Published in final edited form as:

Stat Methods Med Res. 2018 April ; 27(4): 1045–1055. doi:10.1177/0962280216650918.

Exact confidence limits for the response rate in two-stage designs with over- or under-enrollment in the second stage

Guogen Shan

Department of Environmental and Occupational Health, School of Community Health Sciences, University of Nevada Las Vegas, Las Vegas, NV, USA

Abstract

Simon's two-stage design has been widely used in early phase clinical trials to assess the activity of a new investigated treatment. In practice, the actual sample sizes do not always follow the study design precisely, especially in the second stage. When over- or under-enrollment occurs in a study, the original critical values for the study design are no longer valid for making proper statistical inference in a clinical trial. The hypothesis for such studies is always one-sided, and the null hypothesis is rejected when only a few responses are observed. Therefore, a one-sided lower interval is suitable to test the hypothesis. The commonly used approaches for confidence interval construction are based on asymptotic approaches. These approaches generally do not guarantee the coverage probability. For this reason, Clopper-Pearson approach can be used to compute exact confidence intervals. This approach has to be used in conjunction with a method to order the sample space. The frequently used method is based on point estimates for the response rate, but this ordering has too many ties which lead to conservativeness of the exact intervals. We propose developing exact one-sided intervals based on the p -value to order the sample space. The proposed approach outperforms the existing asymptotic and exact approaches. Therefore, it is recommended for use in practice.

Keywords

Clopper-Pearson approach; coverage probability; exact one-sided interval; two-stage design

1 Introduction

Multi-stage designs, especially two-stage designs, have been widely used in clinical trials to allow an on-going study to be stopped early to protect participants' safety and improve studies' efficiency. In cancer clinical trials, Simon's two-stage design¹ is traditionally used to investigate the efficacy and safety of a new therapeutic treatment in a one-arm study. The outcome from such studies is often binary: response or no response to the new treatment. In practice, the actual final sample size could be different from the planned values due to some

Reprints and permissions: sagepub.co.uk/journalsPermissions.nav

Corresponding author: Guogen Shan, School of Community Health Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA. guogen.shan@unlv.edu.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

practical reasons, such as an over- or under-estimated dropout rate, a significant budget cut, or an extended study time. The first interim analysis can always be performed when the planned sample size for the first stage is reached. Therefore, the sample size deviation often happens in the second stage. Due to the unplanned sample size in the second stage, the original critical value is no longer valid for making proper statistical inference about the activity of the new treatment.

Upon completion of the study, the response rate and the associated confidence interval are always reported. The response rate is often estimated by using the maximum likelihood method, which is the number of responses divided by the number of participants in the study. However, this maximum likelihood estimate (MLE) is biased.^{2,3} Later, Jung and Kim² derived the uniformly minimum variance unbiased estimator (UMVUE) for the response rate by using the Rao–Blackwell theorem. They showed that the unbiased estimate has comparable efficiency when compared to the biased MLE. They also showed that the UMVUE creates a natural ordering of the sample space by the total number of responses.

In addition to the point estimate, the confidence interval should also be calculated and reported. This becomes very important in the case when a study does not follow the design precisely. Since a study is designed in a two-stage fashion, the confidence interval calculation should incorporate this design information. For a study with under- or over-enrollment in the second stage, Koyama and Chen⁴ proposed a method to control for the conditional type I error in the unconditional p -value calculation and the confidence interval calculation. Later, Zhao et al.⁵ conducted a simulation study and showed that the confidence interval based on the approach by Koyama and Chen⁴ does not guarantee the nominal coverage probability. In addition, they pointed out that the approach by Koyama and Chen has difficulty in computing confidence interval for some sample points. Then, they proposed a new approach based on the likelihood ratio to construct the confidence interval for the response rate. However, all these intervals are asymptotic, and they do not guarantee the coverage probability.

To make a confidence interval exact, Clopper and Pearson⁶ (CP) approach is traditionally used. This approach has to be used in conjunction with a method to order the sample space. The aforementioned point estimates, including the MLE and the UMVUE, can be used to order the sample space. Although these two point estimates are not the same, the sample space ordering based on them is the same as the sample space is sorted by the total number of responses.⁷ The confidence interval based on this ordering when using the CP approach is exact,⁷ but this may not be a good interval estimate as many ties exist in this ordering. To further improve the confidence interval estimate, we propose using the p -value to form the new sample space ordering. Ordering based on the p -value may be able to break down the ties observed from the exact interval based on the MLE.

The rest of the article is organized as follows. In Section 2, we first introduce the construction of one-sided lower limits based on the approach by Koyama and Chen⁴ for a study with under- or over-enrollment in the second stage. Then, we introduce the exact interval based on the MLE and propose the exact lower limit based on the p -value by using the CP approach. At the end of Section 2, we illustrate the application of the considered

asymptotic and exact intervals by using an example. In Section 3, we compare the two exact lower limits with regard to simple average length (AL) and expected length (EL) by using all sample points in the sample space, and the winner is then compared with Koyama and Chen's approach using the partial of the sample space. We use a Phase II clinical trial to illustrate the application of these approaches to construct exact lower limits. Finally, we provide some remarks in Section 4.

2 Methods

In practice, it is not uncommon that the actual sample size of a clinical trial deviates from the planned design. In a two-stage design setting, the interim analysis can always be conducted when the pre-determined first stage sample size is reached. However, the second stage sample size could be different than planned. Suppose n_1 and n_2 are the planned first stage and second stage sample sizes in Simon's design for testing the one-sided hypotheses

$$H_0: \pi \leq \pi_0$$

against the alternative

$$H_a: \pi \geq \pi_1$$

where π_0 and π_1 are the unacceptable and acceptable response rates, respectively. In addition to the sample sizes, the critical values for the first stage and both stages combined, r_1 and r_t , also need to be determined in the design search to guarantee the type I (α) and II (β) rates. Let X_1 and X_2 be the number of responses from the first stage and the second stage, respectively. As in Simon's design, the study will be stopped in the first stage if $X_1 \leq r_1$. When $X_1 > r_1$, the study proceeds to the second stage. The new treatment is considered promising for further investigation when the total number of responses is more than the critical value: $X_1 + X_2 > r_t$.

When a trial goes to the second stage, the unconditional p -value function is defined as

$$P_U(\pi) = \sum_{a=r_1+1}^{n_1} P_\pi(X_1=a|n_1) P_\pi(X_2 \geq r_t - a + 1 | n_2)$$

where $P_\pi(X_1 = a | n_1) = b(a, n_1, \pi)$, $P_\pi(X_2 \geq r_t - a + 1 | n_2) = (1 - B(r_t - a, n_2, \pi))$, and $b(\cdot)$ and $B(\cdot)$ are the probability density function and the cumulative distribution function of a binomial distribution, respectively. When a study follows the design precisely with sample sizes (n_1, n_2) , type I error rate is guaranteed: $P_U(\pi_0) = \alpha$.

When the actual sample sizes are n_1 and n_2' in a study, the pre-determined critical values are no longer valid to make statistical inference. For this reason, Koyama and Chen⁴ proposed computing the unconditional p -value by controlling the new conditional p -value

$P(X'_2 \geq x'_2 | X_1 = x_1)$ bounded by the original conditional p -value $P(X_2 \leq x_2 | X_1 = x_1)$, where X'_2 is the number of responses observed from n'_2 participants, and it follows a binomial distribution with parameters n'_2 and π . Suppose the observed responses are x_1 and x_2 for the first stage and the second stage for a study with sample sizes n_1 and n'_2 . Koyama and Chen⁴ proposed finding the solution of π^* for any given π

$$P_{\pi^*}(X_2 \geq r_t - x_1 + 1 | n_2) = P_{\pi}(X'_2 \geq x'_2 | n'_2) \quad (1)$$

The solved π^* value is then used in the p -value function for $P_{\pi^*}(X_2 \leq r_t - X_1 + 1 | n_2)$, where $X_1 > r_1$. It should be noted that π^* is the solution based on the one observed data point (x_1, x_2) . Then, the unconditional p -value function for a study with over- or under-enrollment is calculated as

$$P_U(\pi, \pi^*) = \sum_{a=r_1+1}^{n_1} P_{\pi}(X_1 = a | n_1) P_{\pi^*}(X_2 \geq r_t - a + 1 | n_2)$$

The one-sided lower $(1 - \alpha)$ confidence interval is the collection of π such that

$$P_U(\pi, \pi^*) > \alpha.$$

This approach is referred to as the KC approach. In this approach, π^* needs to be determined for each possible π . Therefore, it could be computationally intensive.

It has been noted by other researchers^{5,8} that no confidence interval exists when $x_1 > r_1$, since the left side of equation (1) is always 1 regardless of the value of π^* . For a similar reason, when $x_2 = 0$, the right side of equation (1) is always 1; therefore, there is no confidence interval for sample points $(x_1, 0)$ based on the KC approach. In addition, this approach only applies to the trials proceeding to the second stage. The confidence interval calculation for trials stopped in the first stage is not discussed by Koyama and Chen.⁴

For a one-sided hypothesis, a one-sided confidence interval is statistically desirable to make sure that the hypothesis testing and the confidence interval construction are consistent.⁴ When the significance level is α , a $1 - \alpha$ one-sided interval, $(L, 1)$ should be computed for statistical inference, where L is the $1 - \alpha$ lower limit.

CP⁶ approach is traditionally used to construct exact confidence intervals. This approach has to be used in conjunction with a method to order the sample space. The ordering of the sample space is also called the stochastic ordering. The most commonly used stochastic ordering methods are based on parameter estimates. Let m be the stopping stage. When $m=1$, a trial is stopped in the first stage, and the possible outcomes are $S = \{0, 1, 2, \dots, r_1\}$. When $m=2$, the possible outcomes are $G = \{(X_1, X_2), r_1 + 1 - X_1 - n_1, 0 - X_2 - n_2\}$. The

first estimate considered is the naive response estimate as $\pi_n = X_1/n_1$ when $m=1$, and $\pi_n = (X_1 + X_2)/(n_1 + n_2)$ when $m=2$. Due to the sequential nature of a two-stage design, this naive estimate is biased.² Later, Jung and Kim² derived the UMVUE for the response rate by using the Rao–Blackwell theorem. It is calculated as $\pi_u = \pi_n$ when $m=1$ and

$$\pi_u = \frac{\sum_{a=(r_1+1) \vee (X_1+X_2-n_2)}^{(X_1+X_2) \wedge n_1} \binom{n_1-1}{a-1} \binom{n_2}{X_1+X_2-a}}{\sum_{a=(r_1+1) \vee (X_1+X_2-n_2)}^{(X_1+X_2) \wedge n_1} \binom{n_1}{a} \binom{n_2}{X_1+X_2-a}}, \quad \text{when } m=2$$

where $w \vee q = \max(w, q)$ and $w \wedge q = \min(w, q)$. It can be seen that

$\pi_u(X_1, X_2) = \pi_u(X_1^*, X_2^*)$ is satisfied when $X_1 + X_2 = X_1^* + X_2^*$. Therefore, when the point estimate π_u is used for stochastic ordering, sample points with the same total number of responses would have the same lower limit. In this scenario, the ordering based on the UMVUE π_u is the same as that based on the naive estimate π_n , and the sample space is sorted by the total number of responses. This approach based on the total number of response to compute exact intervals is referred to as the R approach. For sample points in set S , their total numbers of responses are always less than or equal to r_1 , while sample points in set G have the number of responses at least $r_1 + 1$. It follows that the lower limits for sample points in set G are larger than those in set S .

Although the stochastic ordering based on parameter estimates is traditionally used, it is often associated with many ties in the ordering. For this reason, we propose using the p -value calculated from the following tail area to create a new stochastic ordering

$$\Omega(x_1, x_2) = \{(X_1, X_2') : X_1 \geq x_1, X_1 + X_2' \geq x_1 + x_2\}$$

This approach is referred to as the P approach as the p -value is used for sample space ordering. The p -value for the observed sample point (x_1, x_2) in set G is computed as

$$P(x_1, x_2) = \sum_{a=x_1}^{n_1} P_{\pi_0}(X_1=a|n_1) P_{\pi_0}(X_2' \geq x_1 + x_2 - a | n_2')$$

and p -value for sample point x_1 in set S is calculated as

$$P(x_1) = \sum_{a=x_1}^{n_1} P_{\pi_0}(X_1=a|n_1)$$

It is easy to show that the p -value for sample points in set S is always greater than that in set G . A sample point with a smaller p -value indicates stronger evidence against the null hypothesis. Therefore, it should be associated with a larger lower limit for the response rate.

The new ordering is the one sorted by p -value from the largest to the smallest, where the sample point with the smallest p -value has the largest lower limit. It should be noted that other p -values can also be used for the stochastic ordering, such as the p -value based on the UMUVE. However, the ordering based on a parameter estimate and the ordering based on the p -value by using the parameter estimate, are the same. Therefore, it is not necessary to repeat the ordering based on the p -value by using π_n or π_u .

Let $\Gamma_\lambda(X_1, X_2)$ denote the position for the sample point (X_1, X_2) in the sample space when the λ approach is used, $\lambda = \text{KC, R, and P}$. It should be noted that sample points having the same value from the sorting method should have the same position number. For simplicity, the sample space is sorted by the lower limit from the smallest to the largest. Then, the one-side interval $(L_\lambda(x_1, x_2), 1]$ based on the λ approach is the collection of π

$$\{\pi: P(\Omega_\lambda(x_1, x_2) | \pi) > \alpha\} \quad (2)$$

where $\Omega_\lambda(x_1, x_2) = \{(X_1, X_2) : \Gamma_\lambda(X_1, X_2) \leq \Gamma_\lambda(x_1, x_2)\}$. It should be noted that the KC approach can only compute the confidence interval for part of the sample space.

Coverage probability is defined as

$$P_\pi(\pi \in (L(X_1, X_2), 1]) = P((X_1, X_2) : L(X_1, X_2) < \pi | \pi) \quad (3)$$

A confidence interval is called exact when $P_\pi(\pi \in (L(X_1, X_2), 1]) = 1 - \alpha$ is true for any $\pi \in [0, 1]$. