Published in final edited form as:

Stat Biosci. 2015 October 1; 7(2): 348–366. doi:10.1007/s12561-014-9123-3.

Options and Considerations for Adaptive Laboratory Experiments

Lai Wei and David Jarjoura

Center for Biostatistics, The Ohio State University, 2012 Kenny Road, Columbus, OH 43221, U.S.A.

Abstract

Motivated by laboratory experiments that fail to reach significance, we developed a small sample size approach to designing a subsequent experiment that controls overall type I error and achieves sufficient conditional power. We focus on experiments with leukemia cells, and use a specific example in Chronic Lymphocytic Leukemia to discuss unanticipated patient variance and difficult to predict interaction effect sizes. We emphasize the importance of achieving significance in the first run of an experiment, which results in simplifying the multiple considerations usually associated with interim analysis and decision making in adaptive clinical trials. Within the context of combination testing for an adaptive laboratory experiment, we show that a range of reasonable options for the futility cut-off, effect size estimation, and significance level for the first run provide similar power and expected overall sample size. We contrast this approach to a naive procedure in which a second unplanned experiment is run based on non-significance in the first experiment, and data are combined as if they were obtained from one run.

Keywords

Conditional error function; Conditional power; Small sample size; Sample size re-estimation

1 Introduction

Motivating examples for this methodological research were experiments on cells from patients suffering from Chronic Lymphocytic Leukemia. In one experiment, a test condition was expected to promote CD154 expression, a clinically relevant target. Based on previous experimental results of doubtful relevance, a sample size of 6 was proposed. In this type of leukemia cell experiment each patient provides a sufficient number of cells so that their cells can be observed under all conditions of the experiment. Still, prior experience with these leukemia cell experiments had us concerned about variance across patients in the differences between the effects of test vs. control conditions. So we raised the issue of adding replicates, if needed, in a second run in this and other similar leukemia cell experiments. It is common for us to simultaneously plan a whole series of experiments with the scientists. Each is planned to provide a component of a coherent story about the mechanisms that control the disease, and how treatment conditions change the cells behavior and ameliorate their

proliferation. For the whole series of experiments, it is important to reach significance on almost all of them in order to confirm the proposed mechanism, and so this is another motivation for planning a method to recover from non-significance. Clearly considering all experiments simultaneously is beyond the scope of this paper, so we focused on one of the most important ones in the series to motivate our approach. In the example of leukemia cell type experiments only two conditions were tested, so that a comparison of means with a one-sample t-test seemed appropriate. A large scientifically meaningful effect size was assumed to assess a sufficiently powerful sample size. Planning on another run was informal at first, because the developments proposed here were not yet fully developed. We simply used a more conservative alpha and found that after running this experiment, the one-sided p-value was $p_1 = 0.06$. The effect size was smaller than expected, because patient variance was larger than expected. However, the results looked promising and a second run of the experiment was planned.

In adaptive designs for small sample size experiments with animals or human (patient donor) cells, considerations are obviously distinct from clinical trials. For example, there is typically no accrual time in these experiments, because all animals or patients' cells are available to run through an experiment simultaneously, and great emphasis is placed on reaching significance within that first run. For reasons stated above, we do not pre-plan sample size for the second run, and expect that a second run will be rarely be needed. Instead we spend most of α on the first run, and reserve little for the second run. There is a great cost to set up an experiment a second time, and so is considered an inefficient options by scientists. In other words, it would be difficult to justify to scientists a design with a pre-planned second run of an experiment, when they are convinced a single efficient run should suffice. Therefore, the Chen et al (2004) and Gao et al (2008) large sample size procedures, which assume a preplanned second stage, are not directly relevant to our situation. For example, their findings that type I error can be easily controlled with a seamless test, given promising results from the first stage, cannot be compared directly to our findings.

In consideration of these fundamental distinctions and others of lab experiments, we propose that the main situation scientists wish to avoid is having no mitigation plan when significance is not achieved in the first run of an experiment. Because there is usually a great resource cost in repeating a run of an experiment, investigators invest much effort in collecting sufficient preliminary data to ensure that the design of a single full run of an experiment will provide the desired conclusions. External pilot studies and similar prior experiments (e.g. use of a familiar mouse model) provide confidence and refinement during design and planning. Some investigators go so far as to use a much larger sample size than preliminary data suggest in order to avoid running an experiment a second time. Still, when significance is not achieved, but results look promising, a decision will often be made to repeat the experiment. Investigators then tend to ask statisticians to determine the number of additional replications to combine with the first run to achieve significance, with no recognition that type I error was already spent on the first test. Besides the leukemia cell experiments, we have also found that failed significance occurs in animal experiments that require a long observation period (e.g. when outcomes like mortality, disease onset, or tumor size are measured over months). Also in leukemia cell experiments where patient cell-by

condition random effects are present, it is difficult to anticipate accurately both the variance of response and especially the effect size. Even if these were anticipated correctly, power is usually set to 80%, so that 20% of experiments would fail to reach significance when presumed effect sizes are correct. With pre-planning within this context, we suggest an easy to use method that controls overall type I error. It provides confidence when deciding the size of the second run of an experiment, given the results of the first run look promising (but failed to reach significance). Finally, our experience with more recent grant reviews indicates that thorough justification of animal numbers is becoming more critical, and attention to proper design and reporting of animal experiments is increasing (Festing, 2010). We think that the proposed method satisfies the demand for efficiency and sophistication in the planning of these experiments, while maintaining simplicity.

We are not considering the first stage as an internal pilot (Kairalla et al, 2010; Proschan, 2005) because of our common experience in which the investigator feels that he/she has developed the ideal protocol, through preliminary studies and literature, and has sufficient confidence to attempt a single conclusive experiment. The small size of laboratory experiments and the simultaneous running of all animals or tissue samples also make the use of internal pilots less applicable than for clinical trials. For example, if the experiment only needs a sample size of 6 mice per group to achieve desired power, then an internal pilot on one or two mice would not be sufficiently informative about re-designing the study. Instead an external pilot with a few mice is usually used to refine the design or intervention before settling on a final design. The same can be said about donor cell experiments. Internal pilots can be useful with somewhat larger experiments. For example, in the approach suggested by Denne and Jennison (1999), estimating the unknown variance increases power substantially. An internal pilot in larger experiments could be used to provide the first stage sample size in our approach, and then our suggestions below for planning a second run in the face of failed significance would be applicable.

