}) from unity. We denote the latter by MAD$_R$. As expected, its values are mostly close to zero. Column 5 reports the values of $\bar{N}_i$ and $s(N_i)$. Although $\bar{N}_i$ is close to n_{i0}, the deviation of N_i values across replications seems moderate. In column 6, the average of the ratios (N_i/n_{i0}) across replications and the mean absolute deviation of N_i's are shown. As expected, the average of the ratios is close to unity. However, the mean absolute deviation of N_i's is moderate. The former provides an empirical verification of Theorem D1. Additionally, we observed that the overall sampling cost never exceeded $A_0 (= \$\ 5{,}000)$. Column 7 shows the average of the estimated optimal variance of the contrast ($\overline{V}_N$) based on the final sample sizes across replications and the mean absolute deviation of V_N values across replications. Although the deviations of N_i values from n_{i0} are found to be moderate, it is interesting to note that the mean absolute deviation of V_N is close to zero. Thus, V_N is close to the variance based on the optimal sample sizes and known values of the population parameters (V_o).

Next, for each replication, we find the attained power of the test by plugging in the variance (V_N) in the power expression, assuming that the deviation (δ) of the contrast values under the null and alternative hypotheses is equal to 1. The ninth and tenth columns show the average attained power ($\bar{P}$), its standard deviation, and its mean absolute deviation. The results presented in these two columns clearly indicate that the average of the attained maximized powers is high, with a small standard deviation and mean absolute deviation. For other values of δ, one can similarly compute the attained power.

Thus, overall, we find that our proposed sequential procedure works remarkably well in terms of all performance indicators.

Comparison With Luh and Guo Procedure

To compare the performance of the proposed sequential procedure with that of Luh & Guo (2016), we consider the same simulation set-up as above. Also, as above, we suppose the total budget for data collection is \$5,000. To implement the Luh and Guo procedure, a pilot study is conducted, and the optimal sample sizes are estimated by plugging in the estimates of population variances based on the data obtained in the pilot study. For our simulation study, we assume that in the pilot study, data are collected on 10 subjects from each of the three populations. Notice that the Luh and Guo procedure completely ignores the data collected in the pilot study for carrying out the test of hypotheses. Thus, if one applies the Luh & Guo (2016) procedure, then for conducting the hypothesis testing, the budget for data collection reduces to \$5,000 minus the cost of data collection in the pilot study, instead of \$5,000. In contrast, the proposed sequential procedure considers observations collected at all stages, including the pilot stage, for estimating the sample sizes and carrying out the test. Thus, the proposed sequential procedure utilizes the budget more efficiently than that of Luh & Guo (2016). Also, using the data from the pilot stage requires that no changes to the study protocol have been made.

In our simulation study, we ignore the cost of the pilot study of the Luh and Guo procedure. Suppose the estimated optimal sample sizes are denoted by $N_{i\mathrm{LG}}$. The results are presented in Table 2. Notice that the performance indicators of the Luh and Guo procedure presented in columns 2–5 of Table 2 correspond to those of columns 4–7, respectively, of Table 1 for the proposed procedure. In column 6 of Table 2, the deviations (δ) at which the attained power of the test for noninferiority/superiority at the 5% level is either 0.8 or 0.9 are presented. These values are computed based on the sample sizes of the Luh and Guo procedure (δ_{LG}), the proposed sequential procedure (δ_S), and the optimal procedure assuming population variances to be known (δ_{optim}).

Comparing Tables 1 and 2, we observe that the proposed sequential procedure has a smaller MAD$_R$, standard deviation, and mean absolute deviation of the estimated sample sizes as compared to the Luh & Guo (2016) procedure in all cases. In fact, the mean absolute deviations of the estimated sample sizes using Luh & Guo (2016) procedure are substantially higher than the proposed sequential procedure. Thus, the sample sizes estimated by the proposed sequential procedure are closer to the optimum sample sizes compared to those of the Luh & Guo (2016) procedure. Also, the estimated variance of the contrast corresponding to the proposed sequential procedure is less, on average, compared to that of the Luh & Guo (2016) procedure. This phenomenon can be attributed to the fact that the sample size determination in the Luh and Guo procedure is based on a pilot sample.

Also, the average of the estimated δ values corresponding to the proposed sequential procedure ($\overline{\delta}_S$) is closer to that of the optimum procedure (δ_{optim}) as compared to that of the Luh & Guo (2016) procedure ($\overline{\delta}_{\mathrm{LG}}$). Since the estimated δ for the sequential procedure on average is more than that for Luh & Guo (2016), it is easier to detect a true superiority/noninferiority by the sequential procedure than by the Luh & Guo (2016) procedure, although the standard deviation ($SD(\delta_{\mathrm{LG}})$) and the mean absolute deviation ($\mathrm{MAD}_{\delta_{\mathrm{LG}}}$) of the delta values corresponding to the Luh and Guo procedure for all cases are slightly less than that of the proposed sequential procedure.

