Journal of Data Science 13(2015), 115-126

A Type of Sample Size Planning for Mean Comparison in Clinical Trials

Junfeng Liu[1*] and Dipak K. Dey[2]

[1] *GCE Solutions, Inc.*
[2] *Department of Statistics, University of Connecticut*

Abstract: Early phase clinical trials may not have a known variation (σ) for the response variable. In the light of applying t-test statistics, several procedures were proposed to use the information gained from stage-I (pilot study) to adaptively re-estimate the sample size for managing the overall hypothesis test. We are interested in choosing a reasonable stage-I sample size (m) towards achieving an accountable overall sample size (stage-I and later). Conditional on any specified m, this paper replaces σ by the estimated σ (from stage-I with sample size m) to use the conventional formula under normal distribution assumption to re-estimate an overall sample size. The estimated σ, re-estimated overall sample size and the collective information (stage-I and later) would be incorporated into a surrogate normal variable which undergoes hypothesis test based on standard normal distribution. We plot the actual type I&II error rates and the expected sample size against m in order to choose a good universal stage-I sample size (m^*) to start.

Key words: Hypothesis test, normal test, surrogate, Type-I(II) error.

1. Introduction

We assume the well-studied standard treatment has a mean effect μ_0 (known) and the experimental treatment has a mean effect μ_1 (to be tested) with subject-indexed endpoint values $(Y_1, \cdots, Y_{n,} \cdots)$ independently following a normal distribution, i.e., $Y_i \sim N(\mu_1, \sigma^2)$, $i \geq 1$. The hypothesis to be tested is

$$H_0 : \mu_1 = \mu_0 \qquad \text{vs.} \qquad H_1 : \mu_1 \geq \mu_0 + \Delta, \;\; \Delta > 0.$$

Flexible methods for clinical trial design (e.g., the number of planned interim analyses, sample size, test statistic, rejection region) have been developed under different scenarios to improve hypothesis test efficiency with the actual type I & II error rates (α,β) well controlled (e.g., Pocock (1977), O'Brien and Fleming (1979), Lan and DeMets (1983), Wang and Tsiatis (1987), Wittes and Brittain (1990), Ashby and Machin (1993), Whitehead (1993), Joseph and Bélisle (1997), Cui, Hung, and Wang (1999), Müller and Schäfer (2001,2004), Burington and Emerson (2003), Bartroff and Lai (2008)). For instance, several group sequential methods proposed sample size re-estimation based on interim estimate of effect size and/or calculation

* Corresponding author.

of conditional power for confirmatory studies (e.g., phase III) with variance specified. Other design changes (e.g., α-spending function, future interim analysis times) could be based on special combination rules for p-values or conditional rejection error probabilities at the first interim analysis.

Often times, the variation of random variable Y remains unknown at the start of investigating a new treatment. Several methods utilize an estimated variation ($\hat{\sigma}$) from an internal or external pilot study for further sample size estimation used for later stages (e.g., Browne (1995), Kieser and Friede (2000), Posch and Bauer (2000), Coffey and Muller (2001), Proschan (2005), Friede and Kieser (2006), Kairalla *et al*. (2012)). These methods are mainly based on applying certain types of *t*-test statistics. Under similar circumstances, the present work utilizes a simple method to re-estimate the overall sample size as well as an accountable stage-I sample size (m^*).

The rest of the article is organized as follows. Section 2 investigates the relationship between the recommended sample sizes from using normal and *t* tests when σ is known. In Section 3, we use the stage-I (pilot study) sample variation to call the conventional normal test sample size re-estimation formula and apply a surrogate test statistic to the hypothesis test procedure involving the standard normal variable. We numerically study its type-I error rate, power and the expected sample size. Section 4 concludes the paper.

2. A Comparison between Normal and *t* Tests (σ is known)

The σ availability or not determines which distribution is to be utilized for hypothesis testing. An explicit σ yields test statistic $\sqrt{n}(\bar{Y}_n - \mu_0)/\sigma$ with the sample size (n) satisfying

$$\Phi\left(\sqrt{n}\Delta/\sigma - \Phi^{-1}(1-\alpha)\right) \geq 1-\beta$$

Thus,

$$n = \mathrm{SSN}(\alpha,\beta,\sigma,\Delta) = \min\left\{n \geq 1: \Phi_{\sqrt{n}\Delta/\sigma}\left(\Phi^{-1}(1-\alpha)\right) \leq \beta\right\} \quad (1)$$