Posch and Bauer (1999), Lehmacher and Wassmer (1999), Posch et al (2003), and Bretz et al (2009) showed that several adaptive designs can be formulated in terms of a simple conditional error function. We explored two combination approaches in small sample size, one based on the p-value product method introduced by Bauer and Kohn (1994) (BK) and the other based on a p-value inverse normal transformation (IN) discussed in Lehmacher and Wassmer (1999), Posch and Bauer (1999), and Bretz et al (2009). In both approaches we emphasize an upfront choice of α_1 (first run significance level) that is close to α , for reasons stated above. Typically α_1 is set much lower for combination tests, and sometimes is determined as a function of other choices. A futility level (no second run) α_0 must also be chosen for this approach in order to define significance criteria for the second run. For these combination tests, when the second experiment is run because the p-value from the first is between α_1 and α_0 , the p-value from the second experiment is combined with the first, and critical values are chosen to control the overall type I error at α . Here we describe the simple steps in implementing these approaches for small sample sizes and apply it to our motivating example.

In contrast to large sample approaches that fail to control type I error adequately with small sample sizes (Posch and Bauer, 2000; Wassmer, 1998), it is clear that the BK and IN

methods can be easily adapted to control type I error in small sample size experiments. The IN method requires pre-specified weights for the individual run's p-values and numerical integration, but the range of weights can be restricted by the small range we adopt for α_0 to α_1 . In power simulations we show that imprecision of effect size estimates has little consequence for a range of mistaken assumptions about effect size and reasonable choices of α_0 in our context. We were particularly interested in the effect of small degrees of freedom associated with the two stages of the experiment, because in small sample sizes the degrees of freedom associated with the two p-values could reduce power relative to treating the data as if it came from one experiment with larger degrees of freedom. Within our context, power loss is negligible. Our results are therefore similar to what has been shown in large sample sizes (Posch and Bauer, 2000; Wassmer, 1998). The BK and IN methods in small sample size experiments seem to have power that is close to that of an “optimal” t-test from the seamless combination of data from the two experiments as if they were one, despite the possibility of quite different sample sizes in each stage (Bauer and Kohne, 1994). We show that a “naive” method of adding replicates and combining data from two stages seamlessly fails to control type I error, with no important gain in power over BK or IN methods. We also studied the effect of setting α_1 lower than a choice that fits our circumstances in order to uncover any substantial effects on power or total sample size. Finally, we also studied the effects of overly optimistic effect sizes at the first run to determine the ability of the methods to recover from such mistaken assumption.

2 Extension of combination tests for small sample sizes

2.1 Hypothesis Testing

Though our experiment on patient's cells used a one-sample test, we consider first the more common two sample problem. We propose a combination test of the one-sided hypothesis:

$$H_0: \mu_y \leq \mu_x \quad \text{against} \quad H_1: \mu_y > \mu_x,$$

where μ_y and μ_x are the population means from group x and y . Let n_1 denote the proposed sample size in each group for the first run. If the p-value from the first run gives $p_1 \leq \alpha_1$, we stop with a rejection of the null. If $p_1 > \alpha_0$, we stop with no conclusion (presumably loss of interest in pursuing the alternative with a resource demanding second experiment). If $\alpha_1 < p_1 < \alpha_0$, we proceed to the second run. Sample size n_2 can be estimated based on data from the first run. If we proceed to the second run, the null hypothesis H_0 is rejected at the final analysis if $C(p_1, p_2) \leq c$, where p_2 is the p-value based only on the second run data. $C(p_1, p_2)$ is a monotonically increasing combination function of p_1 and p_2 , and c is the critical value that controls overall type I error α .

2.2 Choosing Parameters to Control Type I Error

Overall type I error is controlled by calculation of c , given a preplanned choice of α_0 and α_1 :

$$\alpha = \alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 \mathbf{1}_{[C(p_1 p_2) \leq c]} dp_2 dp_1 = \alpha_1 + \int_{\alpha_1}^{\alpha_0} A(p_1) dp_1, \quad (1)$$

where $\alpha(p_1)$ is the conditional error function defined in the Appendix. For the BK method, H_0 is rejected at the end of the second run if $p_1 \times p_2 \leq c$. With α_0 and α_1 chosen first, c is simply obtained by considering that

$$\alpha = \alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^{c/p_1} dp_2 dp_1 = \alpha_1 + \int_{\alpha_1}^{\alpha_0} \frac{c}{p_1} dp_1 = \alpha_1 + c (\log \alpha_0 - \log \alpha_1). \quad (2)$$

In other references to the BK method, given the overall type I error probability (α) and the utility value for α_0 , a c is chosen to satisfy $c_\alpha = \exp \left\{ - (1/2) \chi_{4,\alpha}^2 \right\}$, which then determines the first run significance level α_1 . In addition in these approaches, first run power is dependent on these choices. Given our need to achieve significance in the first run, our approach must be different. Instead, small sample size experiments (SBK), we choose α_0 and α_1 first so that power of the first run is $1 - \beta$ and the second run sample size does not become unreasonably large.

For the IN method with small sample size (SIN), the weighted inverse normal combination function leads to rejection of H_0 at the end of the second run if

$1 - \Phi \left\{ \omega_1 \Phi^{-1} (1 - p_1) + \omega_2 \Phi^{-1} (1 - p_2) \right\} \leq c$. In order to control type I error at α , c is chosen to satisfy the following equation:

$$\alpha = \alpha_1 + \int_{\alpha_1}^{\alpha_0} 1 - \Phi \left\{ \frac{\Phi^{-1} (1 - c) - \omega_1 \Phi^{-1} (1 - p_1)}{\omega_2} \right\} dp_1. \quad (3)$$

Given α_0 , α_1 , and pre-specified w_1 and w_2 , c must be solved numerically.

2.3 Proposed n_2 Calculation with Small n_1

Let $CP_\delta(n_2, c|p_1)$ denote the conditional probability that p-value, p_2 , based on n_2 observations in each group is less than $A(p_1)$, given p_1 , effect size $(\mu_y - \mu_x)/\sigma = \delta$, and n_1 , where σ is the population standard deviation. This conditional power has been derived for a noncentral t-distribution in the Appendix. Setting $CP_\delta(n_2, c|p_1, n_1) = 1 - \beta$ yields

$$1 - \beta = 1 - G_{2(n_2-1), \eta} \left(t_{2(n_2-1), A(p_1)} \right) \quad (4)$$

$$t_{2(n_2-1), \eta, \beta} = t_{2(n_2-1), 1 - A(p_1)},$$

so that the sample size for the second run, n_2 , should satisfy the above equality. Fixing conditional power and c , one can solve this equation iteratively to obtain n_2 . For combination tests with small sample size, the noncentrality parameter h (or d) must be chosen, but an estimate from the first run will have a wide confidence interval. (Consider that you only need an estimate of δ , when the first run results are not significant). In our context, the β in Equation (4) is the same as overall β . Therefore, we explored how limited information about δ from a small sample size performs in estimating n_2 . With small n_1 , we suggest simply using the estimated effect size from the first run, $\hat{\delta} = (\bar{Y}_1 - \bar{X}_1) / \hat{\sigma}$. We compare this to a more conservative approach, which uses an estimated lower confidence

bound on δ to obtain n_2 . Cumming and Finch (2001) provided a confidence interval for estimated effect size. The lower bound for η , η_L , satisfies the following equation:

$$P \left\{ t_{2(n_1-1), \eta_L} \geq \frac{\bar{Y}_1 - \bar{X}_1}{\hat{\sigma} / \sqrt{n_1/2}} \right\} = p,$$

where p is the probability of the observed t value in the upper tail. Once the lower bound for h is found, the lower bound for δ is $\hat{\eta}_L / \sqrt{n_1/2}$. The confidence intervals are slightly skewed (based on the noncentral t distribution). Despite this asymmetry, the $\hat{\delta}$ and the 50%tile of the confidence interval differs very little when n_1 is above 4. With $\hat{\delta}_L$ from $\hat{\eta}_L$, we solve Equation (4) iteratively to obtain n_2 . We will compare these two choices to show the slight gain in power that is achieved with this more conservative approach, and justify our suggestion of simply using $\hat{\delta}$.

The method to calculate the n_2 is also provided for the one sample problem (e.g. patient donor cell experiments), which is provided in the Appendix. We provide n_2 values in Table 4 for SBK and SIN (in italics) methods as defined in this section for the one-sample problem using the noncentral t for small values of n_1 and the first run result (p_1) assuming a desired power of 80%. This table used $\hat{\delta}$ rather than the lower confidence bound $\hat{\delta}_{25\%}$ because of this conservative approach's higher ratio of n_2/n_1 and its insignificant increase in power described below. For the SBK method, the critical value, $c_{SBK} = 0.0031$ was calculated by solving Equation (2) given $\alpha = 0.025$, $\alpha_1 = 0.02$, and $\alpha_0 = 0.1$. For the SIN method, we used a Riemann sum to approximate the integration and solved Equation (3) numerically. For $\alpha_0 = 0.1$, $c_{SIN} = 0.0145$, given $\omega_1 = \sqrt{1/3}$ and $\omega_2 = \sqrt{2/3}$. Simple programs are available upon request from first author.

2.4 Choice of α_0 and α_1

As argued in the introduction, it is critical to reach significance with one (first stage) run of the experiment, because of set up costs. Therefore, we choose α_1 close to $\alpha = 0.05$ or to 0.025 for the more justifiable one-sided test. n_1 is chosen to provide the desired power (e.g. 80%) for the first run, given α_1 , ignoring the possibility of moving to the second run. Because of the discrete values for power in small sample sizes, it turns out that for the usual choices of effect sizes that generate a small sample size, the n_1 for $\alpha_1 = 0.04$ or 0.02 is either the same as or one larger when α_1 is set to 0.05 or 0.025. Adding one additional animal or patient cell sample in each group is a minor cost relative to the setup of the experiment. In contrast, setting alpha to a much lower level for the first run (and reserving more for the second run) would be unacceptable to scientists, who expect to test at the usual significance level. For example if the p -value from the first run was 0.01, and we required something smaller for α_1 to declare significance, our approach would be rejected. Still, we did explore the effects of a smaller α_1 on second run sample sizes in the simulations.

Choice of α_0 is consequential, because it impacts both overall power and the potential size of n_2 . Larger choice of α_0 allows the second run of an experiment even with an

unexpectedly small $\hat{\delta}$. Given that n_1 was chosen based on a scientifically meaningful d , scientists typically lose interest in expending resources with a much smaller $\hat{\delta}$ than expected. In addition with larger values of α_0 , the n_2 can be so many times larger than n_1 that a scientist would choose to not run a second experiment of that size. Larger α_0 is also associated with a smaller value for the significance criterion c (Equation (1)). In contrast, larger α_0 increases the probability of running a second experiment. This increases power, but not by much unless the true δ is much lower than our first run guess, as we show in simulations below. For example, a reasonable ratio of n_2 to n_1 should be less than 5. For SBK Figure 1(a) provides the ratio of n_2 to n_1 using different α_0 . The plotted n_2 corresponds to $p_1 = \alpha_0$ (maximum n_2 at each α_0). Note that $\alpha_0 \in (0.052, 0.25)$ for $\alpha = 0.05$, $\alpha_1 = 0.04$, and $\alpha_0 \in (0.026, 0.25)$ for $\alpha = 0.025$, $\alpha_1 = 0.02$. For SBK $p_1 \times p_2$, so it only makes sense to consider cases where $\alpha_1 < c$ (because we reject the null if $p_1 \times p_2 < c$). The minimum of $\alpha_0 = 0.052$ or 0.026 were chosen to satisfy $\alpha_1 < c$. From all the curves, $\alpha_0 = 0.1$ seems to keep the n_2/n_1 ratio reasonable. Interestingly Figure 1(a) shows that the n_2/n_1 ratios are almost independent of n_1 . The curves correspond to $n = 4, 6, 8$, and 10 (which correspond to $d = 2.6, 1.9, 1.6$, and 1.4), and are from the bottom to the top in each group, respectively. The just noticeable ordering from bottom to the top for small to large n_1 is explained by subtle effects of degrees of freedom on conditional power. Figure 1(a) used $\hat{\delta}$ rather than the lower confidence bound on $\hat{\delta}(\hat{\delta}_{25\%})$ for calculating n_2 . $\hat{\delta}_{25\%}$ produces much higher and unreasonable ratios as seen in Figure 1(b). We address the slight improvement in power achieved by using the lower bound $\hat{\delta}_{25\%}$ next. With larger choices of α_0 , smaller values of $\hat{\delta}$ will be pursued in a second run, but power improvements seem small in most cases. This is at least partially explained by the fact that higher α_0 produce smaller critical values for c . For example, for $\alpha_1 = 0.04$, $\alpha_0 = 0.1$, $c = 0.011$, while with an increase in $\alpha_0 = 0.2$, $c = 0.0062$. Other examples of the effect are described later in Table 2. Highly similar results could be plotted for SIN.

2.5 Calculation of c for SBK and Choice of Weights for SIN

As shown above, the critical value, c , is calculated by solving the equations to control type I error at α given α_0 , α_1 , and pre-specified conditional error function $\alpha(p_1)$. For the SBK method, Equation (2) gives $c = -(a - \alpha\alpha_1)/\{\log(\alpha_0) - \log(\alpha_1)\}$. For the SIN method, c can only be solved numerically. Moreover, the pre-specified weight w_1 and w_2 must be chosen.