Discussion

The sequential methodology developed in this article yields the sample sizes to be selected from each group to achieve the maximum statistical power for carrying out tests for noninferiority/superiority or equivalence, given a budget constraint. In general, the sample size increases with the increase in variance and decreases with the increase in sampling cost. Thus, if for a particular population or group, it costs more to obtain a new observation compared to the other groups, then we select a smaller number of observations from that group. On the other hand, if the variability of a population

Table 2
Comparison With Luh–Guo's Work for Test for Noninferiority/Superiority

Distributions	n_{i0} MAD_{RNLG}	$\overline{N}_{i\text{LG}}$ $SD\ (N_{i\text{LG}})$	$Av(\frac{N_{i\text{LG}}}{n_{i0}})$ $\text{MAD}_{N_{i\text{LG}}}$	$\overline{V}_{o\text{NLG}}$ $\text{MAD}_{V_{o\text{NLG}}}$	$\overline{\delta}_{\text{LG}}, \overline{\delta}_S, \delta_{\text{optim}}$ $SD\ (\delta_{\text{LG}}), SD(\delta_S)$ $\text{MAD}_{\delta_{\text{LG}}}, \text{MAD}_{\delta_S}$
G(shape = 16, scale = 0.2)	41.2035	42.9117	1.0415		**Power = 0.8**
	0.2549	13.7391	10.6		− 0.3538, − 0.3443, − 0.3487 0.0237, 0.0278
G(shape = 50, scale = 0.125)	83.1149	86.0128	1.0349	0.1948	0.0188, 0.0196
	0.228	24.1336	19.01	0.0208	**Power = 0.9**
					− 0.1600, − 0.1557, − 0.1577
G(shape = 6, scale = 1.50001)	99.7385	98.298	0.9856		0.0107, 0.0126
	0.0768	9.7298	7.691		0.0085, 0.0088
L(location = 1, scale = 2)	42.2309	43.5723	1.0318		**Power = 0.8**
	0.264	14.1704	11.23		− 1.2374, − 1.1978, − 1.2202 0.0894, 0.1067
L(location = 2, scale = 1)	45.1468	46.738	1.0352	2.3856	0.0703, 0.0780
	0.2647	15.3725	12.04	0.2715	**Power = 0.9**
					− 0.5597, − 0.5418, − 0.5519
L(location = 4, scale = 7)	94.6731	93.5516	0.9882		0.0405, 0.0483
	0.0741	8.8671	7.028		0.0318, 0.0353
$N(M = 1, SD = 4.5)$	96.373	95.3967	0.9899		**Power = 0.8**
	0.0664	8.1742	6.426		− 0.4360, − 0.4253, − 0.4302 0.0251, 0.0319
$N(M = 2, SD = 0.9)$	61.3566	62.8708	1.0247	0.2956	0.0198, 0.022
	0.2156	16.8645	13.26	0.0268	**Power = 0.9**
					− 0.1972, − 0.1924, − 0.1946
$N(\text{mean}M = 3, SD = 1)$	41.7479	43.1864	1.0345		0.0114, 0.0144
	0.2379	12.7986	9.986		0.0089, 0.010
0.95 N(1,19) + 0.05 N(20,15)	75.913	76.2801	1.0048		**Power = 0.8**
	0.1095	10.3803	8.291		− 2.3927, − 2.2960, − 2.3389 0.1398, 0.1916
0.95 N(2,5) + 0.05 N(20,8)	81.5976	79.2446	0.9712	8.9039	0.1110, 0.1289
	0.2412	24.1194	19.37	0.8267	**Power = 0.9**
					− 1.0822, − 1.0385, − 1.0579
0.95 N(3,11) + 0.05 N(20,11)	89.1344	91.2688	1.0239		0.0632, 0.0867
	0.2079	23.4618	18.58		0.0502, 0.0583