$$= \min\left\{n \geq 1: n \geq \left(\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\right)^2(\sigma/\Delta)^2\right\}$$

$$= \min\left\{n \geq 1: \Pr\left(Z + \sqrt{n}\Delta/\sigma \geq \Phi^{-1}(1-\alpha)\right) \geq 1-\beta\right\},$$

where Z is the standard normal variable and $\Phi_\theta(\cdot)$ is the cumulative distribution function for $Z+\theta$. If random variable $Y \sim N(\theta, \sigma^2)$ and random variable $V \sim \chi^2_\nu$ with degrees of freedom ν (Y and V are independent), then it is known that $Y/\sqrt{V/\nu}$ follows a non-central t -distribution with ν degrees of freedom and non-centrality parameter θ with probability density function

$$f_{\nu,\theta}(x) = \frac{\nu^{\nu/2}}{\sqrt{\pi}\Gamma(\nu/2)}\frac{e^{-\theta^2/2}}{(\nu+x^2)^{(\nu+1)/2}}\sum_{k=0}^{\infty}\Gamma((\nu+k+1)/2)\frac{\theta^k}{k!}\left(\frac{2x^2}{\nu+x^2}\right)^{k/2}$$

When σ is unknown, one-sample t-test is usually applied

$$(\bar{Y}_n - \mu_0)/(s_n/n) > T_{n-1}^{-1}(1-\alpha) \;\Rightarrow\; \text{reject } H_0.$$

The power is $1 - T_{\sqrt{n}\Delta/\sigma,n-1}\left(T_{n-1}^{-1}(1-\alpha)\right)$, where$T_{\theta,n-1}(\cdot)$ is the cumulative non-central *t* distribution function with degrees of freedom $n-1$ and non-centrality parameter θ. The scale parameter σ is involved in finding sample size (*n*) such that

$$n = \text{SST}(\alpha,\beta,\sigma,\Delta) = \min\left\{n \geq 1: T_{\sqrt{n}\Delta/\sigma,n-1}\left(T_{n-1}^{-1}(1-\alpha)\right) \leq \beta\right\}$$

$$= \min\left\{n \geq 1: \Pr\left(Z + \sqrt{n}\Delta/\sigma \geq T_{n-1}^{-1}(1-\alpha)\sqrt{V/(n-1)}\right) \geq 1-\beta\right\} \qquad (2)$$

where, $V \sim \chi^2_{n-1}$. When Δ is specified (e.g., log(2)), corresponding to a two-fold change, the sample size comparison between using (1) and (2) shows that, *t* test consistently requires 1 or 2 more subjects than normal test as σ varies. When Δ is specified, we are interested in proposing effective ways for recommending the required sample size even if σ remains unknown.

3. A Type of Overall Sample Size Planning

3.1 A Surrogate Normal Variable

For the mean comparison with effect size Δ, when σ remains unknown, we consider replacing σ by stage-I estimation ($\hat{\sigma}_m$) followed by the sample size determination rule (1) with $\hat{\sigma}_m = \left(\frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y}_m)^2\right)^{1/2}$. When α, β and Δ are fixed, we consider function $n^*(\sigma) = D\sigma^2$, where $D = T\Delta^{-2}$ and $T = \left(\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\right)^2$. The recommended overall sample size (denoted by $n(\hat{\sigma}_m)$) is the smallest integer which is no less than $n^*(\hat{\sigma}_m)$, i.e.,

$n(\hat{\sigma}_m) = \lfloor n^*(\hat{\sigma}_m) \rfloor$, if $n^*(\hat{\sigma}_m)$ is an integer (with probability measure 0);
$n(\hat{\sigma}_m) = \lfloor n^*(\hat{\sigma}_m) \rfloor + 1$, if $n^*(\hat{\sigma}_m)$ is not an integer (with probability measure 1).

Where $[\star]$ represents the largest integer not exceeding $\star$. Conditional on $\hat{\sigma}_m$ and $n(\hat{\sigma}_m)$, we study a surrogate normal variable (SNV) which is defined as

$$\text{SNV} = \frac{\bar{Y}_{n(\hat{\sigma}_m)} - \mu_0}{\hat{\sigma}_m/\sqrt{n(\hat{\sigma}_m)}} = \frac{\bar{Y}_{\lfloor D\hat{\sigma}_m^2 \rfloor + 1} - \mu_0}{\hat{\sigma}_m/\sqrt{\lfloor D\hat{\sigma}_m^2 \rfloor + 1}} = \frac{\bar{Y}_{\lfloor D\hat{\sigma}_m^2 \rfloor + 1} - \mu_0}{\sigma/\sqrt{\lfloor D\hat{\sigma}_m^2 \rfloor + 1}} \times \frac{\sigma}{\hat{\sigma}_m}. \qquad (3)$$

