Abstract:

Keywords: Testing for equivalence; Testing for superiority/non-inferiority

1. Introduction

The power of a test increases as the number of replications increase as this reduces the variability. Also, increase in replication size incurs larger sampling cost. ¡ proposed a procedure which can used to find the optimal replication size and read depth so as to attain the maximum power of a test in a case control setup with the variability thought to be from three sources: biological, technical imprecision and Poisson variance under a fixed cost. The procedure adopted by ¡ is a two-stage procedure where the optimal sizes are obtained using pilot sample. The two-stage procedure falls in the domain of sequential analysis area, for details regarding the sequential analysis area, we refer to ¡ and others. The two-stage procedure is less economical than the group-sequential sequential procedure in terms of optimal size (for e.g. refer to Ghosh and Sen, ()) as a two-stage procedure is based on smaller sample size than the purely sequential or group sequential procedures. Also, for testing for a large number of genes belonging to more than two cell types ¡ cannot be implemented as increasing replication size for each tested hypothesis will result in variable replication size. In this article, we maximize the power of the superiority or inferiority tests related to genes belonging to more than two cell types under a cost constraint considering a fixed read depth using purely sequential or group sequential procedure.

2. Tests for equivalence and Non-inferiority/Superiority

For the i^{th} group suppose $X_{i1}, \ldots, X_{in_i}$ be independent and identically distributed random variables, not necessary normal, with means μ_i and variances σ_i^2, thus in total there are $n = \sum_{i=1}^{K} n_i$. Let us consider a parameter, which is a function of K population means, $\delta = \mathbf{c}'\boldsymbol{\mu}\left(\sum_{i=1}^{K} c_i\mu_i\right)$. $\mathbf{c} = (c_1, \ldots, c_K)$ are known coefficients.

An estimator of the parametric function δ is $\widehat{\delta}_n = \mathbf{c}'\widehat{\boldsymbol{\mu}}_n$, where $\widehat{\boldsymbol{\mu}}_n = (\bar{X}_1, \ldots, \bar{X}_K)'$, a vector of group means. If γ_i's are sample size allocation ratios then, $n_i = \gamma_i n_1$ for $j = 1, \ldots, K$ with $\gamma_1 = 1$. The variance of $\widehat{\delta}_n$ is $Var(\widehat{\delta}_n) = \sum_{i=1}^{K} c_i^2\sigma_i^2/n_i = \frac{1}{n_1}\sum_{i=1}^{K} c_i^2\sigma_i^2/\gamma_i$.Thus for not so small n_1, using central limit theorem, we have

$$\frac{\sqrt{n_1}(\widehat{\delta}_n - \delta)}{\sqrt{\sum_{i=1}^{K} c_i^2\sigma_i^2/\gamma_i}} \underset{\mathcal{D}}{\to} N(0,1). \tag{2.1}$$

Using the central limit theorem, we can test claims regarding the parametric function δ without the assumption of normality of the data.

2.1. Testing for equivalence

Suppose, we wish to test hypotheses for equivalence by comparing δ, to two equivalence margins with $\delta_1 < 0 < \delta_2$ (see page 81, Lehmann and Romano, (); Guo et al., (); Luh and Guo, ()). Thus the null and alternative hypotheses are:

$$H_0 : \delta \leq \delta_1, \text{ or } \delta \geq \delta_2 \text{ against } H_A : \delta_1 < \delta < \delta_2 \tag{2.2}$$

In many studies, such as educational testing and drug testing, often we show equivalence of treatments. Then the null hypothesis H_0, defined in Equation 2.2, is rejected at $\alpha\%$ level of significance if,

$$\delta_1 + z_\alpha\sqrt{Var(\widehat{\delta}_n)} < \widehat{\delta}_n < \delta_2 + z_\alpha\sqrt{Var(\widehat{\delta}_n)}, \tag{2.3}$$

where, z_α is the upper α point of standard normal distribution. The approximate power function of the corresponding test of equivalence is given by,

$$P_{TE}(\delta) = \Phi\left(-z_\alpha + \frac{\delta_2 - \delta}{\sqrt{Var(\widehat{\delta}_n)}}\right) - \Phi\left(z_\alpha - \frac{\delta - \delta_1}{\sqrt{Var(\widehat{\delta}_n)}}\right) \tag{2.4}$$

2.2. Testing for superiority/inferiority

Suppose, we wish to test hypotheses for superiority (or inferiority) by comparing δ, to δ_0 (see page 81, Lehmann and Romano, (), Guo et al., ()). provided that the larger value of the response variable is better, the null and alternative hypotheses are:

$$\begin{aligned} &\text{(Inferiority) } H_0 : \delta \leq -|\delta_0| \text{ against } H_A : \delta > -|\delta_0| \\ &\text{(Superiority) } H_0 : \delta \geq -|\delta_0| \text{ against } H_A : \delta > -|\delta_0| \end{aligned} \tag{2.5}$$

Let us consider the inferiority test. Then the null hypothesis H_0, defined in Equation 2.5, is at $\alpha\%$ level of siginificance rejected if,

$$\widehat{\delta}_n > -|\delta_0| + z_\alpha \sqrt{Var(\widehat{\delta}_n)} \tag{2.6}$$

where, z_α is the upper α oint of standard normal distribution. The approximate power function of the corresponding inferiority test is given by,

$$P_{TI}(\delta) = \Phi\left(-z_\alpha + \frac{\delta + |\delta_0|}{\sqrt{Var(\widehat{\delta}_n)}}\right) \tag{2.7}$$

2.3. Sample Size Allocation

In experimental research, limited funding is allotted beforehand to carry out the sampling process. Thus under a cost constraint we have to carry out the test of equivalence or inferiority (or supe-

riority) with minimum errors. Note that we fix the probability of type I error to α, so the idea is to minimize the probability of type II error or equivalently maximizing the power under cost constraints. Recall from Equations 2.4 and 2.7, in test of equivalence and test of inferiority (or superiority), the power gets maximized if $Var(\widehat{\delta}_n)$ gets minimized.

Let us consider that a fixed amount $\$A_0$ is allotted for the sampling process and also consider that $a_1, a_2, \ldots, a_K$ denotes the cost of sampling one observations in groups $1, 2, \ldots, K$ respectively. Thus, $A_0 = \sum_{i=1}^{K} a_i n_i = n_1 \sum_{i=1}^{K} a_i \gamma_i$. Thus, the allotted sample size for group 1 under the cost constraint is

$$n_1 = \frac{A_0}{\sum_{i=1}^{K} a_i \gamma_i} \tag{2.8}$$

Using the expression in Equation 2.8, the variance of $\widehat{\delta}_n$ is

$$Var(\widehat{\delta}_n) = \frac{1}{A_0} \left(\sum_{i=1}^{K} \frac{c_i^2 \sigma_i^2}{\gamma_i} \right) \left(\sum_{i=1}^{K} a_i \gamma_i \right), \tag{2.9}$$

which will be minimized under the optimal allocation ratio

$$\gamma_i = \frac{|c_i|\sigma_i}{|c_1|\sigma_1} \sqrt{\frac{a_1}{a_i}}. \tag{2.10}$$

Thus with the the optimal allocation ratio in Equation 2.10, the optimal sample sizes for the K groups are,

$$n_{io} = \frac{A_0 |c_i| \sigma_i}{\sum_{l=1}^{K} |c_l| \sigma_l \sqrt{a_l a_i}}, \text{ for } i = 1, \ldots, K. \tag{2.11}$$

and the minimized variance of $\widehat{\delta}_n$ under the optimal allocation ratio is

$$V_0 = \frac{1}{A_0} \left(\sum_{i=1}^{K} |c_i| \sigma_i \sqrt{a_i} \right)^2. \tag{2.12}$$

Thus, maximum power is attained, under the cost constraint A_0, if the minimum variance defined in Equation 2.12 is attained or in other words if the optimal sample allocation ratio is used. However the optimal allocation ratio as defined in Equation 2.10 depends on population variances of the K groups. In practice, the values of population variances of each group are unknown, hence we cannot compute the allocation ratio and thereby the sample size required to obtain maximum power of the test. In the next section, we develop a new approach to avoid using a potentially poor estimate of sample size, often obtained by using supposed population standard deviations, which are potentially poor estimates to plan sample size. Our method ensures that we are informed by actual data from the population of interest.