Note. n_{i0} is the theoretical sample size if the population parameters values are used; $\overline{N}_{i\text{LG}}$, $SD(N_{i\text{LG}})$, and $\overline{V}_{o\text{NLG}}$ are, respectively, the mean final sample sizes of the populations, the standard deviation of the final sample sizes of the populations, and the average of the estimated minimized variance using the LG procedure; $\text{MAD}_{\text{RN}_{\text{LG}}}$ is the mean absolute deviation of the ratio ($N_{i\text{LG}}/n_{i0}$) of the final sample size using LG procedure and the optimal sample size; $\text{MAD}_{N_{\text{LG}}}$ and $\text{MAD}_{V_{oN_{\text{LG}}}}$ are, respectively, the mean absolute deviation of the final sample size and variance using the LG procedure; $\overline{\delta}_{\text{LG}}$, $\overline{\delta}_S$, andδ_{optim} are, respectively, denote the average distance value between the null from alternative hypotheses required for obtaining a prespecified (=0.8, 0.9) power while applying the LG, proposed sequential procedures, and when the population parameters values are used; $SD(\delta_{\text{LG}})$ and $SD(\delta_S)$ indicate respective standard errors; $\text{MAD}_{\delta_{\text{LG}}}$ and MAD_{δ_S} indicate the mean absolute deviations of the delta values for two procedures; tabled values are based on 5,000 replications of a Monte Carlo simulation study from Gamma (G), logistic (L), normal (N), and a mixture of normal distributions.

or group is greater compared to the other groups, we select more observations from that group. Usually, the population variances are not known, and thus, for determining the sample sizes from different groups, population parameter estimates need to be used. In the present paper, we develop a sequential procedure to estimate the optimal sample sizes, which updates the estimate of variances as more and more observations are obtained. We show that this procedure leads to optimal sample sizes yielding maximum statistical power given a budget constraint.

The simulation studies show that, on average, the sample sizes determined by the sequential procedure and the (theoretical) optimal sample sizes required to achieve the maximum power are sufficiently close to each other. Furthermore, our procedure does not require the assumption of a particular distribution (e.g., normal) of the data. Our procedure performs well, provided we have enough sample sizes for the central limit theorem to work. Also, our procedure ensures that the total sampling cost does not exceed the prespecified budget amount $\$A_0$. Furthermore, unlike the existing methods, the proposed method does not require any prior knowledge of the value(s) of the population variance(s), which is the case for traditional power analysis (i.e., the population variance of each of the populations or a single value for an assumed homogeneous variance situation ahead of time).

Next, we discuss some limitations of the proposed sequential procedure, which can be considered for further research work.

(1) We have assumed that the random observations drawn from the populations are independent and identically distributed. Therefore, the procedure is not currently amenable to multilevel structures.
(2) Our sequential procedure holds in general and not for a specific data distribution like a normal distribution. However, it does require that the sample sizes in the groups are not too small so that the central limit theorem works. What is "not too small" depends on various factors, like the skewness and kurtosis of the population distributions. For example, in the case of symmetric distributions, small sample sizes may work well. If the populations are distributed normally, the proposed method does not require any sample size restriction. It is required only when the population distributions are nonnormal. The proposed method stops sampling once the stopping rule is satisfied. Based on the sample sizes, when it stops, we carry out the hypothesis testing.
(3) The proposed sequential procedure considers a situation in which the prespecified data collection cost from the populations for every observation remains the same, c_i for the ith population ($i = 1, 2, \ldots, k$). In many situations, there may be potential savings or additional costs in per unit sampling cost for the first stage as compared to the second stage and, subsequently the later stages. As an instance, in an activity monitoring experiment, the $50(= m)$ participants included in the first stage might cost \$20 per participant, while the second stage, which comprises $10(= m')$ participants, might require \$25 for each participant and similarly for the later stages. Our procedure is not equipped to incorporate fluctuating data collection costs for each stage.
(4) The proposed sequential procedure requires additional m' observations in the k population at every stage, after the pilot stage, until the stopping rule is met for a population. Choosing a larger value of m' might lead to oversampling (e.g., ending a study at 130 per group when it could have ended with 105 per group). For fixing a value of m', economic considerations may be taken into account. However, the choice of m' does not affect the performance of the sequential procedure except that for larger m' values, fewer sampling stages are required.
(5) Achieving a high degree of statistical power is certainly an important problem in hypothesis testing. For these types of studies, often an external research grant agency is approached with a specified budget for data collection. The proposed sequential procedure, however, does not guarantee a prespecified statistical power at a given alternative. It achieves the minimum variance or, equivalently, the maximum power at a specified alternative under budget constraints. Thus, our procedure is challenged with fixing a sampling budget a priori for achieving a certain power at a prespecified alternative. For this purpose, a "fixed-n" approach such as Luh & Guo (2016) is useful. There, optimal sample sizes are estimated by considering estimated population standard deviation values from an initial pilot study or using supposed values (e.g., expert opinion, prior study, and meta-analysis). Caution is necessary here, though, because even with a small amount of deviation in standard deviation estimates, the targeted statistical power may not be achieved after the study. Thus, the attained power may differ from the targeted power the study was said to have. Nevertheless, if necessary, the optimal sample size can be determined based on the prespecified values, which are provided in Appendix C, and which are the same as Equations 2, 8, and 9 in Luh & Guo (2016). Provided a sampling budget, the aim should be to maximize power by minimizing the variance of the contrast using our proposed sequential procedure.
(6) We assume no cost of analysis for each check of the stopping rule. In many situations, this is reasonable (e.g., it is negligible to perform a check from data obtained from an online study). However, if a researcher must pay a consultant, for example, for each check of the stopping rule, it would not only take time but also impose a cost.