Its distribution is regulated by parameters *D*, m and δ_μ(defined as $\mu_1 - \mu_0$). When $\mu_1 = \mu_0$, (3) amounts to a central *t*-distribution with degrees of freedom of $m-1$. When $\mu_1 \neq \mu_0$, (3) can be rewritten as

$$\text{SNV} = \left(\frac{\bar{Y}_{\lfloor D\hat{\sigma}_m^2 \rfloor+1} - E(Y)}{\sigma/\sqrt{\lfloor D\hat{\sigma}_m^2 \rfloor+1}} + \frac{E(Y)-\mu_0}{\sigma/\sqrt{\lfloor D\hat{\sigma}_m^2 \rfloor+1}} \right) \Big/ \left(\frac{\hat{\sigma}_m}{\sigma} \right). \quad (4)$$

Although $(E(Y)-\mu_0)(\lfloor D\hat{\sigma}_m^2 \rfloor+1)^{1/2}\sigma^{-1}$ plays the similar role as θ does in the non-central *t*-distribution (Section 2), it is not a constant and (4) is no longer a non-central *t*-distribution. SNV (3).(4) presents certain degree of skewness when $E(Y) \neq \mu_0$.

3.2 A Testing Procedure

We apply the following hypothesis testing rule

$$\text{SNV} \geq \Phi^{-1}(1-\alpha) \;\Rightarrow\; \text{reject } H_0. \quad (5)$$

Under H_0, the actual type-I error rate is

$$\Pr\left(\text{SNV} \geq \Phi^{-1}(1-\alpha)\right) = \Pr\left(\frac{\bar{Y}_{n(\hat{\sigma}_m)} - \mu_0}{\sigma/\sqrt{n(\hat{\sigma}_m)}} \geq \Phi^{-1}(1-\alpha)\frac{\hat{\sigma}_m}{\sigma} \right)$$

$$= 1 - \int \Phi\left(\Phi^{-1}(1-\alpha)\frac{\hat{\sigma}_m}{\sigma}\right) dF_{m,\sigma}(\hat{\sigma}_m) = 1 - \int_{>0} \Phi(\Phi^{-1}(1-\alpha)x) f_m(x)dx, \quad (6)$$

where $f_m(x) = \frac{(m-1)^{(m-1)/2}}{2^{(m-3)/2}\Gamma((m-1)/2)} x^{m-2} e^{-(m-1)x^2/2}$ due to $\hat{\sigma}_m^2/\sigma^2 \sim \chi_{m-1}^2/(m-1))^{1/2}$. Eq.(6) depends on α and m only. The relationship between the actual and planned type-I error rates (Eq.(6)) indicates that an adjusted significance level α_{adj} corresponds to the prescribed nominal level α such that $\alpha = 1 - \int_{>0} \Phi\left(\Phi^{-1}\left(1-\alpha_{\text{adj}}\right)x\right) f_m(x)dx$. The Students' *t*-distribution (under H_0) has probability distribution function $\Pr(\text{SNV} \leq z) = \int_{>0} \Phi(xz) f_m(x)dx$. Given δ_μ, the actual power is

$$\Pr(\text{SNV} \geq \Phi^{-1}(1-\alpha)) = \Pr\left(\frac{\bar{Y}_{n(\hat{\sigma}_m)} - \mu_0}{\sigma/\sqrt{n(\hat{\sigma}_m)}} \geq \Phi^{-1}(1-\alpha)\frac{\hat{\sigma}_m}{\sigma} \right)$$

$$= \int \left(1 - \Phi\left(\Phi^{-1}(1-\alpha)\frac{\hat{\sigma}_m}{\sigma} - \frac{\delta_\mu\sqrt{n(\hat{\sigma}_m)}}{\sigma}\right)\right) \mathrm{d}F_{m,\sigma}(\hat{\sigma}_m). \quad (7)$$

When $E(Y) = \mu_1$, the probability distribution function for this SNV is

$$\Pr(\text{SNV} < x) = 1 - \int \left(1 - \Phi(x \times \frac{\hat{\sigma}_m}{\sigma} - \frac{\delta_\mu}{\sigma} \times (\Delta^{-2}\, T\hat{\sigma}_m^2 + 1)^{1/2})\right) \mathrm{d}F_{m,\sigma}(\hat{\sigma}_m).$$