A typical choice in a conventional two-stage design is $w_1 = \sqrt{n_1/(n_1+n_2)}$ and

$w_2 = \sqrt{n_2/(n_1+n_2)}$. However, in our context, we emphasize achieving significance in the first run, so that the second run sample size, n_2 , is not pre-specified. Therefore, we pre-specify w_1 and w_2 based on our expectation of the ratio of n_2/n_1 within our context. We show later in Figure 1 and Table 2 that for a range of reasonable choices of α_0 and α_1 , the expected value of n_2/n_1 will be around 2. So we chose $w_1 = \sqrt{1/3}$ and $w_2 = \sqrt{2/3}$. If the first experiment fails to achieve significance at α_1 , but also produces a p-value below α_0 , we move onto the second run of the experiment. The n_2 then is calculated based on the first run information by solving Equation (4). The n_2 for the SBK method is obtained by solving $t_{2(n_2-1), \eta, \beta} = t_{2(n_2-1), 1-c/p_1}$ given the conditional power and c . For the SIN method, n_2 can

be obtained by solving

$t_{2(n_2-1), \eta, \beta} = t_{2(n_2-1), 1 - \Phi \left\{ \Phi^{-1}(1-c) - \omega_1 \Phi^{-1}(1-p_1) \right\} / \omega_2}$. Clearly, the SBK method in our specific setting is much simpler to use. Further we show that the n_2 for achieving 80% conditional power are only slightly lower for SIN than for SBK (n_2 tables are not provided, but are available on request).

2.6 Approximate Large Sample Size Method for Choosing n_2

For large n_1 and assuming the variance is known (normality of the test statistic instead of the noncentral t), Equation (4) for SBK for the two sample problem has a closed form for

$n_2 = n_1 \left(z_{c/p_1} + z_\beta \right)^2 / z_1^2$. With $\hat{\delta} = z_1 \sqrt{2/n_1}$ from the first run, $n_2 = n_1 \left(z_{c/p_1} + z_\beta \right)^2 / z_1^2$. This corresponds to the equation for estimating n_2 for the BK method in Posch and Bauer (2000). With this large sample size method of estimating n_2 , we of course expected n_2 to be consistently smaller than our noncentral t approach when using $\hat{\delta}$ for small sample sizes, but we were not sure by how much. When $\alpha = 0.025$, $\alpha_1 = 0.02$, $\alpha_0 = 0.1$, and $n_1 = 15$, 25% of the large sample n_2 are the same as, and the remaining 75% are 1 smaller than the n_2 that the noncentral t method produces. For $\alpha_0 = 0.15$, 10% are the same, 80% are 1 smaller than, and 10% are 2 smaller than the noncentral t method. For the alternative large sample size approach for the one-sample problem, the equation is the same because $\hat{\delta} = z_1 / \sqrt{n_1}$.

3 Applying SBK to the motivating example and Simulations

3.1 Applying SBK to our motivating example

Using SBK in our example with $\alpha = 0.025$, $\alpha_1 = 0.02$, and $\alpha_0 = 0.1$, the corresponding $c = 0.0031$ (Equation (2)). For the first run, with a sample size of 6 (n_1), we had 80% power to detect an effect size of 1.6 standard deviations using $\alpha_1 = 0.02$. Because of the discrete values for power in small sample sizes, the n_1 for one-sided $\alpha_1 = 0.02$ is the same as when α_1 was set to 0.025. The p-value from the first run (based on a one sample t-test on differences) was $p_1 = 0.06$, which was within our second run decision range. The $p_1 = 0.06$ corresponded to an effect size estimate of 0.76 standard deviations, so is smaller than expected. Estimating n_2 by solving the equation $t_{n_2-1, \eta, 0.8-t-1, 1-0.031/0.06}$ iteratively, which is based on 80% conditional power, another 12 (See Table 4) patient's cell samples were needed for the second run. After we ran the second stage, the p-value was $p_2 = 0.000035$. Combining these two runs' p-values, $p_1 \times p_2 = 0.060 \times 0.000035 = 0.0000021 < 0.0031 = c$, which gave us the significance we needed for confidently reporting the results.

3.2 Simulations for $E(n_2)$, Type I and II Error

Simulations were used to compare testing options and conditions regarding type I and II errors, the probability of a second run, and sample size for n_2 and total n . The simulations were based on the two sample problem. In our simulations, we estimated the type I and II error and all other parameters using 10,000 replicates of the procedure, with a Monte Carlo error of 0.005.

3.2.1 Type I error for a “naive” Method—A naive procedure might combine n_1 and n_2 observations in a single seamless test when faced with non-significance in an initial experiment and might make no adjustment to the significance criteria. Based on the effect observed in a first run, which fails the significance test, a total sample size, n_{total} , could be calculated so that $n_2 = n_{total} - n_1$. This might be the approach taken by an investigator who naively expects that the seamless significance test still controls type I error. A similar procedure for calculating a n_2 based on conditional power was outlined in Chen et al (2004) and Gao et al (2008) for large sample size. They found that the type I error is not inflated when the conditional power is above approximately 50% given a pre-planned two-stage design, otherwise the significance criteria must be adjusted. We showed above that reasonable choices of α_0 and α_1 can produce an n_2 that is 5 times larger than n_1 , so we did not expect that type I error would be as well controlled as in Chen et al. Note also that our naive approach should be more biased than the “naive” approach defined by Wittes et al (1999), where only the variance parameter drives second run sample size and stopping is not part of the plan. The results in Table 1 show that type I error is not controlled using this naive procedure even with our choices of low values for α_0 . Note that a significance level of 0.025 was used in the naive procedure for the first run test. Simulated type I errors for our SBK and SIN methods are also provided simply for comparison, and are as expected. This naive procedure did not control type I error, mainly because it makes no adjustment for testing twice nor using the first run data to both test the hypothesis and determine n_2 . In the examples we ran, testing twice without adjusting alpha seemed to contribute 40% to the inflated error rate and the capitalizing on chance by using first run data to drive n_2 contributes around 60%. We also used the significance level of 0.02 at the first run for the naive procedure in order to mimic the significance level used in SBK and SIN methods. We found that the type I error is still not controlled, but it is inflated less by around 20% in our examples.

3.2.2 Effect on power of α_0 for SBK when the presumed value of δ is incorrect in the first run—From Figure 2, we let the true δ range from 20% to 140% for the first run guess of effect size (δ_1). The overall power was calculated and compared between the SBK procedure for small sample sizes using $\hat{\delta}$ and the conservative $\hat{\delta}_{25\%}$. Figure 2(a) provides the power of these two choices under a number of different values of d and n_1 with $\alpha_0 = 0.1$. Note that $n_1 = 4, 6, 8$, and 10 are ordered from the top to the bottom in each group of curves. Figure 2(b) is for $\alpha_0 = 0.15$. With these choices, the SBK procedure with conservative $\hat{\delta}_{25\%}$ (dashed lines) did not increase the overall power by much compared to the SBK procedure with $\hat{\delta}$ (solid lines), even though it increased n_2 substantially (see Figure 1(b)). Figure 2(c) provides the probability of moving to the second run using SBK with $\hat{\delta}$ for $\alpha_0 = 0.1$ and $\alpha_0 = 0.15$, respectively. Of course with $\alpha_0 = 0.15$, the probability of having the second run is higher than with $\alpha_0 = 0.1$, but not by much. The probability of having the second run is quite low when the true δ is too much lower or higher than the initial guess (δ_1).