We believe this article solves an important piece missing from the extant literature. Moving to a sequential framework in a data distribution-free environment in which the cost of sampling is explicitly considered helps advance the design literature in a flexible framework.

References

Anderson, S. F., & Kelley, K. (2022). Sample size planning for replication studies: The devil is in the design. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000520

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Chattopadhyay, B. (2020). Accuracy in insurance billing error estimation using auxiliary information. *South African Statistical Journal*, *54*(2), 153–162. https://doi.org/10.37920/sasj.2020.54.2.2

Chattopadhyay, B., & Banerjee, S. (2021). Estimation of location parameter within pre-specified error bound with second-order efficient two-stage procedure. *South African Statistical Journal*, *55*(1), 45–54. https://doi.org/10.37920/sasj.2021.55.1.4

Chattopadhyay, B., & De, S. K. (2016). Estimation of Gini index within pre-specified error bound. *Econometrics*, *4*(3), Article 30. https://doi.org/10.3390/econometrics4030030

Chattopadhyay, B., & Kelley, K. (2016). Estimation of the coefficient of variation with minimum risk: A sequential method for minimizing sampling error and study cost. *Multivariate Behavioral Research*, *51*(5), 627–648. https://doi.org/10.1080/00273171.2016.1203279

Chattopadhyay, B., & Kelley, K. (2017). Estimating the standardized mean difference with minimum risk: Maximizing accuracy and minimizing cost with sequential estimation. *Psychological Methods*, *22*(1), 94–113. https://doi.org/10.1037/met0000089

Darku, F. B., & Chattopadhyay, B. (2021). Gini's mean difference-based minimum risk point estimator of mean. In N. Mukhopadhyay & P. Pratim Sengupta (Eds.), *Gini inequality index* (pp. 93–106). Chapman and Hall/CRC Press.

Dumas-Mallet, E., Button, K. S., Boraud, T., Gonon, F., & Munafò, M. R. (2017). Low statistical power in biomedical science: A review of three human research domains. *Royal Society Open Science*, *4*(2), Article 160254. https://doi.org/10.1098/rsos.160254

Goeman, J. J., Solari, A., & Stijnen, T. (2010). Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority. *Statistics in Medicine*, *29*(20), 2117–2125. https://doi.org/10.1002/sim.4002

Guo, J. H., Chen, H. J., & Luh, W. M. (2011). Sample size planning with the cost constraint for testing superiority and equivalence of two independent groups. *British Journal of Mathematical and Statistical Psychology*, *64*(3), 439–461. https://doi.org/10.1348/000711010X512408

Gut, A. (2009). *Stopped random walks: Limit theorems and applications*. Springer.

Kelley, K., Darku, F. B., & Chattopadhyay, B. (2018). Accuracy in parameter estimation for a general class of effect sizes: A sequential approach. *Psychological Methods*, *23*(2), 226–243. https://doi.org/10.1037/met0000127

Kelley, K., Darku, F. B., & Chattopadhyay, B. (2019). Sequential accuracy in parameter estimation for population correlation coefficients. *Psychological Methods*, *24*(4), 492–515. https://doi.org/10.1037/met0000203

Keselman, H., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, *68*(3), 350–386. https://doi.org/10.3102/00346543068003350

Lehmann, E. L., & Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Luh, W. M., & Guo, J. H. (2016). Sample size planning for the noninferiority or equivalence of a linear contrast with cost considerations. *Psychological Methods*, *21*(1), 13–34. https://doi.org/10.1037/met0000039

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experiments and analyzing data: A model comparison perspective* (3rd ed.). Routledge.

