# Research Report for 3-stage Procedure of Selecting the Best Population

## 1. Introduction

This report aims to introduce a three-stage selection procedure design for binomial data collected from k treatment groups $\pi_1, \pi_2, \pi_3 \ldots\ldots\ \pi_k$ and a control group $\pi_0,$ as well as to test the effectiveness of a this methodology. Such effectiveness is evaluated through a comparison with the traditional "two-stage selection & testing design" published by Thall, Simon and Ellenberg in 1988. Same with the 2-stage design, the purpose of the three-stage design is to select the group with the best population (the population with the largest parameter $p_{[k]}$). This report will present in detail the three-stage methodology which was initiated from the idea of combining the two-stage procedure and subset selection approach into one experiment containing three stages. Following the thorough introduction, formulas for numerical analysis will be presented with necessary proofs. In the fourth section, comparison results are shown by conducting numerical analysis. Numerical results will be given with comparison with the two-stage method for an equal or similar size of available sample data. Current limitedness and unsolved problems will be discussed in the last section. Due to the application of normal approximation, as well as this three stage method should be applied to the situation where sample size is allowed to be large.

This 3-stage design is based on two research papers—*Selecting a subset containing the best of several binomial populations* by Gupta and Sobel in 1960, and *two-stage selection and testing designs for comparative clinical* trials by Thall, Simon and Ellenberg in 1988. In the three stage procedure, subset selection is conducted in stage 1, an adjusted two-stage selection procedure is conducted in stage 2 and 3. P(CS), alpha level, and expected sample size (EN) are all redefined in the overall three stages. The final goal of this 3 stage design is to minimize the expected sample size subject to the constraints of P(CS) and Type I error rate alpha.

## 2. Literature review

Subset selection is brought into the first stage in this 3 stage methodology. Early research by Gupta and Sobel in 1960 has designed procedure R to select the subset of several binomial populations that successfully include the best population with pre-assigned probability P(CS). The procedure and P(CS) are defined for both equal sample size and unequal sample sizes. Expected size of selected subset is presented in both general configurations and the least favorable configuration.

In this report, the three-stage design is introduced for binomial populations with k treatment means and a control group. Such case is widely applied to clinical trials, where the parameter to be compared is usually denoted as the effectiveness of therapies, or mortality of drugs. Early methodologies date back to 1984 when Dunnett proposed a single stage selection procedure for the case of binomial(Bernoulli) populations, for which an arcsine transformation is used to transfer the case to a normal population with common known variance.

Later in 1988, Thall, Simon and Ellenberg published their methodology of "two-stage selection and testing design" using the same normalized data through arcsine transformation. This new approach at that time combines hypothesis testing with selection procedures, allowing an early termination of experiment at the first stage if the statistic defined in stage 1 fails to pass the cutoff value. As a result, this two stage method enjoys obvious advantage over Dunnett's method in reducing the expected sample size. Specifically, "even the maximum possible sample sizes of the corresponding two-stage designs are smaller than the fixed sample sizes of Dunnett's balanced designs in each case".

Specifically, it is worthy to be mentioned that the two stage design is best applied to situations where large sample size is allowed, due to the data transformation for the sake of normal approximation.

## 3. Methodology

In this part, a brief summary of the two papers' approach is introduced in section 3.1. The procedures of the three stages will be described in section 3.2 with necessary statistical proofs.

### 3.1 Brief summary of subset selection approach and 2-stage selection approach

In the 2-stage selection and testing design developed by Thall, Simon and Ellenberg in 1988, treatment v is selected in state 1 if the maximal statistic $T_1 := \max\left(\frac{\boldsymbol{z_{i1}-z_{01}}}{\sqrt{\boldsymbol{2}}}\right) > \boldsymbol{y_1}$ holds. Otherwise, terminate the experiment at this stage and claim null hypothesis to be true. Once treatment v is selected, bring that treatment to stage 2 together with the control, and take new samples from the two populations. In this second stage, the statistic $T_1$ becomes the combination of the two stages : $T_2 := \sqrt{\frac{n_1}{n_1+n_2}}\left(\frac{z_{v1}-z_{01}}{\sqrt{2}}\right) + \sqrt{\frac{n_2}{n_1+n_2}}\left(\frac{z_{v2}-z_{02}}{\sqrt{2}}\right)$, where both $\frac{z_{v1}-z_{01}}{\sqrt{2}}$ $and$ $\frac{z_{v2}-z_{02}}{\sqrt{2}}$ follow standard normal distribution, resulting in the statistic $T_2$ also follows standard normal distribution obviously. If $T_2 > y_2$, $H_0$ is rejected and conclude $p_v > p_0; if\ T_2 \le y_2, accept\ H_0$.

The least favorable configuration is defined as the configuration $\theta^*(p_{[1]} = p_{[2]} \dots = p_{[k-1]} = p_0 + \delta_1, p_{[k]} = p_0 + \delta_2)$ that minimizes $1 - \beta^*\ or\ P(CS)$. In the procedure, both stages are conducted under hypothesis testing, where $H_0$ is defined to be the configuration all $p_{[i]} = p_0$. Size α, the Type I error, is computed as the probability that any treatment is selected when $H_0$ is true. According to the authors, size α is computed by evaluating P(CS) under the configuration $H_0$, and multiplying the result by k. The expected sample size EN is derived as the averaged value under $H_0$ and $\theta^*$ as $\frac{1}{2}E(N|H_0) + \frac{1}{2}E(N|\theta^*)$. The numerical result is to choose the design parameters n1,n2,y1,y2 to minimize EN subject to the constraints imposed by specifying α and $1 - \beta^*$.