3.2.3 Comparison with other options—In Table 2, the power, the average sample size of n_2 given a second run ($E(n_2|0.02 < p_1 < \alpha_0)$), the average sample size of n_2 overall (including $n_2 = 0$, $E(n_2) = E(n_2|0.02 < p_1 < \alpha_0) \times p_{1 \rightarrow 2}$) and the average total sample size

($E(n)$) are compared among two SBK procedures (using $\hat{\delta}$ or $\hat{\delta}_{25\%}$), the SIN procedure with pre-specified weight, two standard tests (with significance level of 0.02 or 0.025), and the naive procedure (with significance level of 0.025 at the first run). Here $P_{1 \rightarrow 2}$ means the probability of a second run. The first run sample size is $n_1 = 4$ for all procedures, $\alpha_0 = 0.1, 0.15$, and 0.2 , and the true d are 40%, 60%, 80%, 100%, and 120% of the first run guess of δ (δ_1). Table 2 shows the increase in power using the adaptive approach as compared to relying only on the first run (standard test) when δ_1 is incorrect. With its lack of type I error control, the naive procedure had slightly higher power than the SBK procedure, but not much higher in these cases. Compared to the SIN procedure, the SBK procedure using $d^\wedge$ showed similar performance, but slightly larger sample sizes for the second run when the true d is much lower than our first run guess. The expected value of n_2/n_1 will be around 2 with $\alpha_1 = 0.02$. With larger α_0 , there is of course a higher probability of a second run as seen in Figure 2(c). The average sample size of n_2 for SBK with $\hat{\delta}_{25\%}$ were larger than those with $\hat{\delta}$ but showed only a slight power improvement. Posch et al (2004) discuss conditional power using combination tests as compared to a seamless t-test. Their simulation showed considerable loss of power of combination testing at small sample size $n_1 = 2$, but negligible for $n_1 = 10$. In our simulation with $n_1 = 4$, only small differences in power were observed between the biased naive test and the SBK test.

The power, the average sample size of n_2 given a second run ($E(n_2 | \alpha_1 < p_1 < \alpha_0)$), the average total n , and the probability of a second run are also compared between the SBK and SIN procedures using $\hat{\delta}$ for different choices of α_1 . Although we argued that α_1 should be close to α to emphasize the high importance of reaching significance in the first run (e.g. $\alpha = 0.02$), we still wanted to compare this to lower values of α_1 (e.g. $\alpha_1 = 0.01, 0.001$). In Figure 3, the first run sample size is $n_1 = 4$ for all procedures. For fixed α_1 , the SBK and SIN procedures have similar power in Figure 3(a) (For the SBK procedure, $p_1 \times p_2$, it only makes sense to consider cases where $\alpha_1 < c$, because we reject the null if $p_1 \times p_2 < c$. So the critical value, c , for SBK is not available for the case of small α_1 (e.g., $\alpha_1 = 0.001$). There is no similar limitation for SIN.) Where they can be compared, the SBK procedure produces slightly larger sample sizes than SIN for the second run when the true d is much lower than our first run guess, which is similar to what we observed in Table 2. Surprisingly, for both methods, the power is slightly decreased as the α_1 gets smaller. Of course, the probability of having a second run for both SBK and SIN with smaller α_1 is much higher than those with $\alpha_1 = 0.02$ especially when the true d is close to our first run guess in Figure 3(c), and as expected the second run sample size is smaller with smaller α_1 in Figure 3(d). However the overall (total) expected sample size is similar for all choices of α_1 in Figure 3(b). Sometimes this total n is slightly larger and sometimes smaller with smaller α_1 . Therefore, using a smaller α_1 than we suggest (thus conflicting with our emphasis on achieving significance in the first run), provides no efficiency benefit in our context. We did notice that the expected value of n_2/n_1 is around 1 when $\alpha_1 = 0.01$ or 0.001 , which is smaller than the ratio for the larger α_1 . For this reason, we also ran the simulations for SIN with equal weights

($\omega_1 = \omega_2 = \sqrt{1/2}$). However, the results for power, $E(n | \alpha_1 < p_1 < \alpha_0)$, and $P_{1 \rightarrow 2}$ are similar

to the SIN procedure with the weights of $\omega_1 = \sqrt{1/3}$ and $\omega_2 = \sqrt{2/3}$. Quite similar results are found for much larger first run sample size ($n_1 = 48$).

4 Discussion

The extra effort of applying our proposed approach to laboratory experiments presumes that type I error control is critical for reaching unambiguous conclusions about results of these experiments. In our experience, scientific reviewers require this level of control in both publications and the planning of grants. Thus this method provides a sufficiently sophisticated approach for such reviews, but is not so complex that it is beyond the intuitive understanding by scientists of the statistical issues raised by adding replications. It also responds to calls for more efficient use of animals.

Within our context, which highly values running an experiment once to avoid the cost of repeating the experimental set up, considerations for sample size estimation for a potential second run are surprisingly simple. Based on reasonable choices of parameters for experiments and our simulations, we showed that the multi-dimensional considerations that could be complex in other contexts were not so in our context. Although conditional power could be low when the initial guess of effect size is overly optimistic, our requirement that p_1 be small (corresponding to lower futility values of α_0) before considering a second run, appropriately limits second run sample size and reduces the differences in power among options such as conservative choices of effect size estimation, or a naive method that does not control type I error with two tests.

The choice of weights for SIN discussed in section 2.5 is based on an expectation of the second stage sample size, n_2 . If the actual new sample size is substantially different from expected, the weights could lower the efficiency of the test. Note that subjects in one stage will have more weight than in the other stage, which has been controversial in human trials.

In planning a potential adaptive experiment with emphasis on achieving significance in the first run, this SBK method also requires choice of the first run significance level (α_1), which then determines the combined results significance criterion c . Choosing α_1 close to the usual significance level rather than a much smaller value, as is typical in adaptive clinical trials, is consistent with the scientists need to avoid setting up a costly experiment a second time. In addition, this choice produces slightly more power than choosing smaller α_1 with no increase in total sample size in our context. If one wishes to use a non-parametric test to obtain p-values, the sample sizes can be obtained through the inflation factor method as outlined in PASS 2008 (NCSS LLC, Kaysville, Utah, USA).

When a second run is needed to reach significance, we know that the estimate of effect size will be biased when results from the two runs are combined (Bauer and Kohne, 1994). We suggest that only the results from the first run be used for estimation, and the combination of data be reserved for hypothesis testing. Of course, the focus of laboratory experiments is on hypotheses testing rather than point estimation. In addition it is unusual to use effect estimates from laboratory experiments for planning future ones, but coefficients of variation often are used for sample size calculations, and obtaining these from the first run makes

most sense. Although estimation based on first stage data is unbiased, it does exclude additional data from the second stage, and this could produce a confidence interval that is inconsistent with the hypothesis test.