Welch, B. L. (1947). The generalization of Student's problem when several different population variances are involved. *Biometrika*, *34*(1–2), 28–35. https://doi.org/10.1093/biomet/34.1-2.28

(*Appendices follow*)

Appendix A

Statistical Formulation

We estimate the contrast $\psi = \sum_{i=1}^{k} c_i \mu_i$ using $\hat{\psi}_n$ as defined in Equation 4. From Equation 5, the variance of $\hat{\psi}_n$ is

$$\mathrm{Var}(\hat{\psi}_n) = \sum_{i=1}^{K} c_i^2 \sigma_i^2 / n_i. \tag{A1}$$

Let γ_i be the ith sample size allocation ratios defined as

$$n_i = \gamma_i n_1 \tag{A2}$$

for $i = 1, \ldots, K$ with $\gamma_1 = 1$. Thus, using Equation A2 in Equation A1, we obtain

$$\mathrm{Var}(\hat{\psi}_n) = \frac{1}{n_1} \sum_{i=1}^{K} c_i^2 \sigma_i^2 / \gamma_i. \tag{A3}$$

If the smallest of the per-group sample sizes (i.e., the K n_i values) is not too small,[4] using the Central Limit Theorem, we have

$$\frac{\sqrt{n_1}(\hat{\psi}_n - \psi)}{\sqrt{\sum_{i=1}^{K} c_i^2 \sigma_i^2 / \gamma_i}} \xrightarrow[\mathcal{D}]{} N(0, 1). \tag{A4}$$

This result would help to draw inferences about the contrast parameter ψ without requiring any additional assumption about the distributions of the parent populations.

Appendix B

Proof of Theorem 1

Proof. Using Equation A2, we have, $A_0 = \sum_{i=1}^{K} a_i n_i = n_1 \sum_{i=1}^{K} a_i \gamma_i$. That is,

$$n_1 = \frac{A_0}{\sum_{i=1}^{K} a_i \gamma_i}. \tag{B1}$$

Plugging the value of n_1, obtained from Equation B1, in Equation 5, we have

$$\mathrm{Var}(\hat{\psi}_n) = \frac{1}{A_0} \left(\sum_{i=1}^{K} \frac{c_i^2 \sigma_i^2}{\gamma_i} \right) \left(\sum_{i=1}^{K} a_i \gamma_i \right), \tag{B2}$$

which will be minimized under the optimal allocation ratio given as

$$\gamma_i = \frac{|c_i| \sigma_i}{|c_1| \sigma_1} \sqrt{\frac{a_1}{a_i}}. \tag{B3}$$

For justification of Equation B3, see Luh & Guo (2016). Thus, the optimal sample sizes for the K populations under cost constraints are given as

$$n_{i0} = \frac{A_0 |c_i| \sigma_i}{\sum_{l=1}^{K} |c_l| \sigma_l \sqrt{a_l a_i}} \quad \text{for } i = 1, \ldots, K \tag{B4}$$

and the minimized variance of $\hat{\psi}_n$ under the optimal allocation ratio is

$$V_o = \frac{1}{A_0} \left(\sum_{i=1}^{K} |c_i| \sigma_i \sqrt{a_i} \right)^2.$$

[4] By "not too small" here and elsewhere we mean a sample size that is large enough so that the noted properties hold well enough for the statement to be true. The exact value of "not too small" is context specific.

(*Appendices continue*)

Appendix C

Testing for Noninferiority/Superiority

For testing for noninferiority, we test

$$\begin{aligned}&\text{(Inferiority) } H_0: \psi \le -|\psi_0|\\ &\text{against(Noninferiority) } H_A: \psi > -|\psi_0|\end{aligned} \tag{C1}$$

where the value of $|\psi_0|$ represents the least relevant value of the contrast.

Similarly testing for superiority, we test

$$\begin{aligned}&\text{(Non-superiority) } H_0: \psi \le |\psi_0|\\ &\text{against(Superiority) } H_A: \psi > |\psi_0|.\end{aligned} \tag{C2}$$

From Equation C1, it is evident that the hypothesis of “Inferiority” H_0 is to be rejected at a level of significance α (see Guo et al., 2011; Lehmann & Romano, 2006, p. 81) if

$$\hat{\psi}_n > -|\psi_0| + z_\alpha\sqrt{\mathrm{Var}(\hat{\psi}_n)}, \tag{C3}$$

where z_α is the upper 100α percentile point of standard normal distribution. Similarly, the hypothesis of nonsuperiority is to be rejected at level α if

$$\hat{\psi}_n > |\psi_0| + z_\alpha\sqrt{\mathrm{Var}(\hat{\psi}_n)}. \tag{C4}$$

The power of a test is defined as the probability of rejecting the null hypothesis when it is false. For finding the power of the test for inferiority (see Equation C1), note that a value of ψ, specified under the alternative hypotheses, can be written as $-|\psi_0| + \delta$, where $\delta > 0$. The power of the test at $\psi = -|\psi_0| + \delta$ can then be written as

$$\begin{aligned}\text{Power} &= P(\hat{\psi}_n > -|\psi_0| + z_\alpha\sqrt{\mathrm{Var}(\hat{\psi}_n)}|\psi = -|\psi_0| + \delta)\\ &= P\left(\frac{\hat{\psi}_n - (-|\psi_0| + \delta)}{\sqrt{\mathrm{Var}(\hat{\psi}_n)}} > \frac{-\delta}{\sqrt{\mathrm{Var}(\hat{\psi}_n)}} + z_\alpha\right)\\ &= \Phi\left(\frac{\delta}{\sqrt{\mathrm{Var}(\hat{\psi}_n)}} - z_\alpha\right),\end{aligned} \tag{C5}$$

where $\Phi(.)$ is the cumulative distribution function of a standard normal variable, and $P(A|\psi = -|\psi_0| + \delta)$ denotes that the probability of the event A under the assumption that $-|\psi_0| + \delta$ is the true value of ψ. Using Equation 5 in Equation C5, we obtain the power at $\psi = -|\psi_0| + \delta$ equal to

$$\Phi\left(\frac{\sqrt{n_1}\delta}{\sqrt{\sum_{i=1}^K c_i^2\sigma_i^2/\gamma_i}} - z_\alpha\right). \tag{C6}$$

Suppose our problem is to determine the K sample sizes n_i following the allocation in Equation B3 such that the power at $\psi = -|\psi_0| + \delta$ is equal to a prespecified value $1-\beta$, that is,

$$\Phi\left(\frac{\sqrt{n_1}\delta}{\sqrt{\sum_{i=1}^K c_i^2\sigma_i^2/\gamma_i}} - z_\alpha\right) = \Phi(z_\beta). \tag{C7}$$

With this, we obtain

$$n_1 = \frac{(z_\alpha + z_\beta)^2}{\delta^2}\left(\sum_{i=1}^K c_i^2\sigma_i^2/\gamma_i\right). \tag{C8}$$

Thus, given the contrast coefficients c_i's, level α, power $1-\beta$, and a fixed alternative $-|\psi_0| + \delta$, we obtain $n_i = n_1\gamma_i$, assuming that the variances σ_i^2's are known. The same sample size allocation formula given in Equation C8 is valid for testing superiority at level α for attaining power $1-\beta$ at a fixed alternative $|\psi_0| + \delta$.

Given the contrast coefficients c_i's, level α, achieved minimized variance (V_{oN}) and a fixed alternative $-|\psi_0| + \delta$, the achieved power is

$$\Phi\left(\frac{\delta}{\sqrt{V_{oN}}} - z_\alpha\right). \tag{C9}$$

Testing for Equivalence

The null and alternative hypotheses for testing equivalence are, respectively,

$$H_0: \psi \le \psi_1, \quad \text{or} \quad \psi \ge \psi_2 \quad \text{and} \quad H_A: \psi_1 < \psi < \psi_2. \tag{C10}$$

In order to motivate the testing procedure to be adopted for testing equivalence, it helps to formulate the above hypothesis testing problem in terms of two one-sided hypothesis testing problems as follows. Consider the following testing problems:

$$H_{01}: \psi \le \psi_1 \quad \text{against} \quad H_{A1}: \psi > \psi_1 \tag{C11}$$

and

$$H_{02}: \psi \ge \psi_2 \quad \text{against} \quad H_{A2}: \psi < \psi_2. \tag{C12}$$

Here, the rejection regions for testing H_{01} and H_{02} with a Type 1 error rate of α, that is, the sets of values of $\hat{\psi}_n$, which lead to the rejection of the hypotheses H_{01} and H_{02}, are, respectively, given by

$$\hat{\psi}_n > \psi_1 + z_\alpha\sqrt{\mathrm{Var}(\hat{\psi}_n)}, \tag{C13}$$

$$\hat{\psi}_n < \psi_2 - z_\alpha\sqrt{\mathrm{Var}(\hat{\psi}_n)}. \tag{C14}$$

Let us denote the regions given in Equations C13 and C14 by $\mathcal{R}_1$ and $\mathcal{R}_2$, respectively.