3. Sample Size Determination under Independent Case

As opposed to use of supposed population values, the new approach that we momentarily develop, do not fix the sample sizes in advance, rather we estimate in stages. In statistics, this procedure is called sequential procedure. In sequential procedure, in the first stage, a small sample called a pilot sample is observed and then the parameter(s) are estimated to check a pre-defined condition in a pre-specified rule, known as the stopping rule. Further sampling of observations is carried out if the pre-defined condition is not met and further sampling is stopped once the pre-defined condition is satisfied. At a particular stage, if the pre-defined condition is not met, the researcher collects one or more additional observation(s) and then estimates the parameter of interest. This process is repeated until the pre-defined condition is met. For details about the general theory of sequential estimation procedures, we refer interested readers to Sen (()), Ghosh and Sen (()) and others.

As discussed, the optimal sample sizes for the K groups, n_{io} for $i = 1, \ldots, K$, is unknown because it depends on corresponding population standard deviations σ_i, which is itself unknown in practice. Thus, in order to estimate n_{io}, an estimator of σ_i is required, henceforth we will use $s_{iN_{io}}$ as an estimator of the population standard deviation based on N_{io} observations drawn from each of the K groups, where, N_{io} is the estimated optimal final sample size for the i^{th} group. We now develop an algorithm to find an estimate of the optimal sample size via the sequential estimation

procedure.

Stage 1: First, $m(\geq 2)$ randomly selected observations are collected from each of the K groups, often called pilot sample size and set $N_{io} = m$ for each of the K groups. From these m observations from each of the K groups, we estimate the corresponding population standard deviations. Let us define, s_{i1}^2 as the sample standard deviation computed from m observations collected from the i^{th} group in the first stage. If for the i^{th} group $(i = 1, \ldots, K)$, $m \geq \frac{A_0|c_i|s_{iN_{io}}}{\sum_{l=1}^{K}|c_l|s_{lN_{lo}}\sqrt{a_l a_i}}$, stop sampling from the i^{th} group and set the sample size for the i^{th} group equal to $N_{io} = m$. If for the i^{th} group $(i = 1, \ldots, K)$, $m < \frac{A_0|c_i|s_{iN_{io}}}{\sum_{l=1}^{K}|c_l|s_{lN_{lo}}\sqrt{a_l a_i}}$ then proceed to the next step.

Stage 2: Obtain $m^{'}$ additional observations from the remaining groups, where we set $m^{'} = 1$ in general drawn from a particular group. Thus there are $N_{io} = m + m^{'}$ observations from each of the remaining groups (i.e., the pilot sample size and an additional $m^{'}$ observations per group, for a total sample size of $\sum_{i=1}^{K} N_{io}$). If $m + m^{'} \geq \frac{A_0|c_i|s_{iN_{io}}}{\sum_{l=1}^{K}|c_l|s_{lN_{lo}}\sqrt{a_l a_i}}$ stop further sampling and set the final sample size equal to $\sum_{i=1}^{K} N_{io}$. If $m + m^{'} < \frac{A_0|c_i|s_{iN_{io}}}{\sum_{l=1}^{K}|c_l|s_{lN_{lo}}\sqrt{a_l a_i}}$ then continue the sampling process by sampling $m^{'}$ more observations per group.

This data collecting process after the pilot sample stage continues until the sample size condition is satisfied, that is, $N_{io} \geq \frac{A_0|c_i|s_{iN_{io}}}{\sum_{l=1}^{K}|c_l|s_{lN_{lo}}\sqrt{a_l a_i}}$. At this stage, we stop further sampling and report that the final sample size is $N_o = \sum_{i=1}^{K} N_{io}$.