Acknowledgements

This work is partially supported by Award Number UL1RR025755 from the National Center for Research Resources. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center for Research Resources or the National Institutes of Health. We appreciate the comments of Drs. Xiaobai Li and Lianbo Yu.

Appendix

Conditional Error function

Posch and Bauer (1999) and Bretz et al (2009) defined a conditional error function for adaptive combination tests as:

$$A(p_1) = \begin{cases} 0 & \text{if } p_1 \geq \alpha_0 \\ \max_{x \in [0,1]} \left\{ x \cdot \mathbf{1}_{[C(p_1, x) \leq c]} \right\} & \text{if } \alpha_1 < p_1 < \alpha_0, \\ 1 & \text{if } p_1 \leq \alpha_1 \end{cases},$$

where the indication function $\mathbf{1}_{[\cdot]}$ is 1 if $C(p_1, x) \leq c$ and 0 otherwise. For example, the BK approach with Fisher's product test and the IN approach with the weighted inverse normal combination have the following conditional error functions,

$$A(p_1) = \begin{cases} 0 & \text{if } p_1 \geq \alpha_0 \\ \frac{c}{p_1} & \text{if } \alpha_1 < p_1 < \alpha_0 \\ 1 & \text{if } p_1 \leq \alpha_1 \end{cases}$$

and

$$A(p_1) = \begin{cases} 0 & \text{if } p_1 \geq \alpha_0 \\ 1 - \Phi \left\{ \frac{\Phi^{-1}(1-c) - \omega_1 \Phi^{-1}(1-p_1)}{\omega_2} \right\} & \text{if } \alpha_1 < p_1 < \alpha_0 \\ 1 & \text{if } p_1 \leq \alpha_1 \end{cases},$$

where w_1 and w_2 are pre-specified weight such that $\omega_1^2 + \omega_2^2 = 1$ and Φ denotes the cumulative distribution function of the standard normal distribution.

Conditional Power

Under normality

$$\begin{aligned} CP_\delta(n_2, c | p_1, n_1) &= \{p_2 < A(p_1) | p_1, n_1, \delta\} \\ &= P \left\{ \frac{(\bar{Y}_2 - \bar{X}_2)}{\sqrt{\frac{2\sigma^2}{n_2}}} > t_{A(p_1)} | p_1, n_1, \delta \right\}. \end{aligned}$$

where $\bar{Y}_2 - \bar{X}_2 \sim N(\delta\sigma, 2\sigma^2/n_2)$ and $2(n_2 - 1)\hat{\sigma}^2/\sigma^2 \sim \chi_{2(n_2-1)}^2$ are independent, and $\hat{\sigma} = \sqrt{(\hat{\sigma}_{x_2}^2 + \hat{\sigma}_{y_2}^2)/2}$ ($\hat{\sigma}_{x_2}^2$ and $\hat{\sigma}_{y_2}^2$ are the sample variances from the second run data for group x and y , respectively). Thus $(\bar{Y}_2 - \bar{X}_2) / \sqrt{2\hat{\sigma}^2/n_2}$ has a noncentral t -distribution with degrees freedom of $2(n_2 - 1)$ and noncentrality parameter $\eta = \delta\sigma / \sqrt{2\sigma^2/n_2} = \delta \sqrt{n_2/2}$. We can then write

$$CP_{\delta}(n_2, c|p_1, n_1) = 1 - G_{2(n_2-1), \eta}(t_{2(n_2-1), A(p_1)}),$$

where $G_{2(n_2-1), \eta}(\cdot)$ denotes the cumulative distribution function of the noncentral t -distribution with $2(n_2-1)$ degrees of freedom and noncentrality parameter η .

Determination of n_2 for One Sample Patient Donor Cell Experiments

For our experiment on patient's cells, the null and alternative hypotheses are $H_0: \mu \leq 0, H_1: \mu > 0$, and the per patient contrast has $C \sim N(\mu, s^2)$, where μ is the population mean and s^2 is the population variance. The conditional power is

$$CP_{\delta}(n_2, c|p_1, n_1) = P\left\{\frac{\bar{C}_2}{\sqrt{\hat{\sigma}^2/n_2}} > t_{A(p_1)} | p_1, n_1, \delta\right\},$$

where effect size $\delta = \mu/\sigma$, and $\bar{C}_2 \sim N(\delta\sigma, \sigma^2/n_2)$ and $(n_2 - 1)\hat{\sigma}^2/\sigma^2 \sim \chi_{n_2-1}^2$ are independent. Therefore, $\bar{C}_2 / \sqrt{\hat{\sigma}^2/n_2}$ has a noncentral t -distribution with degrees freedom of n_2-1 and noncentrality parameter $\eta = \sqrt{n_2}\delta$. This leads to $CP_{\delta}(n_2, c|p_1, n_1) = 1 - G_{n_2-1, \eta}(t_{n_2-1, A(p_1)})$. Therefore, the sample size for the second run, n_2 , can be calculated by solving the equation $t_{n_2-1, \eta, \beta} = t_{n_2-1, 1-A(p_1)}$. For the SBK method, n_2 is obtained by solving the equation $t_{n_2-1, \eta, \beta} = t_{n_2-1, 1-c/p_1}$.

References

- Bauer P, Kohne K. Evaluation of experiments with adaptive interim analyses. *Biometrics*. 1994; 50:1029–1041. [PubMed: 7786985]
- Bretz F, Koenig F, Brannath W, Glimm E, Posch M. Tutorial in biostatistics: Adaptive designs for confirmatory clinical trials. *Statistics In Medicine*. 2009; 28:1181–1217. [PubMed: 19206095]
- Chen YHJ, DeMets DL, Lan KKG. Increasing the sample size when the un-blinded interim result is promising. *Statistics in Medicine*. 2004; 23:1023–1038. [PubMed: 15057876]
- Cumming G, Finch S. A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*. 2001; 61:532–574.
- Denne JS, Jennison C. Estimating the sample size for a t-test using an internal pilot. *Statistics In Medicine*. 1999; 18:1575–1585. [PubMed: 10407230]
- Festing M. Statistics and animals in biomedical research. *Significance*. 2010; 7:176–177.

- Gao P, Ware JH, Mehta C. Sample size re-estimation for adaptive sequential design in clinical trials. *Journal of Biopharmaceutical Statistics*. 2008; 18:1184–1196. [PubMed: 18991116]
- Kairalla J, Muller K, Coffey C. Combining an internal pilot with an interim analysis for single degree of freedom tests. *Communication in Statistics - Theory and Methods*. 2010; 39:3717–3738.
- Lehmacher W, Wassmer G. Adaptive sample size calculation in group sequential trials. *Biometrics*. 1999; 55:1286–12,901. [PubMed: 11315085]
- Posch M, Bauer P. Adaptive two stage designs and the conditional error function. *Biometrical Journal*. 1999; 41:689–696.
- Posch M, Bauer P. Interim analysis and sample size reassessment. *Biometrics*. 2000; 56:1170–1176. [PubMed: 11129475]
- Posch M, Bauer P, Brannath W. Issues in designing flexible trials. *Statistics In Medicine*. 2003; 22:953–969. [PubMed: 12627412]
- Posch M, Timmesfeld N, Konig F, Muller HH. Conditional rejection probabilities of student's t-test and design adaptations. *Biometrical Journal*. 2004; 46:389–403.
- Proschan MA. Two-stage sample size re-estimation based on a nuisance parameter: a review. *Journal of Biopharmaceutical Statistics*. 2005; 15:559–574. [PubMed: 16022163]
- Wassmer G. A comparison of two methods for adaptive interim analyses in clinical trials. *Biometrics*. 1998; 54:696–705. [PubMed: 9629649]
- Wittes J, Schabenberger O, Zucker D, Brittain E, M P. Internal pilot studies i: type i error rate of the naive t-test. *Statistics In Medicine*. 1999; 18:3481–3491. [PubMed: 10611620]

Fig. 1.

The effect of α_0 on the ratio n_2/n_1 . (a) used $\hat{\delta}$ for calculating n_2 ; (b) used the lower confidence bound on $\hat{\delta}$ ($\hat{\delta}_{25\%}$) and $\hat{\delta}$ curves from (a) simply carried over for reference. For each group of curves, from the bottom to the top are for $n_1 = 4, 6, 8$, and 10, respectively. n_2 achieves conditional power of 80%.

Fig. 2.

The overall power and probability of a second run using SBK with $\alpha = 0.025$. Horizontal axis represents the true value of d as a % of the first run guess of d (d_1). (a) shows the overall power with $\alpha_0 = 0.1$ using $\hat{\delta}$ or $\hat{\delta}_{25\%}$; (b) shows the overall power with $\alpha_0 = 0.15$ using $\hat{\delta}$ or $\hat{\delta}_{25\%}$; (c) shows the probability of a second run using $\hat{\delta}$ with $\alpha_0 = 0.1$ and $\alpha_0 = 0.15$. For each group of curves, from the top to the bottom are $n_1 = 4, 6, 8$, and 10 , respectively.

Fig. 3.

Simulation results using SBK and SIN for different choices of $\alpha_1 = 0.02, 0.01, 0.001$. First run guesses of d ($d_1 = 2.6$ standard deviations), $n_1 = 4$, $\alpha = 0.025$, and $\alpha_0 = 0.1$. Horizontal axis represents the true value of d as a % of the first run guess of δ (δ_1). $\omega_1 = \sqrt{1/3}$ and $\omega_2 = \sqrt{2/3}$ were used for SIN. (a) shows the overall power; (b) shows the average total sample size; (c) shows the probability of a second run; (d) shows the average sample size of the second run given a second run.

Table 1

Simulation results comparing the type I error of the naive procedure, SBK, and SIN using $\hat{\delta}$ with $\alpha = 0.025$, $\alpha_1 = 0.02$, and $\alpha_0 = 0.1, 0.15$ for the two sample problem. Standard error of simulation is 0.0016. For SBK, the critical value, $c_{SBK} = .0031$ for $\alpha_0 = 0.1$ and $c_{SBK} = .0025$ for $\alpha_0 = 0.15$. For SIN, $c_{SIN} = .0145$ for $\alpha_0 = 0.1$ and $c_{SIN} = .0109$ for $\alpha_0 = 0.15$ are numerically obtained with $\omega_1 = \sqrt{1/3}$ and $\omega_2 = \sqrt{2/3}$.

		Naive Procedure (0.025)	SBK	SIN
$\alpha_0 = 0.1$	$n_1 = 4$	0.0405	0.0252	0.0258
	$n_1 = 6$	0.0341	0.0246	0.0235
	$n_1 = 8$	0.0332	0.0239	0.0249
	$n_1 = 10$	0.0342	0.0246	0.0251
$\alpha_0 = 0.15$	$n_1 = 4$	0.0402	0.0266	0.0252
	$n_1 = 6$	0.0373	0.0243	0.0242
	$n_1 = 8$	0.0377	0.0250	0.0243
	$n_1 = 10$	0.0329	0.0242	0.0241

Table 2

Simulation results comparing the power of different procedures and different first run guesses of δ ($\delta_1 = 2.6$ standard deviations), $\alpha = 0.025$, $\alpha_1 = 0.02$, and $\alpha_0 = 0.1, 0.15, 0.2$ for the two sample problem. For SBK, the critical value, $c_{SBK} = .0031$ for $\alpha_0 = 0.1$, $c_{SBK} = .0025$ for $\alpha_0 = 0.15$, and $c_{SBK} = .0022$ for $\alpha_0 = 0.2$. For SIN, $c_{SIN} = .0145$ for $\alpha_0 = 0.1$, $c_{SIN} = .0109$ for $\alpha_0 = 0.15$, and $c_{SIN} = .0095$ for $\alpha_0 = 0.2$ are numerically obtained with $\omega_1 = \sqrt{1/3}$ and $\omega_2 = \sqrt{2/3}$. $P_{1 \rightarrow 2}$ is the probability of a second run.