(7) could be rewritten as

$$\Pr\left(\text{SNV} \geq \Phi^{-1}(1-\alpha)\right)$$

$$= \int \left(1 - \Phi\left(\Phi^{-1}(1-\alpha)\frac{\hat{\sigma}_m}{\sigma} - \frac{\Delta+(\delta_\mu-\Delta)}{\sigma}\sqrt{n^*(\hat{\sigma}_m)} + \frac{\delta_\mu}{\sigma}\left(\sqrt{n^*(\hat{\sigma}_m)} - \sqrt{n(\hat{\sigma}_m)}\right)\right)\right) \mathrm{d}F_{m,\sigma}(\hat{\sigma}_m)$$

$$= \int \left(1 - \Phi\left(-\Phi^{-1}(1-\beta)\frac{\hat{\sigma}_m}{\sigma} - \frac{\delta_\mu-\Delta}{\sigma}\sqrt{n^*(\hat{\sigma}_m)} + \frac{\delta_\mu}{\sigma}\left(\sqrt{n^*(\hat{\sigma}_m)} - \sqrt{n(\hat{\sigma}_m)}\right)\right)\right) \mathrm{d}F_{m,\sigma}(\hat{\sigma}_m), \quad (8)$$

which depends on α, β, Δ, σ and m. The power increases as δ_μ increases. When $\delta_\mu > \Delta$, the power has a lower bound

$$\int \left(1 - \Phi\left(-\Phi^{-1}(1-\beta) \times \frac{\hat{\sigma}_m}{\sigma}\right)\right) dF_{m,\sigma}(\hat{\sigma}_m) = \int \Phi(\Phi^{-1}(1-\beta)x) f_m(x)dx, \quad (9)$$

which depends on β and m only. At different nominal α and β values, Figure 1 demonstrates the actual type-I error rate (6), power lower bound (9) and adjusted significance level profiles as m varies. The observed type-I error rate inflation has also been reported by previous studies which used other sample size re-estimation approaches (e.g., Kairalla *et al*., 2012). As m increases, the actual type-I error rate and power lower bound monotonically approach the respective nominal values. When m is moderate (e.g., $m = 15$), neither the difference between the nominal and actual type-I error rates nor the difference between the nominal and actual powers is substantial.

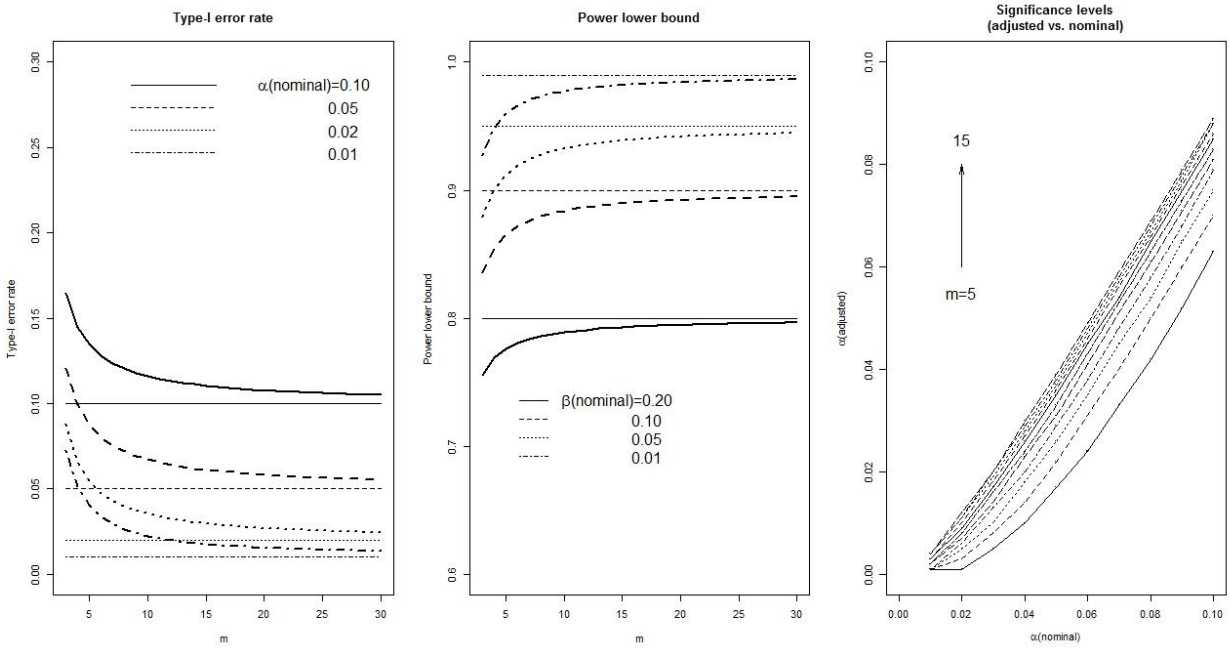

Figure 1: α-specific type-I error rate profiles (the left panel). β-specific power lower bound profiles (the middle panel). The horizontal lines are the respective nominal ones.

3.3 Power

We numerically study the properties of the power function (7),(8).

(1) When $\delta_\mu = \Delta = \log(2)$, Figure 2 (the left panel) shows the resultant power profile as m varies at $\sigma = 2^{\mathrm{i}-3}$, $\mathrm{i} = 0,\ldots,5$. Power increases as σ decreases. When σ is small (e.g., $\sigma = 1/4$), the power could be larger than the nominal value (80%) and the power profile may not be monotonically increasing with m. When σ increases (e.g., $\sigma > 1$), power profiles monotonically increase as m increases and profiles cluster with each other.

(2) When $\delta_\mu = \Delta = \log(2)$, Figure 2 (the right panel) shows the resultant power profile as σ varies at $\mathrm{m} = 5 \times \mathrm{i}$, $\mathrm{i} = 0,\ldots,5$. When σ is small (e.g., $\sigma < 1/2$), smaller ms lead to larger powers conditional on σ. When σ gets larger (e.g., $\sigma \geq 1/2$), smaller m leads to smaller powers conditional on σ.

(3) When $\sigma = 1$, Figure 3 (the left panel) shows the resultant power profile as $\delta_\mu = \Delta$ varies at $m = 5 \times \mathrm{i}, \mathrm{i} = 1,\cdots 5$. Each power profile roughly monotonically increases as Δ increases.

(4) When $\Delta = \log(2)$ and $\sigma = 1$, Figure 3 (the right panel) shows the resultant power profile as δ_μ varies at $m = 5 \times i, i = 1,\cdots 5$. Larger ms have larger powers.

Instead of considering the probability of achieving the planned power, we observe that the achieved actual power is well above or close to the planned value as m increases to certain value (e.g., 15).

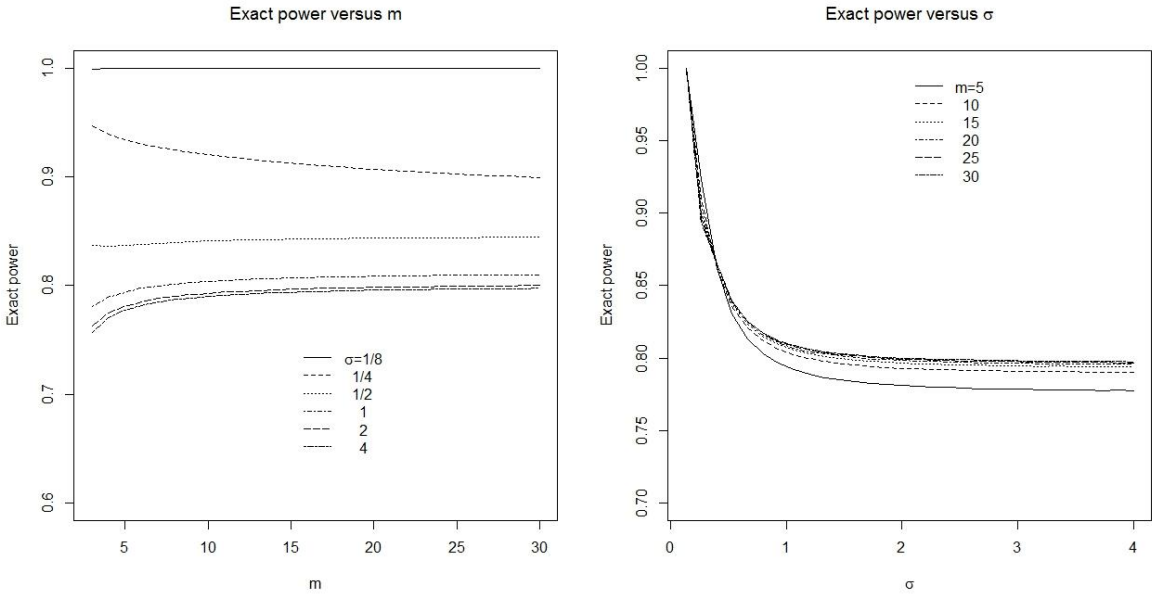

Figure 2: The exact power versus m (the left panel) and σ (the right panel). The power is calculated at $\delta_\mu = \Delta = \log(2)$.