In the subset selection methodology by Gupta and Sobel in 1960, procedure R for equal sample sizes is designed to select population $\pi_i$ into the subset if and only if $x_i \ge x_{max} - d$, where d=d$(n_1, k, P_1^*)$ . The constant d is the smallest non-negative integer satisfying the P(CS) requirement: $\inf_{0\le p\le 1} P\{CS; p, d\} = \inf_{0\le p\le 1} \{\sum_{\alpha=0}^{n} \binom{n}{\alpha} p^\alpha (1-p)^{n-\alpha} \left[\sum_{j=0}^{\alpha+d} \binom{n}{j} p^j (1-p)^{n-j}\right]^{k-1}\} \ge P^*$. The expected size of selected subset is computed as the expectation of the summation of a chance variable Y, which denoted as an indicator function which equals 1 if $\pi_i$ is selected into the subset and 0 otherwise. Therefore for any given value of n,p, k, P*, $E\{S\} = E\{\sum_{i=1}^{k} Y_i\} = \sum_{i=1}^{k} E\{Y_i\} = \sum_{i=1}^{k} P(CS; p, d) = \sum_{x=0}^{n_i} \{ \binom{n}{x} (p+\delta)^x (q-\delta)^{n-x} \left[\sum_{j=0}^{x+d} \binom{n}{j} p^j q^{n-j}\right]^{k-1} \} + (k-1) \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} \left[\sum_{j=0}^{x+d} \binom{n}{j} p^j q^{n-j}\right]^{k-2} \left[\sum_{j=0}^{x+d} \binom{n}{j} (p+\delta)^j (q-\delta)^{n-j}\right]^{k-1}\}$. The expected proportion of populations retained in the selected subset by procedure R is presented in Table 3 of the 1960's paper, which will be used in the computation of EN in this 3-stage report.

### 3.2 Three-stage procedure design

The procedure described below is a combined application of Gupta and Sobel's subset selection approach and Thall, Simon and Ellenberg's 2-stage selection and testing design. The experiment's goal is to select the "best" population with comparison to a control group. Subset selection is applied in stage 1, ending up with m selected populations in the subset. Stage 2 and 3 corresponds to the traditional 2-stage procedure in 1988's methodology. For the sake of alignment throughout all three stages, the least favorable configuration $\theta^*$ is defined as: $p_{[1]} = p_{[2]} = \cdots = p_{[k-1]}: = p,\ p_{[k]}: = p + \delta$.

### 3.2.1 Three-stage selection procedure

**Stage 1: subset selection**

Take $n_1$samples from each of k treatments and the control group. Observe the number of successes $X_{is}$ $(i = 1,2 \dots k;\ s = 1,2,3)$ in each sample. Define $z_{is}$=$2\sqrt{n_i}a(\frac{X_{is}}{n_s})$, where $z_{is}$ follows standard normal distribution with mean $2\sqrt{n_i}a(p_i)$ and variance 1. The data and statistic in this stage is as follows.

| Population $\pi_i$ | $\pi_0$ | $\pi_1$ | $\pi_2$ | | $\pi_k$ |
|---|---|---|---|---|---|
| Observed #success $X_{i1}$ | $X_{01}$ | $X_{11}$ | $X_{21}$ | … | $X_{k1}$ |
| $z_{i1}$ $\coloneqq 2\sqrt{n_i}a(\frac{X_{i1}}{n_1})$ | $z_{01}$ $=\boldsymbol{2\sqrt{n_1}a(\frac{X_{01}}{n_1})}$ | $z_{11}$ $=\boldsymbol{2\sqrt{n_1}a(\frac{X_{11}}{n_1})}$ | $z_{21}$ $=\boldsymbol{2\sqrt{n_1}a(\frac{X_{21}}{n_1})}$ | … … | $z_{k1}$ $=\boldsymbol{2\sqrt{n_1}a(\frac{X_{k1}}{n_1})}$ |

<u>Selection procedure:</u>

If :

$$x_i \geq x_{max} - d,\ \text{ where d=d}(n_1, k, P_1^*)$$

Select the population i into the subset.

Conduct this procedure on all k treatment samples, and got selected subset $\{\pi_A, \pi_B, \pi_C \dots \pi_M\}$ containing m out of k treatment groups. Now enter stage 2 with m treatment groups.

**Stage 2: m choose 1 under hypothesis testing: $\boldsymbol{H_0(p_{[1]} = p_{[2]} = p_{[3]} \dots = p_{[k]} = p_0)}$**

Take $n_2$ samples from each of m treatments and the control group. $X_{i2}$ *and* $z_{i2}$ are defined in the same way as in stage 1. Below is the data and statistic for this stage.

| Population $\pi_i$ | $\pi_0$ | $\pi_A$ | | $\pi_i$ | | $\pi_M$ |
|---|---|---|---|---|---|---|
| Observed #success in stage 2 :=$X_{i2}$ | $X_{02}$ | $X_{A2}$ | … | $X_{i2}$ | … | $X_{M2}$ |
| $z_{i2}$ $\coloneqq 2\sqrt{n_i}a\left(\frac{X_{i1}}{n_1}\right)$ | $z_{02}$ $=\boldsymbol{2\sqrt{n_2}a(\frac{X_{02}}{n_2})}$ | $z_{A2}$ $=\boldsymbol{2\sqrt{n_2}a(\frac{X_{A2}}{n_2})}$ | … … | $z_{i2}$ $=\boldsymbol{2\sqrt{n_2}a(\frac{X_{i2}}{n_2})}$ | … … | $z_{M2}$ $=\boldsymbol{2\sqrt{n_2}a(\frac{X_{M2}}{n_2})}$ |

<u>Selection Procedure:</u>

If $T_2 = max\left(\frac{z_{i2}-z_{02}}{\sqrt{2}}\right) > y_2$, Reject $H_0$. denote $\pi_i$ as population v, *where* $\frac{z_{v2}-z_{02}}{\sqrt{2}} \sim n(0,1)$. Take Treatment V and control group to stage 3. Otherwise terminate.

*(Note: In this stage, we could have combine the result of both stage 1 and stage 2, as what we do in the procedure in stage 3 described below—combing data from all three stages by assigning weights to data in each stage. However, we chose to discard such stage combination for the sake of simplifying the calculation for randomization. Such randomization will be discussed more in detail in section 3. )*

**Stage 3: Double check the significance of treatment v under hypothesis testing:** $H_0\big(p_{[1]} = p_{[2]} = p_{[3]} \dots . = p_{[k]} = p_0\big)$

Take $n_3$ samples from the selected population v and the control group. $X_{i3}$ *and* $z_{i3}$ are defined in the same way as in stage 1. Below is the data and statistic for this stage.