	True δ	$\alpha_0 = 0.1$					$\alpha_0 = 0.15$					$\alpha_0 = 0.2$				
		0.4 δ_1	0.6 δ_1	0.8 δ_1	δ_1	1.2 δ_1	0.4 δ_1	0.6 δ_1	0.8 δ_1	δ_1	1.2 δ_1	0.4 δ_1	0.6 δ_1	0.8 δ_1	δ_1	1.2 δ_1
power	standard (0.02)	0.20	0.41	0.64	0.83	0.94	0.20	0.41	0.64	0.83	0.94	0.20	0.41	0.64	0.83	0.94
	standard (0.025)	0.24	0.46	0.69	0.86	0.95	0.24	0.46	0.69	0.86	0.95	0.24	0.46	0.69	0.86	0.95
	SBK ($\hat{\delta}$)	0.43	0.73	0.92	0.98	1	0.52	0.81	0.95	0.99	1	0.59	0.86	0.97	1	1
	SBK ($\hat{\delta}_{25\%}$)	0.52	0.77	0.93	0.98	1	0.62	0.85	0.96	0.99	1	0.70	0.90	0.98	1	1
	SIN	0.42	0.74	0.92	0.98	1	0.51	0.81	0.95	0.99	1	0.60	0.86	0.97	1	1
	Naïve	0.46	0.76	0.92	0.98	1	0.56	0.82	0.96	0.99	1	0.64	0.88	0.97	1	1
$E(n_2 0.02 < p_1 < \alpha_0)$	SBK ($\hat{\delta}$)	8.02	7.31	6.74	6.02	5.51	12.1	10.3	9.10	7.55	6.54	16.9	13.4	10.7	8.80	6.78
	SBK ($\hat{\delta}_{25\%}$)	26.0	22.7	20.0	16.8	14.7	55.8	42.9	35.3	24.9	19.1	143	94.2	60.5	39.2	21.3
	SIN	7.78	7.20	6.65	6.08	5.77	11.7	10.3	8.93	7.79	6.47	16.1	13.3	10.6	9.04	7.11
	Naïve	6.80	6.43	5.89	5.50	4.93	9.36	8.55	7.30	6.43	5.47	12.5	10.3	8.50	6.77	5.81
$P_{1 \rightarrow 2}$	SBK	0.33	0.38	0.28	0.15	0.06	0.44	0.44	0.32	0.17	0.06	0.52	0.50	0.34	0.17	0.06
	SIN	0.32	0.37	0.28	0.15	0.06	0.44	0.45	0.32	0.17	0.06	0.51	0.49	0.34	0.17	0.06
	Naïve	0.30	0.33	0.23	0.12	0.04	0.41	0.40	0.27	0.13	0.05	0.49	0.44	0.29	0.14	0.04
	SBK ($\hat{\delta}$)	2.65	2.75	1.91	0.91	0.35	5.33	4.58	2.91	1.29	0.39	8.74	6.70	3.63	1.47	0.44
$E(n_2)$	SBK ($\hat{\delta}_{25\%}$)	8.59	8.54	5.65	2.55	0.93	24.5	19.0	11.3	4.26	1.15	74.3	47.0	20.5	6.56	1.38
	SIN	2.52	2.64	1.87	0.91	0.36	5.11	4.65	2.84	1.31	0.41	8.26	6.52	3.58	1.52	0.46
	Naïve	2.02	2.10	1.37	0.67	0.21	3.84	3.39	1.98	0.85	0.26	6.09	4.53	2.45	0.92	0.25
	SBK ($\hat{\delta}$)	6.65	6.75	5.91	4.91	4.35	9.33	8.58	6.91	5.29	4.39	12.74	10.70	7.63	5.47	4.44
$E(n)$	SBK ($\hat{\delta}_{25\%}$)	12.59	12.54	9.65	6.55	4.93	28.5	23.0	15.3	9.26	5.15	78.3	51.0	24.5	10.56	5.38
	SIN	6.52	6.64	5.87	4.91	4.36	9.11	8.65	6.84	5.31	4.41	12.26	10.52	7.58	5.52	4.46
	Naïve	6.02	6.10	5.37	4.67	4.21	7.84	7.39	5.98	4.85	4.26	10.09	8.53	6.45	4.92	4.25

Table 3

Critical values for the SBK and SIN procedures, with $\alpha = 0.025$, $\alpha_1 = 0.02, 0.01, 0.001$, and $\alpha_0 = 0.1, 0.15, 0.2$ for the two sample problem.

	<u>$\alpha_1 = 0.02$</u>			<u>$\alpha_1 = 0.01$</u>			<u>$\alpha_1 = 0.001$</u>		
	$\alpha_0 = 0.1$	$\alpha_0 = 0.15$	$\alpha_0 = 0.2$	$\alpha_0 = 0.1$	$\alpha_0 = 0.15$	$\alpha_0 = 0.2$	$\alpha_0 = 0.1$	$\alpha_0 = 0.15$	$\alpha_0 = 0.2$
c_{SBK}	0.0031	0.0025	0.0022	0.0065	0.0055	0.0050	NA	NA	NA
c_{SIN}	0.0145	0.0109	0.0095	0.0392	0.0292	0.0248	0.0556	0.0416	0.0353

Table 4

The second run sample size, n_2 , that achieves conditional power of 80% as a function of p_1 and n_1 , for a one sample t-test with one-sided $\alpha = 0.025$ and $\alpha_0 = 0.1$. For each n_1 , the first row uses the SBK method and the second row in italics uses the SIN method. For SBK, the critical value, $c_{SBK} = .0031$. For SIN, $c_{SIN} = .0145$ is numerically obtained with $\omega_1 = \sqrt{1/3}$ and $\omega_2 = \sqrt{2/3}$.

		p_1																
		n_1	.025	.03	.035	.04	.045	.05	.055	.06	.065	.07	.075	.08	.085	.09	.095	.1
$\alpha=.025$	3	3	3	3	3	4	4	4	5	5	5	6	6	7	7	8	8	9
		3	3	3	3	4	4	4	4	5	5	5	6	6	7	7	8	8
$\alpha_1=.02$																		
$\alpha_0=.1$	4	3	4	4	4	5	5	6	7	7	8	9	9	10	11	12	12	13
$\beta=0.2$	4	4	4	5	5	5	5	6	6	7	8	8	9	9	10	11	12	12
	5	4	5	6	6	6	7	8	9	10	11	12	13	14	15	16	17	18
		5	5	6	6	7	7	8	9	9	10	11	12	13	14	15	16	17
	6	5	6	7	8	9	10	10	11	12	14	15	16	17	18	20	21	23
		6	7	7	8	9	10	11	12	13	14	15	16	17	18	20	21	
	7	6	7	9	10	11	12	14	15	16	18	19	21	22	24	26	27	
		7	8	9	10	11	12	13	14	16	17	18	19	21	22	24	25	
	8	7	9	10	11	13	14	16	18	19	21	23	24	26	28	30	32	
		8	9	10	12	13	14	15	17	18	20	21	23	24	26	28	29	
	9	8	10	11	13	15	17	18	20	22	24	26	28	30	32	34	36	
		9	11	12	13	15	16	18	19	21	23	24	26	28	30	32	34	
	10	9	11	13	15	17	19	21	23	25	27	29	31	33	36	38	41	
		10	12	13	15	17	18	20	22	24	25	27	29	31	33	36	38	
	11	10	12	14	16	19	21	23	25	28	30	32	35	37	40	42	45	
		12	13	15	17	18	20	22	24	26	28	30	32	35	37	40	42	
	12	11	14	16	18	21	23	25	28	30	33	35	38	41	44	47	50	
		13	15	16	18	20	22	24	27	29	31	33	36	38	41	43	46	
	13	12	15	17	20	22	25	28	30	33	36	39	42	45	48	51	54	
		14	16	18	20	22	24	27	29	31	34	36	39	42	44	47	50	
	14	13	16	19	22	24	27	30	33	36	39	42	45	48	52	55	59	
		15	17	20	22	24	26	29	31	34	37	39	42	45	48	51	55	

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

	59	55	52	49	45	42	39	37	34	13	29	26	24	21	19	16	
63	59	56	52	49	45	42	39	35	32	29	26	23	20	17	14	15	
1	560	09	580	08	570	07	590	90	550	50	540	04	535	03	525	1 u	
$\overline{1d}$																	