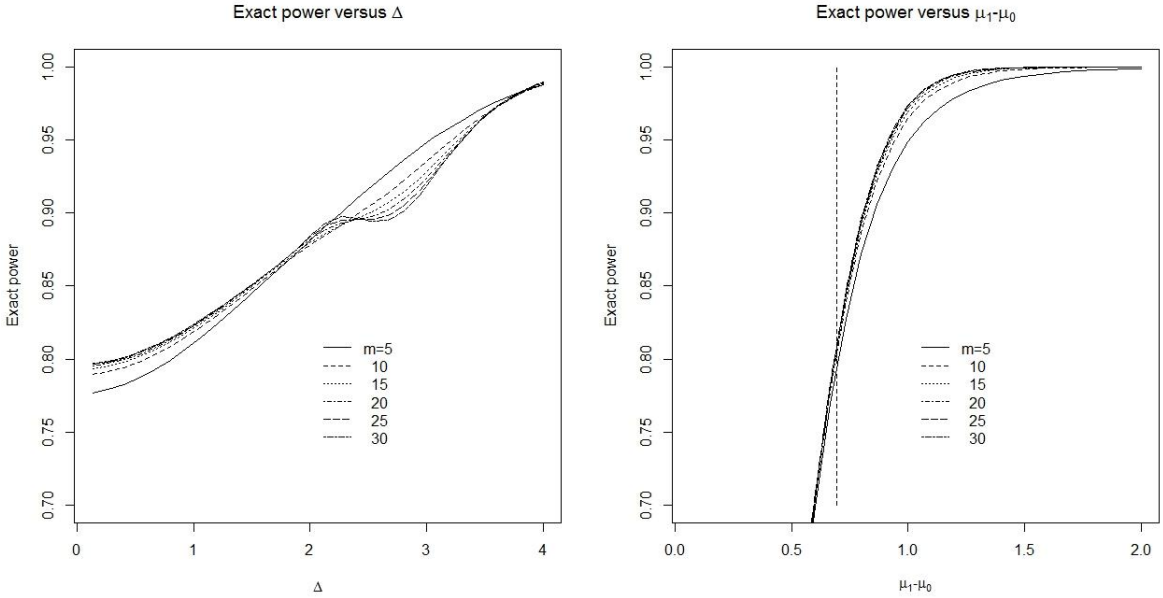

Figure 3: The exact power versus Δ (the left panel) and δ_μ (the right panel). The left panel shows the power profiles at $\delta_\mu = \Delta$ and $\sigma = 1$. The right panel shows the power profiles with $\Delta = \log(2)$ and $\sigma = 1$, where the vertical dotted line is $\delta_\mu = \Delta$.

3.4 Expected Sample Size

Since sample mean and variances are independent of each other given sample size m, the recommended overall sample size (Section 3.1) may include the m stage-I subjects before using the testing rule (5).

(1) If $n(\hat{\sigma}_m) > m$, we recruit additional $n(\hat{\sigma}_m) - m$ subjects (additional to m) for the overall test. $\bar{Y}_{n(\hat{\sigma}_m)}$ is subsequently available.

(2) If $n(\hat{\sigma}_m) \leq m$, we reuse a random sample ($n(\hat{\sigma}_m)$) out of the existent m stage-I subjects (e.g., the first $n(\hat{\sigma}_m)$ out of m) to obtain $\bar{Y}_{n(\hat{\sigma}_m)}$;

For each stage-I size (m), the actual expected overall sample size (n) depends on (m, σ, Δ, α, β). Specifically,

$$E(n|m) = \Pr(n(\hat{\sigma}_m) \geq m) \times E(n(\hat{\sigma}_m)|n(\hat{\sigma}_m) \geq m) + m \times \Pr(n(\hat{\sigma}_m) < m)$$

Given σ, the naive expected sample size comes from (1). We are interested in the difference between the expected sample sizes using SNV and (1). Figure 4 shows the resultant expected sample size difference under several settings. When σ gets larger, the difference becomes ignorable. Stage-I sample recycle or not are compared and our results indicate that the powers from two scenarios are close to each other (similar to the left panel in Figure 2). However, Figure 5 shows that, sample reuse may achieve a smaller type-I error rate compared to sample recruit (not reuse) when σ is smaller (e.g., $= 1/8$, $1/4$, $1/2$).