| DATA | | | |
|---|---|---|---|
| Population $\pi_i$ | $\pi_0$ | $\pi_V$ | Under $H_0$: |
| Observed #success In stage 3 :=$X_{i3}$ | $X_{03}$ | $X_{V3}$ | |
| $z_{i3}$ $:= 2\sqrt{n_3}a\left(\frac{X_{i3}}{n_3}\right)$ | $z_{03}$ $=\mathbf{2}\sqrt{\boldsymbol{n_3}}\boldsymbol{a}(\frac{\boldsymbol{X_{03}}}{\boldsymbol{n_3}})$ | $z_{V3}$ $=\mathbf{2}\sqrt{\boldsymbol{n_3}}\boldsymbol{a}(\frac{\boldsymbol{X_{A3}}}{\boldsymbol{n_3}})$ | $z_{i3}{\sim}n(\mathbf{2}\sqrt{\boldsymbol{n_3}}\boldsymbol{a}(\boldsymbol{p_0}),\mathbf{1})$ |
| Recall stage 1 data | $X_{01}$ | $X_{V1}$ | |
| $z_{i1}$ $:= 2\sqrt{n_i}a\left(\frac{X_{i1}}{n_1}\right)$ | $z_{01}$ $=\mathbf{2}\sqrt{\boldsymbol{n_1}}\boldsymbol{a}(\frac{\boldsymbol{X_{01}}}{\boldsymbol{n_1}})$ | $z_{V1}$ $=\mathbf{2}\sqrt{\boldsymbol{n_1}}\boldsymbol{a}(\frac{\boldsymbol{X_{V1}}}{\boldsymbol{n_1}})$ | $z_{i1}{\sim}n(\mathbf{2}\sqrt{\boldsymbol{n_1}}\boldsymbol{a}(\boldsymbol{p_0}),\mathbf{1})$ |
| Recall stage 2 data | $X_{02}$ | $X_{V2}$ | |
| $z_{i2}$ $:= 2\sqrt{n_i}a\left(\frac{X_{i2}}{n_2}\right)$ | $z_{02}$ $=\mathbf{2}\sqrt{\boldsymbol{n_2}}\boldsymbol{a}(\frac{\boldsymbol{X_{02}}}{\boldsymbol{n_2}})$ | $z_{V2}$ $=\mathbf{2}\sqrt{\boldsymbol{n_2}}\boldsymbol{a}(\frac{\boldsymbol{X_{A2}}}{\boldsymbol{n_2}})$ | $z_{i2}{\sim}n(\mathbf{2}\sqrt{\boldsymbol{n_2}}\boldsymbol{a}(\boldsymbol{p_0}),\mathbf{1})$ |

Selection Procedure:

If :

$$\mathrm{T}_3 = \sqrt{\frac{n_2}{n_1+n_2+n_3}}\left(\frac{z_{v2}-z_{02}}{\sqrt{2}}\right) + \sqrt{\frac{n_1+n_3}{n_1+n_2+n_3}}\left(\frac{z_{v13}-z_{013}}{\sqrt{2}}\right) > y_3,$$

claim $\pi_V$ is the population with $p_{[k]}$.

Note:

$$\sqrt{\frac{n_2}{n_1+n_2+n_3}}\left(\frac{z_{v2}-z_{02}}{\sqrt{2}}\right) + \sqrt{\frac{n_1+n_3}{n_1+n_2+n_3}}\left(\frac{z_{v13}-z_{013}}{\sqrt{2}}\right) \sim \mathrm{n}(0,1) \text{ under } H_0 \;\; \langle 0\rangle,$$

where $z_{v13} = \sqrt{\frac{n_1}{n_1+n_3}}z_{v1} + \sqrt{\frac{n_3}{n_1+n_3}}z_{v3}$.

### 3.2.2 Statistical Formulas

In this session, the statistical formulas is presented in part (i). Statistical proof for each of the formula claimed in (a) will be presented in part (ii).

## (i) Formulas for numerical analysis

This 3-stage method is to choose the parameters $n_3, y_3$ that minimize EN (expected total sample size), subject to the constraints imposed by specifying α-level and P(CS). The formulas below are exactly the statistical expression of this approach:

Minimize EN: $= (k+1)n_1 + (\mathrm{EM}+1)n_2 + n_3(1-\Phi(y_2)) + n_3\left(1-\Phi\left(y_2-\sqrt{2n_2}\Delta_2\right)\right)$ $, where\ EM = Prop * k\ (Prop\ is\ given\ in\ Table\ 3\ of\ 1960's\ paper),$ <1>

such that:

$$1-\beta^* = P_{\theta^*}(CS) := P\left(\pi_{[k]} with\ p_{[k]} = p+\delta\ is\ selected\right) = \sum_{x_0=0}^{n_2}\sum_{x_k=0}^{n_2}\{1-\Phi(y_3\sqrt{\frac{n_1+n_2+n_3}{n_1+n_3}} - \frac{\sqrt{2}n_2}{\sqrt{n_1+n_3}}\left(a\left(\frac{x_k}{n_2}\right)-a\left(\frac{x_0}{n_2}\right)\right) - \sqrt{2(n_1+n_3)}\Delta_2)\}\ \{\sum_{j=0}^{EM-1}\frac{1}{j+1}\binom{EM-1}{j}\{b(x_k,n_2,p)\}^j\{B(x_k-1,n_2,p)\}^{EM-1-j}\}\ \{\mathrm{I}\left(a\left(\frac{x_k}{n_2}\right)-a\left(\frac{x_0}{n_2}\right) > \frac{y_2}{\sqrt{2n_2}}\right)\}\{\mathrm{b}\{x_k,\mathrm{p}+\delta,n_2\}\mathrm{b}\{x_0,p_0,n_2\}\}\ P(*)$$ <2>

and

$$\text{size } \alpha = \mathrm{K}\sum_{x_0=0}^{n_2}\sum_{x_k=0}^{n_2}\left\{1-\Phi\left\{y_3\sqrt{\frac{n_1+n_2+n_3}{n_1+n_3}} - \frac{\sqrt{2}n_2}{\sqrt{n_1+n_3}}\left(a\left(\frac{x_k}{n_2}\right) - a\left(\frac{x_0}{n_2}\right)\right)\right\}\right\}\left\{\sum_{j=0}^{EM-1}\frac{1}{j+1}\binom{EM-1}{j}\{b(x_k,n_2,p_0)\}^j\{B(x_k-1,n_2,p_0)\}^{EM-1-j}\right\} \mathrm{I}\left(a\left(\frac{x_k}{n_2}\right)-a\left(\frac{x_0}{n_2}\right) > \frac{y_2}{\sqrt{2n_2}}\right)\mathrm{b}\{x_k,p_0,n_2\}\mathrm{b}\{x_0,p_0,n_2\}\ P(*)$$ <3>

holds.