At each stage of the algorithm, we check whether the sample size collected up to that stage is at least as large as the estimated value of n_{io} using observations from the i^{th} group collected until that stage. Based on the algorithm just outlined, a sampling stopping rule can be defined as follows:

$$N_{io} \text{ is the smallest integer } n_i(\geq m) \text{ such that } n \geq \frac{A_0|c_i|s_{in_i}}{\sum_{l=1}^{K}|c_l|s_{ln_l}\sqrt{a_l a_i}}, \tag{3.1}$$

The total cost of sampling $\sum_{i=1}^{K} N_{io}$ observations, N_{io} being the estimated final optimal sample

size for i^{th} group computed using Equation 3.1, is A_0. The proof is a special case of lemma 7.1

4. Tests for a Set of Genes

For the i^{th} group suppose $X_{ij1}, \ldots, X_{ijn_i}$ be indpedent and identically distributed random variables, not necessary normal, with means μ_{ij} and variances σ^2_{ij}, thus in total there are $n = \sum_{i=1}^{K} n_i$. So, the function of K population means, $\delta_j = \mathbf{c}'\boldsymbol{\mu_j}\left(\sum_{i=1}^{K} c_i\mu_{ij}\right)$. $\mathbf{c} = (c_1, \ldots, c_K)$ are known coefficients.

An estimator of the parametric function δ_j is $\widehat{\delta}_{nj} = \mathbf{c}'\widehat{\boldsymbol{\mu}}_{nj}$, where $\widehat{\boldsymbol{\mu}}_{nj} = (\bar{X}_{1j}, \ldots, \bar{X}_{Kj})'$, a vector of group means for $j = 1, \ldots, M$. The variance of $\widehat{\delta}_{nj}$ is $Var(\widehat{\delta}_{nj}) = \sum_{i=1}^{K} c_i^2\sigma_{ij}^2/n_i$.Thus for large n_i's, using central limit theorem, we have

$$\frac{(\widehat{\delta}_{nj} - \delta_j)}{\sqrt{\sum_{i=1}^{K} c_i^2\sigma_{ij}^2/n_i}} \underset{\mathcal{D}}{\to} N(0,1) \text{ for } j = 1, \ldots, M. \tag{4.1}$$

Using the central limit theorem, we can test claims regarding the parametric function δ_j without the assumption of normality of the data.

Suppose, we wish to test hypotheses for equivalence by comparing δ_j, $j = 1, 2, ..., M$ for a set of M genes for K cell types, δ_j for $j = 1, 2, ..., M$, to two equivalence margins with $\delta_{01} < 0 < \delta_{02}$. Thus the null and alternative hypotheses for the test of equivalence are:

$$H_{0j} : \delta_j \leq \delta_{01j}, \text{ or } \delta_j \geq \delta_{02j} \text{ against } H_{Aj} : \delta_{01j} < \delta_j < \delta_{02j} \text{ for } j = 1, 2, ..., M. \tag{4.2}$$

Then the null hypothesis H_{0j}, defined in Equation 2.2, is rejected at $\alpha\%$ level of siginificance if,

$$\delta_{01j} + z_\alpha\sqrt{Var(\widehat{\delta}_{nj})} < \widehat{\delta}_{nj} < \delta_{02j} + z_\alpha\sqrt{Var(\widehat{\delta}_{nj})} \text{ for } j = 1, 2, ..., M, \tag{4.3}$$

where, z_α is the upper α point of standard normal distribution. The approximate power function of

the corresponding test of equivalence is given by,

$$P_{TE}(\delta_j^*) = \Phi\left(-z_\alpha + \frac{\delta_{02j} - \delta_j^*}{\sqrt{Var(\widehat{\delta}_{nj})}}\right) - \Phi\left(z_\alpha - \frac{\delta_j^* - \delta_{01j}}{\sqrt{Var(\widehat{\delta}_{nj})}}\right) \text{ for } j = 1, 2, ..., M. \quad (4.4)$$

The null and alternative hypotheses for the Superiority or Inferiority are:

$$\begin{aligned} &\text{(Inferiority) } H_{0j} : \delta_j \leq -|\delta_{0j}| \text{ against } H_A : \delta_j > -|\delta_{0j}| \\ &\text{(Superiority) } H_{0j} : \delta_j \geq -|\delta_{0j}| \text{ against } H_A : \delta_j > -|\delta_{0j}| \text{ for } j = 1, 2, ..., M. \end{aligned} \quad (4.5)$$

Let us consider the inferiority test. Then the null hypothesis H_{0j}, defined in Equation 2.5, is at $\alpha\%$ level of siginificance rejected if,

$$\widehat{\delta}_{nj} > -|\delta_0 j| + z_\alpha \sqrt{Var(\widehat{\delta}_{nj})} \text{ for } j = 1, 2, ..., M \quad (4.6)$$

where, z_α is the upper α point of standard normal distribution. The approximate power function of the corresponding inferiority test is given by,

$$P_{TI}(\delta_j^*) = \Phi\left(-z_\alpha + \frac{\delta_j^* + |\delta_{0j}|}{\sqrt{Var(\widehat{\delta}_{nj})}}\right) \text{ for } j = 1, 2, ..., M. \quad (4.7)$$

4.1. Sample Size Allocation

Our goal is to carry out the test of equivalence or inferiority (or superiority) with maximum power under the cost constraint. Recall from Equations 4.4 and 4.7, in test of equivalence and test of inferiority (or superiority), the power of a test gets maximized if $Var(\widehat{\delta}_{nj})$ gets minimized for each of the M genes. Minimizing variance or increasing sample size for each tested hypothesis will result in variable sample size. This is not desired as for a particular experiment we get observations for the whole set of M genes. So, instead of maximizing the power for each test, we maximize the sum of the power of tests corresponding to the whole set of M genes, $\sum_{j=1}^{M} P_{TI}(\delta_j^*)$.

Let us consider that a fixed amount $\$A_0$ is alloted for the sampling process and also consider that $a_1, a_2, \ldots, a_K$ denotes the cost of sampling one observation in cell types $1, 2, \ldots, K$ respectively. Thus, $A_0 = \sum_{i=1}^{K} a_i n_i$. Using Equation 4.7, the sum of powers for inferiority test for a set of M genes is given by,

$$SP(\delta_j^*) = \sum_{j=1}^{M} P_{TI}(\delta_j^*) = \sum_{j=1}^{M} \Phi\left(-z_\alpha + \frac{\delta_j^* + |\delta_{0j}|}{\sqrt{\sum_{i=1}^{K} \frac{c_i^2 \sigma_{ij}^2}{n_i}}}\right) \text{ for } j = 1, 2, ..., M. \tag{4.8}$$

Thus the optimal sample sizes for the K cell types can be found by maximizing the sum of power for inferiority test for a set of M genes, subject to the cost constraint $A_0 = \sum_{i=1}^{K} a_i n_i$ is given by

$$n_{io} = \underset{n_2,\ldots,n_K}{argmax} SP(\delta_j^*), \text{ for } i = 1, \ldots, K. \tag{4.9}$$

We note that, n_1 depends on $n_2, n_3, ..., n_K$. Thus, using the optimal sample size found in Equation 4.9, maximum power can be attained, under the cost constraint A_0, if we know the population variances related to the M genes in the K cell types. In practice, the values of population variances of each group are unknown, hence we cannot compute the optimal sample sizes by directly maximizing $SP(\delta_j^*)$. In the next section, we develop a new approach to avoid using a potentially poor estimate of sample size, often obtained by using supposed population standard deviations, which are potentially poor estimates to plan sample size. Our method ensures that we are informed by actual data from the population of interest.

4.2. Sample Size Determination under Independent Case

As discussed in section 3, in this subsection we propose a sequential procedure to estimate the optimal number of cells required obtain the maximum sum of powers for testing M hypotheses. As noted earlier the optimal number of cells depends on corresponding population standard deviations σ_{ij}, which is itself unknown in practice. Thus, in order to estimate n_{io}, an estimator of σ_{ij} is required, henceforth we will use $s_{ijN_{io}}$ as an estimator of the population standard deviation of the

relative abundance of the j^{th} gene in the i^{th} cell type based on N_{io} cells, where, N_{io} is the estimated optimal sequenced cells required from all M genes for the i^{th} cell type. We now develop an algorithm to find an estimate of the optimal number of sequenced cells via the sequential estimation procedure.

Stage 1: First, $m(\geq 2)$ sequenced cells are collected from all M genes belonging to all the K cell types, often called pilot sample size and set $N_{io} = m$ for each of the K cell types. From these m relative abundances related to the M genes in each of K cell types, we estimate the corresponding population standard deviations. Let us define, s^2_{ij1} as the sample standard deviation computed from m relative abundances related to the j^{th} gene in the i^{th} cell type in stage 1 and thereby maximize the $SP(\delta^*_j)$ by using s^2_{ij1}.

If for the i^{th} cell type ($i = 1, \ldots, K$), having observed m relative abundances related to the j^{th} gene in the i^{th} cell type, we find $SP(\delta^*_j)$ in terms of $n_2, ..., n_K$, suppose that be denoted by $SP(\delta^*_{s1j})$ and find

$$\widehat{m}_{i1} = \underset{T_1}{argmax} SP(\delta^*_{s1j}), \text{ for } i = 1, \ldots, K. \tag{4.10}$$

where, the set $T_1 = \{n_1, n_2, ..., n_K\}$. Then we compare $m \geq \widehat{m}_{i1}$, stop sampling from the i^{th} cell type and set the sample size for the i^{th} cell type equal to $N_{io} = m$. If for the i^{th} cell type ($i = 1, \ldots, K$), $m < \widehat{m}_{i1}$ then proceed to the next step.

Stage 2: Obtain $m^{'}$ additional observations from the remaining cell types, where we set $m^{'} = 1$ in general drawn from a particular cell type. Thus there are $N_{io} = m + m^{'}$ observations from each of the remaining cell types (i.e., the pilot sample size and an additional $m^{'}$ observations per cell type, for a total sample size of $\sum_{i=1}^{K} N_{io}$). Next, we find $SP(\delta^*_j)$ in terms of $n_1, ..., n_K$ by using sample standard deviations computed from the $m + m^{'}$ observations for cell types for which $m < \widehat{m}_{i1}$ in stage 1 and we will use the sample standard deviations computed in stage 1 for cell types for

which $m \geq \widehat{m}_{i1}$ in stage 1. Suppose that the estimated sum of the power function be denoted by $SP(\delta^*_{s2j})$ and then find

$$\widehat{m}_{i2} = \underset{T_2}{argmax} SP(\delta^*_{s2j}), \text{ for } i = 1, \ldots, K. \tag{4.11}$$

where, the set T_2 is the set of all n_i's for which we had $m < \widehat{m}_{i1}$ in stage 1. Then we compare $m \geq \widehat{m}_{i2}$, stop sampling from the i^{th} cell type and set the sample size for the i^{th} cell type equal to $N_{io} = m + m^{'}$. If for the i^{th} cell type ($i = 1, \ldots, K$), $m < \widehat{m}_{i2}$ then proceed to the next step.

At each stage of the algorithm, we check whether the sample size collected up to that stage is at least as large as the estimated value of n_{io} using observations from the i^{th} group collected until that stage. Based on the algorithm just outlined, a sampling stopping rule can be defined as follows:

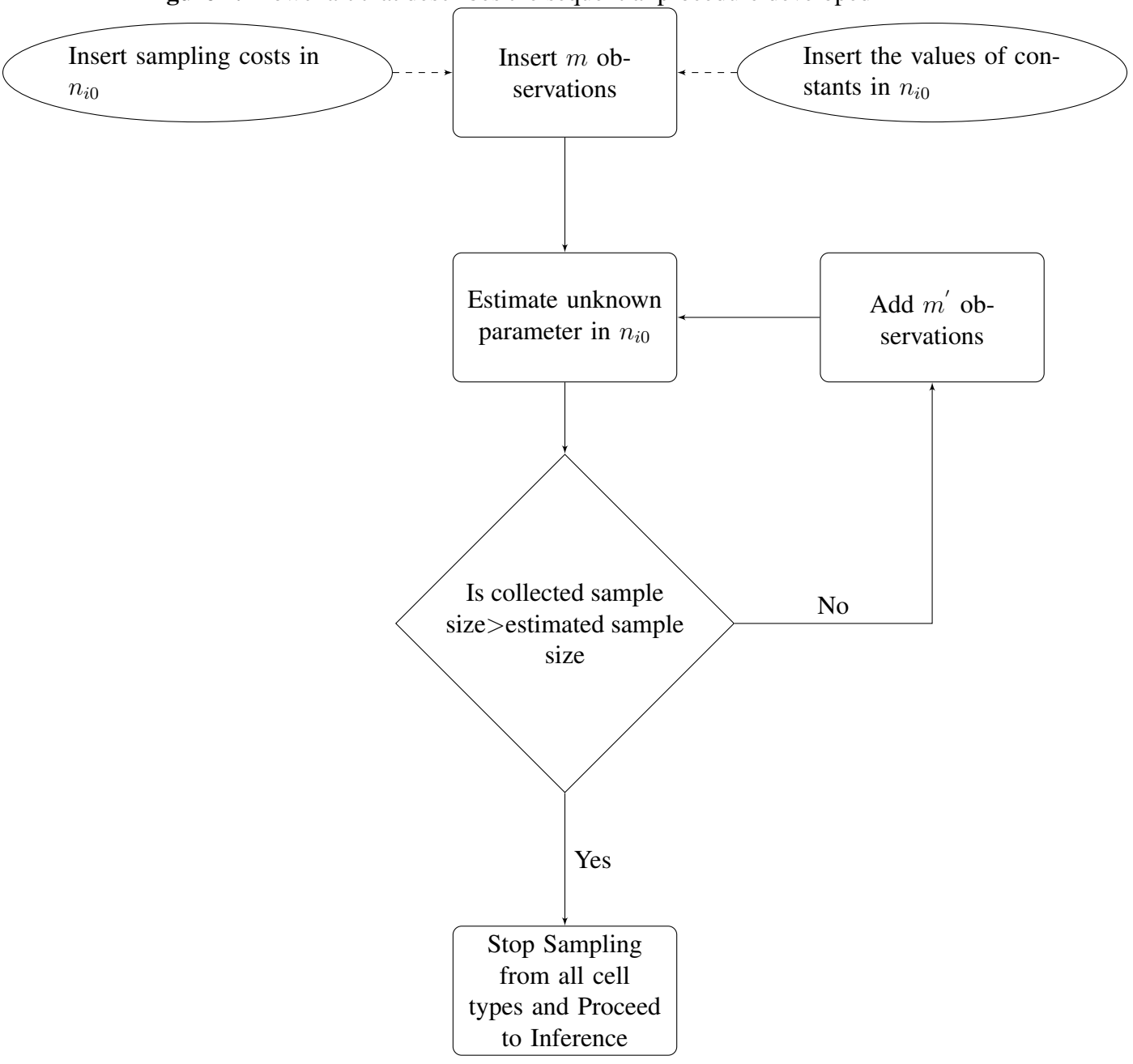
$$N_{io} \text{ is the smallest integer } n_i (\geq m) \text{ such that } n_i \geq argmax SP(\delta^*_{sj}), \text{ for } i = 1, \ldots, K. \tag{4.12}$$

5. Some Discussions

The methodology developed in this paper gives the estimated sample sizes from each group that can be drawn to achieve the maximum power in order to carry out the mentioned hypothesis testing problem with a cost constraint. Consider an example in which there are three groups and the cost of sampling from each group is same, say \$30 $(= a_i, i = 1, 2, 3)$. Suppose that the total amount $A_0 = \$100$ is allotted for the sampling process. Then, it can be deduced that $n = \sum_{i=1}^{3} n_i = 3.33$, which is not an integer. So, a realistic and objective assumption is that n, the total sample size, must be an integer.