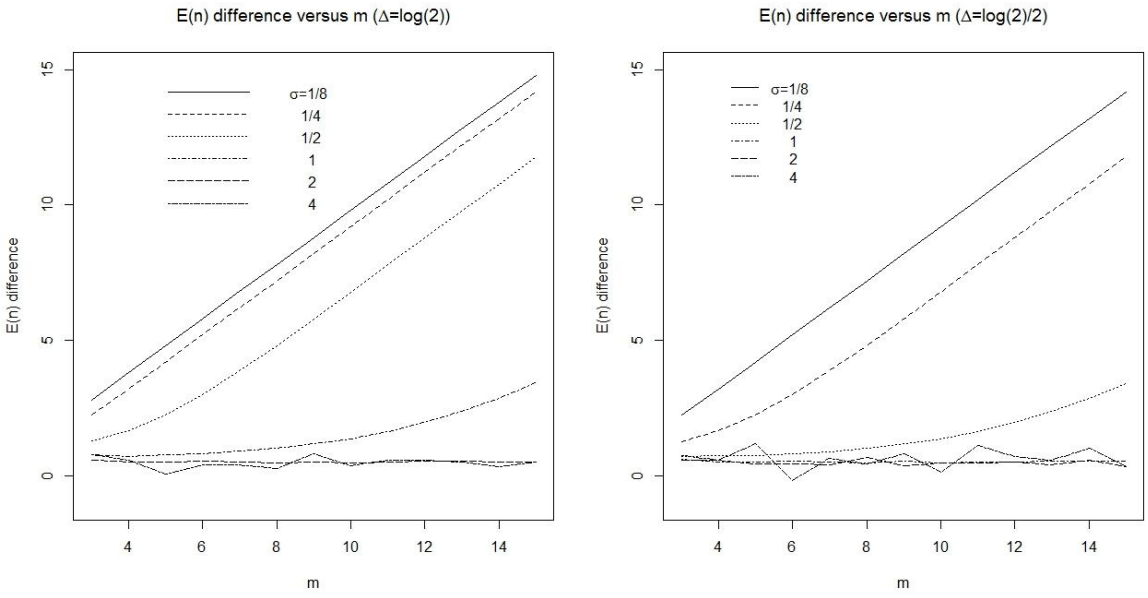

Figure 4: The expected sample size difference between SNV test (σ unknown) and naive normal test (σ known)