Notice that the set of values of ψ specified under the null hypotheses H_0 of nonequivalence can be expressed as the union of the sets of values of ψ under H_{01} and H_{02}, respectively. This is equivalent to saying that hypothesis H_0 is the union of hypotheses H_{01} and H_{02}. Similarly, the alternative hypothesis H_A is the intersection of the

(*Appendices continue*)

hypotheses H_{A1} and H_{A2}. These facts entail that the rejection region of a test at level α for testing H_0 is the intersection of the rejection regions $\mathcal{R}_1$ and $\mathcal{R}_2$. Accordingly, we say that the null hypothesis H_0 of nonequivalence is rejected at a level α if

$$\psi_1 + z_\alpha\sqrt{\mathrm{Var}(\hat{\psi}_n)} < \hat{\psi}_n < \psi_2 - z_\alpha\sqrt{\mathrm{Var}(\hat{\psi}_n)}, \tag{C15}$$

where $\hat{\psi}_n$ is the usual estimator of ψ, and z_α is the upper 100α percentile point of the standard normal distribution.

The power of this test at a given value of ψ specified under the alternative hypotheses H_A ($\psi_1 < \psi < \psi_2$) is the probability of rejecting the null hypotheses H_0 when ψ is the true value of the contrast. Using the normal approximation given in Equation A4, the power of the test at ψ, say, $P(\psi)$, is approximately given by

$$\begin{aligned} P(\psi) &= P(\psi_1 + z_\alpha\sqrt{\mathrm{Var}(\hat{\psi}_n)} < \hat{\psi}_n < \psi_2 - z_\alpha\sqrt{\mathrm{Var}(\hat{\psi}_n)}) \\ &= P\left(z_\alpha - \frac{\psi-\psi_1}{\sqrt{\mathrm{Var}(\hat{\psi}_n)}} < \frac{\hat{\psi}_n-\psi}{\sqrt{\mathrm{Var}(\hat{\psi}_n)}} < -z_\alpha + \frac{\psi_2-\psi}{\sqrt{\mathrm{Var}(\hat{\psi}_n)}}\right) \\ &= \Phi\left(-z_\alpha + \frac{\psi_2-\psi}{\sqrt{\mathrm{Var}(\hat{\psi}_n)}}\right) - \Phi\left(z_\alpha - \frac{\psi-\psi_1}{\sqrt{\mathrm{Var}(\hat{\psi}_n)}}\right). \end{aligned} \tag{C16}$$

In line with Luh & Guo (2016), for finding the sample sizes required to attain a prespecified statistical power $(1-\beta)$ by a level α test at a fixed alternative ψ ($\psi_1 < \psi < \psi_2$), we consider the following two cases separately, $\psi = (\psi_1+\psi_2)/2$ and $\psi \neq (\psi_1+\psi_2)/2$.

For $\psi = (\psi_1+\psi_2)/2$, the sample size n_1 is obtained by considering the equation:

$$\Phi\left(-z_\alpha - \frac{\psi_2 - [(\psi_2+\psi_1)/2]}{\sqrt{\mathrm{Var}(\hat{\psi}_n)}}\right) = 1-\beta/2 = \Phi(z_{\beta/2}), \tag{C17}$$

or equivalently,

$$\Phi\left(z_\alpha - \frac{[(\psi_2+\psi_1)/2]-\psi_1}{\sqrt{\mathrm{Var}(\hat{\psi}_n)}}\right) = \beta/2 = \Phi(z_{1-\beta/2}), \tag{C18}$$

which leads to

$$n_1 = \frac{(z_\alpha + z_{\beta/2})^2}{(\psi_2-\psi_1)^2/4}\left(\sum_{i=1}^{K} c_i^2\sigma_i^2/\gamma_i\right). \tag{C19}$$

Thus, the required sample sizes for different groups are given by $n_i = n_1\gamma_i$, $i = 1, 2, \ldots, k$ with $\gamma_1 = 1$.

For $\psi \neq (\psi_1+\psi_2)/2$, we can have the approximate group size (e.g., Luh & Guo, 2016)

$$n_1 = \frac{(z_\alpha + z_\beta)^2}{\eta^2}\left(\sum_{i=1}^{K} c_i^2\sigma_i^2/\gamma_i\right), \tag{C20}$$

where $\eta = \min(\psi-\psi_1, \psi_2-\psi)$, and consequently $n_i = n_1\gamma_i$, $i = 1, 2, \ldots, k$ with $\gamma_1 = 1$. Notice that $z_{\beta/2}$ in Equation C19 is replaced by z_β in Equation C20. This is a consequence of approximating $\Phi(-z_\alpha + [(\psi_2-\psi)/\sqrt{\mathrm{Var}(\hat{\psi}_n)}])$ by 1 for $\psi < (\psi_1+\psi_2)/2$, and $\Phi(z_\alpha - [(\psi-\psi_1)/\sqrt{\mathrm{Var}(\hat{\psi}_n)}])$ by 0 for $\psi > (\psi_1+\psi_2)/2$ in Equation C16. Except for ψ lying close to $\frac{\psi_1+\psi_2}{2}$, these approximations seem to work well.

Given the contrast coefficients c_i's, level α, achieved minimized variance (V_{oN}), and fixed values of ψ, ψ_1, and ψ_2, the achieved power is

$$\Phi\left(-z_\alpha + \frac{\psi_2-\psi}{\sqrt{V_{oN}}}\right) - \Phi\left(z_\alpha - \frac{\psi-\psi_1}{\sqrt{V_{oN}}}\right). \tag{C21}$$

Appendix D

Theorem D1

Theorem D1. *For the minimum sample size $m_0 = 2$, the stopping rule in Equation 18 yields that, on average, the final sample size of the populations is asymptotically the same as the corresponding optimal sample size under some assumptions. That is, for a high total sampling cost, $A_0 \to \infty$), $E(N_i/n_{i0}) \to 1$, provided that for all the populations $E[s^2_{iN_i}] < \infty$* and *$E[s^2_{ino}] < \infty$.*

Proof. In order to prove the theorem, we first need to prove the following lemma.

Lemma D1. *If the distribution(s) is(are) such that* $\mathrm{E}[s^2_{\mathrm{in}}]$ *exists, then the stopping rule in Equation 18 enjoys the following property*:

$$\frac{N_i}{n_{i0}} \xrightarrow{\text{a.s.}} 1 \text{ as } \mathrm{A}_0 \to \infty, \tag{D1}$$

where $\xrightarrow{a.s}$ indicates almost sure convergence.

Proof. Using Equation 18 yields

$$\frac{A_0|c_i|s_{iN_i}}{\sum_{l=1}^{K}|c_l|s_{lN_l}\sqrt{a_l a_i}} \le N_i \le m + \frac{A_0|c_i|s_{i(N_i-1)}}{\sum_{l=1}^{K}|c_l|s_{l(N_l-1)}\sqrt{a_l a_i}} \quad \text{a.s.} \tag{D2}$$

Now, $N_i \xrightarrow{\text{a.s.}} \infty$ as $A_0 \to \infty$. Now for each $i = 1, 2, \ldots, K$, since $A_0 \to \infty$, $s^2_{in_{i0}} \xrightarrow{\text{a.s.}} \sigma_i^2$. Using Theorem 2.1 of Gut (2009), $s^2_{iN_i} \xrightarrow{\text{a.s.}} \sigma_i^2$ and $s^2_{i(N_i-1)} \xrightarrow{\text{a.s.}} \sigma_i^2$ as $A_0 \to \infty$. Hence, dividing all sides of Equation D2 by n_{i0}, we prove $N_i/n_{i0} \xrightarrow{\text{a.s.}} 1$ for larger total sampling cost.

(*Appendices continue*)

Now, $N_i \geq m$ with probability 1, dividing Equation D2 by n_{i0} yields

$$N_i/n_{i0} - m/n_{i0} \leq \sup_{A_0>0} \frac{s_{i(N_i-1)} \sum_{l=1}^{K} |c_l| \sigma_l \sqrt{a_l}}{\sigma_i \sum_{l=1}^{K} |c_l| s_{l(N_l-1)} \sqrt{a_l}} \quad \text{a.s.} \tag{D3}$$

Provided, $\mathrm{E}[\sup_{A_0>0} s_{i(N_i-1)} / \sum_{l=1}^{K} |c_l| s_{l(N_l-1)} \sqrt{a_l}] < \infty$ for all $i = 1, 2, \ldots, K$ and $N_i/n_{i0} - m/n_{i0} \xrightarrow{\text{a.s.}} 1$ as $A_0 \to \infty$. Thus, the sequence $\{N_i/n_{i0} - m/n_{i0}\}$ converges a.s. 1 and is dominated by the integrable function by $\sup_{A_0>0} \{s_{i(N_i-1)} \sum_{l=1}^{K} |c_l| \sigma_l \sqrt{a_l}\} / \{\sigma_i \sum_{l=1}^{K} |c_l| s_{l(N_l-1)} \sqrt{a_l}\}$ in the sense that Equation D3 holds. Using the dominated convergence theorem (for details, we refer to Chattopadhyay & De, 2016; Darku & Chattopadhyay, 2021), we conclude that for large total sampling cost, $E(N_i/n_{i0}) \to 1$.

Received January 11, 2022
Revision received January 5, 2023
Accepted January 9, 2023 ■