In many biological laboratory experiments, to get measurements, often samples from several groups are fed into machines. These machines often have some number of experiments that can be performed in one go. Consider, in a particular situation, that M be the minimum number of

Figure 1. Flowchart that describes the sequential procedure developed



experiments that can be performed in one go. So, in the pilot sampling stage, for getting Km observations from $K(< M)$ groups (m from each group), we need to run Km experiments. So, it is better to consider a pilot sample size, i.e. $m = M/K$ or some integer multiple of M/K to make it an integer.

Now in the subsequent stages, we need to get $m^{'}$ observations from each group if the pre-defined condition in the algorithm is not satisfied. If at any stage $M >> m^{'}$, then we will consider, each of the remaining groups will get $m^{'}$ slots. If at some stage, except the pilot sampling stage, the cell types which require additional observations is greater than M then choose 1 slot for each group.

In general, the sample size is proportional to the variance and is inversely proportional to the sampling cost. So, if in a certain situation, it costs more to get a new observation, we select smaller number of observations and if there is too much variability in the data, we select more observations. Suppose at some stage, the number of groups requiring additional observations is less than M. Consider a situation, in which there are three remaining groups from which we need one more observation. Samples from these groups are fed into a machine in which M=5. Thus the first three slots will be filled by the samples from the three groups and the other two slots will be filled by the groups with higher sample standard-deviation to cost ratio than the group with the lowest ratio. Please note in such situations, we may exceed the total cost by a small margin.

In many situations there are some genes for which we don't want to continue hypothesis testing as the relative abundances collected upto a certain stage seems to behave arbitrarily. Then we just exclude the observations related to that gene for the remaining stages and then continue with our procedure as usual for other genes.

6. Conclusion

7. Appendix

Lemma 7.1. *The total cost of sampling* $\sum_{i=1}^{K} N_{io}$ *observations is* A_0.

Proof. We note that N_{io} is the estimated final optimal sample size for the i^{th} group. We know from Equation 3.1, that

$$\frac{A_0|c_i|s_{iN_{io}}}{\sum_{l=1}^{K}|c_l|s_{lN_{lo}}\sqrt{a_l a_i}} \leq N_{io} \leq \frac{A_0|c_i|s_{i(N_{io}-1)}}{\sum_{l=1}^{K}|c_l|s_{l(N_{lo}-1)}\sqrt{a_l a_i}} \tag{7.1}$$

Multiplying by a_i on all three sides in Equation 7.1 and summing them over we get,

$$\frac{\sum_{i=1}^{K} A_0 a_i|c_i|s_{iN_{io}}}{\sum_{l=1}^{K}|c_l|s_{lN_{lo}}\sqrt{a_l a_i}} \leq \sum_{i=1}^{K} a_i N_{io} \leq \frac{\sum_{i=1}^{K} A_0 a_i|c_i|s_{i(N_{io}-1)}}{\sum_{l=1}^{K}|c_l|s_{l(N_{lo}-1)}\sqrt{a_l a_i}}$$
$$\frac{\sum_{i=1}^{K} A_0\sqrt{a_i}|c_i|s_{iN_{io}}}{\sum_{l=1}^{K}\sqrt{a_l}|c_l|s_{ljN_{lo}}} \leq \sum_{i=1}^{K} a_i N_{io} \leq \frac{\sum_{i=1}^{K} A_0\sqrt{a_i}|c_i|s_{i(N_{io}-1)}}{\sum_{l=1}^{K}\sqrt{a_l}|c_l|s_{l(N_{lo}-1)}}$$

Thus, we can say using Sandwich theorem that $\sum_{i=1}^{K} a_i N_{io} = A_0$, the total cost of sampling $\sum_{i=1}^{K} N_{io}$ observations. □

References

Ghosh, B. K. and Sen, P. K. (1991), *Handbook of sequential analysis*, vol. 118, CRC Press.

Guo, J.-H., Chen, H. J., and 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, 439–461.

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

Luh, W.-M. and Guo, J.-H. (2015), "Sample size planning for the noninferiority or equivalence of a linear contrast with cost considerations." *Psychological Methods*, 21, 13–34.

Sen, P. K. (1981), *Sequential nonparametrics: Invariance principles and statistical inference*, Wiley New York.