Specifically, EN in <1> is the expected sample size in three stages weighted under two configurations $H_0$ and least favorable configuration $\theta^*$, where the sample size in stage 1 ($\boldsymbol{n_1}$) and stage 2 ($\boldsymbol{n_2}$) are pre-assigned. Formula <2> computes the probability of correct selection of the three stage procedure. Here the P(CS) of state 1 has been pre-assigned to P(*), meaning the probability of selecting a subset that successfully containing the best population is required to be at least P(*). Size α in formula <3> is the Type I error of the three stage procedure, defined as the probability of selecting any of the k treatments when $H_0$ is true.

**(ii) Statistical Proof for <0><1><2><3>**

Claim <0>:

$$\sqrt{\frac{n_2}{n_1+n_2+n_3}}\left(\frac{z_{v2}-z_{02}}{\sqrt{2}}\right) + \sqrt{\frac{n_1+n_3}{n_1+n_2+n_3}}\left(\frac{z_{v13}-z_{013}}{\sqrt{2}}\right) \sim \mathrm{n}(0,1) \text{ under } H_0\ ,$$

where $z_{v13} = \sqrt{\frac{n_1}{n_1+n_3}}z_{v1} + \sqrt{\frac{n_3}{n_1+n_3}}z_{v3}$.

Proof:

Under $H_0$

Mean($z_{v13}$)= $\sqrt{\frac{n_1}{n_1+n_3}}2\sqrt{n_1}a(p_0) + \sqrt{\frac{n_3}{n_1+n_3}}2\sqrt{n_3}a(p_0)$,

Var($z_{v13}$)= $\frac{n_1}{n_1+n_3}var(z_{v1}) + \frac{n_3}{n_1+n_3}var(z_{v3}) = 1$;

Therefore, $z_{v13} \sim n(\sqrt{\frac{n_1}{n_1+n_3}}2\sqrt{n_1}a(p_0) + \sqrt{\frac{n_3}{n_1+n_3}}2\sqrt{n_3}a(p_0), 1)$;

Similarly, $z_{013} \sim \mathrm{n}(\sqrt{\frac{n_1}{n_1+n_3}}2\sqrt{n_1}a(p_0)+\sqrt{\frac{n_3}{n_1+n_3}}2\sqrt{n_3}a(p_0),1)$;

Therefore, $z_{v13} - z_{013} \sim \mathrm{n}(0,2) \Rightarrow \frac{z_{v13}-z_{013}}{\sqrt{2}} \sim n(0,1)$,

Since $\frac{z_{v2}-z_{02}}{\sqrt{2}} \sim n(0,1)$,

Then it follows that $T_3 = \sqrt{\frac{n_2}{n_1+n_2+n_3}}\left(\frac{z_{v2}-z_{02}}{\sqrt{2}}\right) + \sqrt{\frac{n_1+n_3}{n_1+n_2+n_3}}\left(\frac{z_{v13}-z_{013}}{\sqrt{2}}\right) \sim \mathrm{n}(0,1)$ under $H_0$.

> Claim <1>:
>
> $$\mathrm{EN} = (k+1)n_1 + (\mathrm{EM}+1)n_2 + n_3(1-\Phi(y_2)) + n_3\left(1-\Phi\left(y_2-\sqrt{2n_2}\Delta_2\right)\right), where\ EM = Prop * k\ (Prop\ is\ given\ in\ Table\ 3\ of\ 1960's\ paper)$$

Proof:

$$under\ H_0, \frac{z_{V2}-z_{02}}{\sqrt{2}} \sim N(0,1),$$

$$P_{H_0}(T_2 > y_2) = P_{H_0}\left(\frac{z_{V2}-z_{02}}{\sqrt{2}} > y_2\right) = 1-\Phi\ (y_2)$$

$$\text{under LFC}, \frac{z_{V2}-z_{02}}{\sqrt{2}} \sim N\left(\sqrt{2n_2}\Delta_2, 1\right)$$

$$P_{LFC}\left(T_2 = \frac{z_{V2}-z_{02}}{\sqrt{2}} > y_2\right) = P\left(\frac{z_{V2}-z_{02}}{\sqrt{2}} - \sqrt{2n_2}\Delta_2 > y_2 - \sqrt{2n_2}\Delta_2\right) = 1-\Phi\ \left(y_2-\sqrt{2n_2}\Delta_2\right)$$

$$\mathrm{EN}=\frac{1}{2}E(N|H_0)+\frac{1}{2}E(N|LFC)$$

$$=\frac{1}{2}[(\mathrm{k}+1)n_1+(EM+1)n_2+2n_3P_{H_0}(T_2>y_2)]+\frac{1}{2}[(\mathrm{k}+1)n_1+(EM+1)n_2+2n_3P_{LFC}(T_2>y_2)]$$

$$=(\mathrm{k}+1)n_1+(\mathrm{EM}+1)n_2+n_3P_{H_0}(T_2>y_2)+n_3P_{LFC}(T_2>y_2)$$

$$=(\mathrm{k}+1)n_1+(\mathrm{EM}+1)n_2+n_3(1-\Phi\ (y_2)\ )+n_3\left(1-\Phi\left(y_2-\sqrt{2n_2}\Delta_2\right)\right)$$

> Claim <2>:
>
> $$1-\beta^* = P_{\theta^*}(CS) = P(\pi_{[k]} with\ p_{[k]} = p+\delta\ is\ selected) = \sum_{x_0=0}^{n_2}\sum_{x_k=0}^{n_2}\{1-\Phi(y_3\sqrt{\frac{n_1+n_2+n_3}{n_1+n_3}} - \frac{\sqrt{2}n_2}{\sqrt{n_1+n_3}}\left(a\left(\frac{x_k}{n_2}\right)-a\left(\frac{x_0}{n_2}\right)\right)-\sqrt{2(n_1+n_3)}\Delta_2)\}\ \{\sum_{j=0}^{EM-1}\frac{1}{j+1}\binom{EM-1}{j}\{b(x_k,n_2,p)\}^j\{B(x_k-1,n_2,p)\}^{EM-1-j}\{\mathrm{I}\left(a\left(\frac{x_k}{n_2}\right)-a\left(\frac{x_0}{n_2}\right)>\frac{y_2}{\sqrt{2n_2}}\right)\}\{\mathrm{b}\{x_k,\mathrm{p}+\delta,n_2\}\mathrm{b}\{x_0,p_0,n_2\}\ \boldsymbol{P(*)}$$
>
> where (*) denotes the situation of correct selection in stage 1.

Proof:

Under LFC $\theta^*$(define $\arcsin\sqrt{p+\delta} - arcsin\sqrt{p_0} = a(p+\delta) - a(p_0) = \Delta_2$):

(LFC $\theta^*$: $p$ (control)=$p_0$ , $p_{[k]} = p+\delta, p_{[1]} = p_{[2]} = p_{[3]} \ldots. = p_{[k-1]} = p$):

$$z_{v13} \sim n(\sqrt{\frac{n_1}{n_1+n_3}}2\sqrt{n_1}a(p+\delta)+\sqrt{\frac{n_3}{n_1+n_3}}2\sqrt{n_3}a(p+\delta),1)$$

$$z_{013} \sim \mathrm{n}(\sqrt{\frac{n_1}{n_1+n_3}}2\sqrt{n_1}a(p_0)+\sqrt{\frac{n_3}{n_1+n_3}}2\sqrt{n_3}a(p_0),1)$$

$$\Rightarrow z_{v13} - z_{013} \sim \mathrm{n}(2\sqrt{n_1+n_3}\Delta_2, 2)$$

$=> \frac{z_{v13}-z_{013}}{\sqrt{2}} \sim n(\sqrt{2(n_1+n_3)}\Delta_2, 2)$………<2.1>

$P_{\theta^*}(CS) = P(\pi_{[k]} with\ p_{[k]} = p+\delta\ is\ selected)$

=*P(Treatment V passed all 3 stages| *)* ***P(*)***

$=\sum_{x_0=0}^{n_2}\sum_{x_k=0}^{n_2} P_{\theta^*}(X_{02}=x_0, X_{v2}=x_k, T_2>y_2, T_3>y_3, M\ \mid *)$ ***P(*)***

$=\sum_{x_0=0}^{n_2}\sum_{x_k=0}^{n_2} P_{\theta^*}(T_3>y_3 \mid M,\ T_2>y_2,\ X_{02}=x_0, X_{v2}=x_k,\ *) * \ P(M \mid T_2>y_2,\ X_{02}=x_0, X_{v2}=x_k, *) \ * \ I(T_2>y_2 \mid X_{02}=x_0, X_{v2}=x_k, *) \ * \ P(X_{02}=x_0, X_{v2}=x_k \mid *) \ * \ P(*)$ *<2.2>*

In <2.2>, both P($*$) and $P_{\theta^*}(CS)$ are preassigned.
Now computing the four parts respectively:

$\underline{P_{\theta^*}(T_3>y_3 \mid M, \quad T_2>y_2, \quad X_{02}=x_0, X_{v2}=x_k, \quad *)}$

$=P\{\sqrt{\frac{n_2}{n_1+n_2+n_3}}\left(\frac{z_{v2}-z_{02}}{\sqrt{2}}\right)+\sqrt{\frac{n_1+n_3}{n_1+n_2+n_3}}\left(\frac{z_{v13}-z_{013}}{\sqrt{2}}\right) > y_3 \mid M,\ T_2>y_2,\ X_{02}=x_0, X_{v2}=x_k, *\}$

$=P(\frac{z_{v13}-z_{013}}{\sqrt{2}} > [y_3-\sqrt{\frac{n_2}{n_1+n_2+n_3}}\left(\frac{z_{v2}-z_{02}}{\sqrt{2}}\right)]\sqrt{\frac{n_1+n_2+n_3}{n_1+n_3}})$

$\xRightarrow{by<2.1>} =P(\frac{z_{v13}-z_{013}}{\sqrt{2}}-\sqrt{2(n_1+n_3)}\Delta_2) > [y_3-\sqrt{\frac{n_2}{n_1+n_2+n_3}}\left(\frac{z_{v2}-z_{02}}{\sqrt{2}}\right)]\sqrt{\frac{n_1+n_2+n_3}{n_1+n_3}}-\sqrt{2(n_1+n_3)}\Delta_2)$

$=P(\frac{z_{v13}-z_{013}}{\sqrt{2}}-\sqrt{2(n_1+n_3)}\Delta_2) > y_3\sqrt{\frac{n_1+n_2+n_3}{n_1+n_3}}-\sqrt{\frac{n_2}{n_1+n_3}}\left(\frac{z_{v2}-z_{02}}{\sqrt{2}}\right)-\sqrt{2(n_1+n_3)}\Delta_2)$

$=P(\frac{z_{v13}-z_{013}}{\sqrt{2}}-\sqrt{2(n_1+n_3)}\Delta_2) > y_3\sqrt{\frac{n_1+n_2+n_3}{n_1+n_3}}-\sqrt{\frac{n_2}{n_1+n_3}}\left(\frac{2\sqrt{n_2}[a\left(\frac{X_{V2}}{n_2}\right)-a\left(\frac{X_{02}}{n_2}\right)]}{\sqrt{2}}\right)-\sqrt{2(n_1+n_3)}\Delta_2)$

$=1-\Phi(y_3\sqrt{\frac{n_1+n_2+n_3}{n_1+n_3}}-\frac{\sqrt{2}n_2}{\sqrt{n_1+n_3}}\left(a\left(\frac{x_k}{n_2}\right)-a\left(\frac{x_0}{n_2}\right)\right)-\sqrt{2(n_1+n_3)}\Delta_2)$

$\underline{P_{\theta^*}(M \mid T_2>y_2, \quad X_{02}=x_0, X_{v2}=x_k, *)}$

$$= \sum_{j=0}^{EM-1}\frac{1}{j+1}\binom{EM-1}{j}\{b(X_{v2},n_2,p)\}^j\{B(X_{v2}-1,n_2,p)\}^{EM-1-j}$$

$$= \sum_{j=0}^{EM-1}\frac{1}{j+1}\binom{EM-1}{j}\{b(x_k,n_2,p)\}^j\{B(x_k-1,n_2,p)\}^{EM-1-j}$$

$\underline{I(T_2>y_2)} = I\left(\frac{z_{V2}-z_{02}}{\sqrt{2}}>y_2\right) = I\left\{2\sqrt{n_2}\left(a\left(\frac{x_k}{n_2}\right)-a\left(\frac{x_0}{n_2}\right)\right)\right\} > \sqrt{2}y_2) = I\left(a\left(\frac{x_k}{n_2}\right)-a\left(\frac{x_0}{n_2}\right)>\frac{y_2}{\sqrt{2n_2}}\right)$

$\underline{P(X_{02}=x_0, X_{v2}=x_k \mid *)}$=b$\{x_k, p+\delta, n_2\}$b$\{x_0, p_0, n_2\}$

Claim <3>:

size $\alpha = k\sum_{x_0=0}^{n_2}\sum_{x_k=0}^{n_2}\left\{\mathbf{1}-\mathbf{\Phi}\left\{\boldsymbol{y_3}\sqrt{\frac{\boldsymbol{n_1+n_2+n_3}}{\boldsymbol{n_1+n_3}}}-\frac{\sqrt{2}\boldsymbol{n_2}}{\sqrt{\boldsymbol{n_1+n_3}}}\left(\boldsymbol{a}\left(\frac{\boldsymbol{x_k}}{\boldsymbol{n_2}}\right)-a\left(\frac{\boldsymbol{x_0}}{\boldsymbol{n_2}}\right)\right)\right\}\right\}\left\{\sum_{j=0}^{EM-1}\frac{1}{j+1}\binom{EM-1}{j}\{b(x_k,n_2,p_0)\}^j\{B(x_k-1,n_2,p_0)\}^{k-1-j}\right\} I\left(\boldsymbol{a}\left(\frac{\boldsymbol{x_k}}{\boldsymbol{n_2}}\right)-\boldsymbol{a}\left(\frac{\boldsymbol{x_0}}{\boldsymbol{n_2}}\right)>\frac{\boldsymbol{y_2}}{\sqrt{\boldsymbol{2n_2}}}\right)$ b$\{x_k, p_0, n_2\}$b$\{x_0, p_0, n_2\}$ $P(*)$

Proof:

size α
=P(any one of the k treatments is claimed as the best | $H_0$ is true)
=k $\sum_{x_0=0}^{n_2} \sum_{x_k=0}^{n_2} P_{H_0}(X_{02} = x_0, X_{v2} = x_k, T_2 > y_2, T_3 > y_3, M \mid *)$ ***P(*)***
=k $\sum_{x_0=0}^{n_2} \sum_{x_k=0}^{n_2} P_{H_0}(T_3 > y_3 \mid M,\ T_2 > y_2,\ X_{02} = x_0, X_{v2} = x_k,\ *) * P_{H_0}(M \mid T_2 > y_2,\ X_{02} = x_0, X_{v2} = x_k, *) * I(T_2 > y_2 \mid X_{02} = x_0, X_{v2} = x_k, *) * P_{H_0}(X_{02} = x_0, X_{v2} = x_k \mid *) * P_{H_0}(*)$ ,

$\underline{P_{H_0}(T_3 > y_3 \mid M,\quad T_2 > y_2,\quad X_{02} = x_0, X_{v2} = x_k,\quad *)}$
=P$\left\{\sqrt{\frac{n_2}{n_1+n_2+n_3}}\left(\frac{z_{v2}-z_{02}}{\sqrt{2}}\right) + \sqrt{\frac{n_1+n_3}{n_1+n_2+n_3}}\left(\frac{z_{v13}-z_{013}}{\sqrt{2}}\right) > y_3 \mid \frac{z_{v13}-z_{013}}{\sqrt{2}} \sim n(0,1)\right\}$
=P$\left(\frac{z_{v13}-z_{013}}{\sqrt{2}} > y_3\sqrt{\frac{n_1+n_2+n_3}{n_1+n_3}} - \sqrt{\frac{n_2}{n_1+n_3}}\left(\frac{z_{v2}-z_{02}}{\sqrt{2}}\right)\right)$
=1-Φ$\left\{y_3\sqrt{\frac{n_1+n_2+n_3}{n_1+n_3}} - \frac{\sqrt{2}n_2}{\sqrt{n_1+n_3}}\left(a\left(\frac{x_k}{n_2}\right) - a\left(\frac{x_0}{n_2}\right)\right)\right\}$

$\underline{P_{H_0}(M \mid T_2 > y_2,\ X_{02} = x_0, X_{v2} = x_k, *)}$
$$= \sum_{j=0}^{EM-1} \frac{1}{j+1}\binom{EM-1}{j}\{b(x_k, n_2, p_0)\}^j\{B(x_k - 1, n_2, p_0)\}^{EM-1-j}$$

$\underline{I(T_2 > y_2)}$=I$\left(a\left(\frac{x_k}{n_2}\right) - a\left(\frac{x_0}{n_2}\right) > \frac{y_2}{\sqrt{2n_2}}\right)$

$\underline{P_{H_0}(X_{02} = x_0, X_{v2} = x_k \mid *)=}$b$\{x_k, p_0, n_2\}$b$\{x_0, p_0, n_2\}$

*(Note: EM=k(EM/k)=kP(sub), where P(sub) is given in Table 3 of Gupta and Sobel's subset selection approach paper published in 1960.)*

## 4 Numerical Result

Table 1 presents the comparison between the three-stage design presented above and the two-stage procedure by Tall, Simon, and Ellenberg(1988). It should be noted that the value of EN in the table, rather than the global minimum from numerical analysis, is the local minimum for specified $y_1, y_2, n_1, n_2$, with the constraint of error rate α and power $1-\beta^*$ pre-assigned. The values for $n_1$, k and δ for the three-stage design is pre-assigned at stage one (subset selection stage) for a higher P(*) and lower ratio of $\frac{EM}{k}$. This ratio is given in table 3 of Gupta and Sobel's paper of binomial subset selection in 1960. Given $\delta = \delta_2 - \delta_1$, P(*), p+ δ= $p_0 + \delta_2$, the value of $p_0$, $\delta_2$ and $\delta_1$ became fixed. $n_3$ and $y_2$ can be obtained by solving the equation<2>and <3> with different pairs of $y_1, n_1, n_2$ and fixed α, $1-\beta^*$. The corresponding EN can be calculated from equation <1>. Observing $n_2$ and $n_3$ increase with k for fixed $\theta_0, n_1$, δ, and P(*), whereas EN decreases with δ for fixed k, P(*), α and 1-β*. The stage 3 cut-off values $y_3$ is more extreme than the one-sided test cut-off 1.645. This is due to the stage 1 and stage 2 selection process adjustment, $y_3$ cut-off value also increases with k.