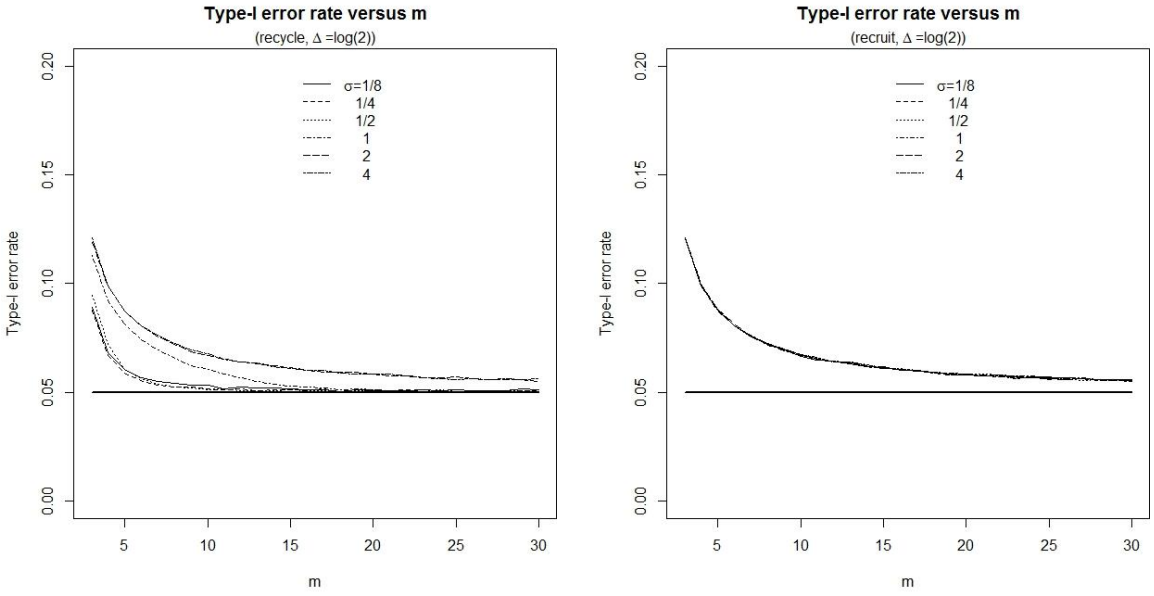

Figure 5: The actual type-I error rate versus m by simulation ($\delta_\mu = \Delta = \log(2)$). The left panel is stage-I sample reuse when the overall sample size is less thean *m.* The right panel is the recruiting new samples. When stage-I subjects are reused, the actual type-I error rate gets smaller (compared to sample recruit) as σ becomes smaller (e.g., $\sigma = 1/8,\ 1/4, 1/2$).

4. Conclusion

The proposed SNV approach to sample size planning (stage-I and overall) and hypothesis testing does not require pre-estimation of the variation from an external pilot study. This method is simple to implement and the collected information used for hypothesis testing comes from each stage of the trial to improve the efficiency. The achieved type-I error rate and power lower bound do not depend on the variation and are close to the nominal ones. Although the resultant expected overall sample size may be larger than that derived under the naive situation where the variation is known, the maximum difference is at most stage-I sample size when the variation is small and ignorable when the variation gets large. A universal selection of stage-I sample size (e.g., m^*=15) is feasible in view of the considered evaluation criteria.

References

[1] Ashby, D. and Machin, D. (1993). Stopping rules, interim analysis and data monitoring committees. *British Journal of Cancer* **68**, 1047-1050.

[2] Bartroff, J. and Lai, T.L. (2008). Efficient adaptive designs with mid-course sample size adjustment in clinical trials. *Statistics in Medicine* **27**, 659-663.

[3] Browne, R.H. (1995). On the use of a pilot sample for sample size determination. *Statistics in Medicine* **14**, 1933-1940.

[4] Burington, B.E. and Emerson, S.S. (2003). Flexible implementations of group sequential stopping rules using constrained boundaries. *Biometrics* **59**, 770-777.

[5] Coffey, C.S. and Muller, K.E. (2001). Controlling test size while gaining the benefits of an internal pilot design. *Biometrics* **57**, 625-631.

[6] Cui, L., Hung, H.M.J., and Wang, S.-J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics* **55**, 853-857.

[7] Friede, T. and Kieser, M. (2006). Sample size recalculation in internal pilot study designs: A review. *Biometrical Journal* **48(4)**, 537-555.

[8] Joseph, L. and Bélisle, P. (1997). Bayesian sample size determination for normal means and difference between normal means. *The Statistician* **46**, 209-226.

[9] Kairalla, J.A., Coffey, C.S., Thomann, M.A. and Muller, K.E. (2012). Adaptive trial designs: a review of barriers and opportunities. *Trials* **13**, 145.

[10] Kieser, M. and Friede, T. (2000). Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in Medicine* **19**, 901-911.

[11] Lan, K.K.G. and DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659-663.

[12] Müller, H.-H. and Schäfer H. (2001). Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* **57**, 886-891.

[13] Müller, H.-H. and Schäfer H. (2004). A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine* **23**, 2497-2508.

[14] O'Brien, P.C. and Fleming, T.R. (1979). A multiple testing procedure for clinical trials.*Biometrics* **35**, 549-556.

[15] Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191-199.

[16] Posch, M. and Bauer, P. (2000). Interim analysis and sample size reassessment. *Biometrics* **50**, 1170-1176.

[17] Proschan, M.A. (2005). Two-stage sample size re-estimation based on a nuisance parameter: A review. *Journal of Biopharmaceutical Statistics* **15**, 559-574.

[18] Wang, S.K. and Tsiatis, A.A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* **43**, 193-200.

[19] Whitehead, J. (1993). Interim analysis and stopping rules in cancer clinical trials. *British Journal of Cancer* **68**, 1179-1185.

[20] Wittes, J. and Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine* **9**, 65-72.

Received December 12, 2013; accepted July 26, 2014.

Junfeng Liu
GCE Solutions, Inc.
Bloomington, IL 61701, USA
jeff.liu@gcesolutions.com
1-(908)367-1391

Dipak K. Dey
Department of Statistics
University of Connecticut
Storrs, Ct 06269-4120, USA
dipak.dey@uconn.edu