Given K, $\delta_1$, $\delta_2$, $\theta_0$, α=0.05 and 1-$\beta^*$, we can obtain the corresponding EN using the two-stage designs method in Tall, Simon and Ellenberg(1988).For the same k, $\delta_1$, $\delta_2$, $\theta_0$, α and 1-$\beta^*$, most two-stage design shows a smaller EN. This is probably due to the waste of stage 1's sample data in the second stage, where $T_2$ only comes from data of stage 2. However, extreme cases exist when $n_1$ in two-stage design is fairly small and $n_2$ is comparably large. In such situations, the three stage turns out to require a smaller EN than two-stage, fixing the α and 1-$\beta^*$ equal. Therefore, the three-stage design may outperform the two-stage design even when data in stage 1 are not used in the three-stage procedures. If we reduce the accuracy of 1-$\beta^*$ by keeping one decimal place, and continue setting extreme values for $n_1$ and $n_2$ for the 2-stage design, more "three-stage outperforming" situations are detected. Furthermore, the stage 3 cut-off values of $y_3$ are more extreme than the two-stage design cut-off value $y_3$. Thus, intuitively, three-stage design gives more exact result than two-stage design.

## 5 Further Discussion

The proposed three-stage design is a combination of binomial subset selection and two-stage designs. Samples are collected in each stage, however, the stage 1's data is used only in the third stage, instead of both stage 2 and 3. The motivation for such design is due to fact that the randomization probability P(M)—the probability that the "truly best" population wins randomization— is almost impossible to be computed once $T_2$ contains data from the first two stages, where the $T_2$ has to be discussed and computed in an unknown number of complex situations.

A possible solvency for this difficulty could be assuming $n_1 = n_2$ , under which the tie would appear only at two situations: (1) ties appear at each of two stages: $X_{i1} = X_{j1}$ $and$ $X_{i2} = X_{j2}$, or (2)ties appear "across" the stages: $X_{i1} = X_{j2}$ $and$ $X_{i2} = X_{j1}$. However, even in such simplified situation, computation of the probability for the best treatment winning randomization is still a challenging task.

On the other hand, subset selection not always save sample size by reducing the number of population being tested. Especially when δ is below 0.2, subset selection always selects more than eighty percent of populations into the subset. This will lead to a dramatic waste of data of stage 1. For instance, you have 5 experimental groups (besides the control group), each of these have 20 samples for the first stage, if all these 5 are selected into the subset, then these 100 data for stage 1 would be wasted. Because in 2-stage approach, you have the same number of experimental groups to test, but will be able to make full use of more data without having to waste data on an additional stage.

We tried to avoid such situation by choosing numbers where the size of subset is at least reduced by 1. Table 1 below shows the result of comparison of the two approaches, where we only tested the cases where subset selection successfully reduced k to some extent. In order to control the P(*) for stage 1 at a high level, $n_1$is chosen to be 20 and k is set to 3 while P*=0.95 and P(sub) = 0.792. Thus, the subset size that enters stage 2 is: k* P(sub) = 3* 0.792, which is approximately 2. But even in such cases, two-stage approach performs much better in reducing E(N) than three-stage design, and the differences between E(N)s increase as k increases. This is the case when δ is 0.25. As δ gets larger, the power of two-stage design can be comparable large. In such cases, we claim the traditional two-stage design is good enough to detect the improved treatment. In summary, our three-stage approach with subset selection suffers obvious disadvantage when δ is less than 0.2, while the two-stage will perform extremely good when δ is greater than 0.25. Therefore, it is really challenging to search for a specific δ between 0.2 to 0.25, and a specific combination of $n_1, n_2, n_3$ and $y_1, y_2$that enable the three-stage to win.

For further study, we will continue working on making full use of the the first stage data in stage 2 and 3. If we managed to update $T_2$ by using data of the first two stage, the power for this three stage design should be increased, which may be able to generate a smaller EN subject to the same power and Type I error rate with the two-stage approach by Tall, Simon, and Ellenberg(1988). Such improvement should hold in a general situation, rather than in extreme and rare cases when 2-stage design did a poor job.

**Table 1. Designs having E(N) for given K, $\delta_1$, $\delta_2$, $\delta$, $\theta_0$, $\alpha$, P(*), and P(sub)**

| $\delta$ | 0.25 | | | | | | 0.5 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\delta_1$ | 0.1 | | | | | | 0.05 | | | |
| $\delta_2$ | 0.35 | | | | | | 0.55 | | | |
| $\theta_0$ | 0.4 | | | | | | 0.2 | | | |
| k | 3 | | | | 5 | | 3 | | 5 | |
| Designs | 3-stage | 2-stage | 3-stage | 2-stage | 3-stage | 2-stage | 3-stage | 2-stage | 3-stage | 2-stage |
| y2 | 0.482 | 0.43 | 1.9 | 0.85 | 0.65 | 0.45 | 0.6 | 0.79 | 0.75 | 0.9 |
| y3 | 1.9 | 1.88 | 0.8 | 1.8 | 2.03 | 2.04 | 2.08 | 1.92 | 2.18 | 2.07 |
| n1 | 20 | | 15 | | 20 | | 20 | | 20 | |
| n2 | 34 | 24 | 38 | 21 | 35 | 30 | 50 | 39 | 55 | 30 |
| n3 | 52 | 61 | 58 | 116 | 60 | 61 | 55 | 56 | 75 | 60 |
| EN | 197 | 153 | 154 | 187 | 256 | 238 | 165 | 190 | 246 | 219 |
| $\alpha$ | 0.0479 | 0.0499 | 0.04 | 0.04 | 0.0499 | 0.0497 | 0.0493 | 0.05 | 0.0494 | 0.04933 |
| 1-$\beta^*$ | 0.92 | 0.925 | 0.883 | 0.883 | 0.910 | 0.927 | 0.95 | 0.99 | 0.95 | 0.99 |
| P(*) | 0.95 | | 0.9 | | 0.95 | | 0.95 | | 0.95 | |
| P(sub) | 0.792 | | 0.735 | | 0.836 | | 0.402 | | 0.343 | |

**Table 2. Expected Proportion of Populations Retained in the Selected Subset in Stage One**

| $n$ | $p+\delta$ \ $k$ | $P^* = .75$: 2 | 3 | 5 | 10 | $P^* = .90$: 2 | 3 | 5 | 10 | $P^* = .95$: 2 | 3 | 5 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\delta = .25$ | | | | | | | | | | | | |
| 5 | .50 | .762 | .850 | .796 | .925 | .896 | .850 | .946 | .925 | .971 | .959 | .946 | .992 |
| | .75 | .762 | .849 | .794 | .916 | .896 | .849 | .941 | .916 | .971 | .958 | .941 | .988 |
| | .95 | .798 | .908 | .883 | .974 | .934 | .908 | .978 | .974 | .988 | .983 | .978 | .998 |
| | 1.00 | .816 | .931 | .917 | .986 | .948 | .931 | .987 | .986 | .992 | .990 | .987 | .999 |
| 10 | .50 | .742 | .643 | .715 | .816 | .838 | .883 | .853 | .926 | .915 | .883 | .939 | .926 |
| | .75 | .742 | .643 | .713 | .808 | .838 | .882 | .849 | .919 | .915 | .882 | .936 | .919 |
| | .95 | .754 | .668 | .783 | .895 | .870 | .928 | .912 | .968 | .947 | .928 | .973 | .968 |
| | 1.00 | .763 | .684 | .821 | .930 | .888 | .948 | .937 | .982 | .961 | .948 | .984 | .982 |
| 15 | .50 | .651 | .628 | .671 | .607 | .805 | .735 | .792 | .864 | .875 | .821 | .884 | .935 |
| | .75 | .651 | .628 | .670 | .604 | .805 | .734 | .789 | .858 | .875 | .830 | .882 | .931 |
| | .95 | .632 | .635 | .719 | .678 | .827 | .768 | .850 | .924 | .908 | .877 | .934 | .973 |
| | 1.00 | .618 | .641 | .749 | .718 | .843 | .791 | .881 | .949 | .926 | .901 | .955 | .984 |
| 20 | .50 | .596 | .528 | .528 | .580 | .714 | .705 | .747 | .812 | .781 | .792 | .837 | .892 |
| | .75 | .596 | .529 | .527 | .577 | .714 | .705 | .745 | .808 | .781 | .792 | .836 | .888 |
| | .95 | .566 | .503 | .531 | .628 | .709 | .728 | .798 | .877 | .796 | .832 | .892 | .943 |
| | 1.00 | .546 | .483 | .532 | .655 | .707 | .745 | .849 | .908 | .809 | .857 | .919 | .963 |
| 25 | .50 | .562 | .527 | .516 | .557 | .703 | .682 | .615 | .669 | .763 | .761 | .798 | .770 |
| | .75 | .562 | .527 | .516 | .555 | .703 | .682 | .614 | .666 | .763 | .761 | .797 | .766 |
| | .95 | .533 | .497 | .509 | .587 | .694 | .696 | .634 | .721 | .772 | .795 | .851 | .832 |
| | 1.00 | .516 | .476 | .503 | .605 | .689 | .707 | .649 | .754 | .781 | .818 | .880 | .866 |
| | $\delta = .50$ | | | | | | | | | | | | |
| 5 | .50 | .594 | .667 | .600 | .831 | .750 | .667 | .850 | .831 | .906 | .875 | .850 | .972 |
| | .75 | .610 | .646 | .568 | .769 | .737 | .646 | .801 | .769 | .878 | .836 | .801 | .948 |
| | .95 | .600 | .657 | .586 | .806 | .744 | .657 | .831 | .806 | .896 | .860 | .831 | .964 |
| | 1.00 | .594 | .667 | .600 | .831 | .750 | .667 | .850 | .831 | .906 | .875 | .850 | .972 |
| 10 | .50 | .527 | .370 | .338 | .439 | .586 | .585 | .502 | .661 | .688 | .585 | .698 | .661 |
| | .75 | .551 | .400 | .369 | .440 | .607 | .588 | .505 | .625 | .691 | .588 | .667 | .625 |
| | .95 | .537 | .383 | .351 | .439 | .595 | .585 | .502 | .645 | .689 | .585 | .685 | .645 |
| | 1.00 | .527 | .370 | .338 | .439 | .586 | .585 | .502 | .661 | .688 | .585 | .698 | .661 |
| 15 | .50 | .502 | .345 | .247 | .153 | .530 | .373 | .321 | .373 | .575 | .434 | .443 | .550 |
| | .75 | .511 | .367 | .284 | .194 | .553 | .404 | .357 | .393 | .598 | .464 | .461 | .537 |
| | .95 | .505 | .353 | .263 | .170 | .539 | .385 | .336 | .381 | .585 | .447 | .450 | .543 |
| | 1.00 | .502 | .345 | .247 | .153 | .530 | .373 | .321 | .378 | .575 | .434 | .443 | .550 |
| 20 | .50 | .500 | .334 | .205 | .119 | .503 | .347 | .246 | .218 | .510 | .372 | .305 | .327 |
| | .75 | .502 | .341 | .221 | .149 | .513 | .370 | .283 | .261 | .527 | .402 | .343 | .356 |
| | .95 | .501 | .336 | .210 | .130 | .506 | .356 | .261 | .236 | .517 | .384 | .321 | .339 |
| | 1.00 | .500 | .334 | .205 | .119 | .503 | .347 | .246 | .218 | .510 | .372 | .305 | .327 |
| 25 | .50 | .500 | .334 | .202 | .107 | .501 | .338 | .206 | .119 | .504 | .348 | .243 | .148 |
| | .75 | .501 | .338 | .211 | .126 | .507 | .352 | .223 | .150 | .514 | .370 | .279 | .188 |
| | .95 | .500 | .335 | .205 | .113 | .503 | .343 | .212 | .131 | .507 | .356 | .257 | .165 |
| | 1.00 | .500 | .334 | .202 | .107 | .501 | .338 | .206 | .119 | .504 | .348 | .243 | .148 |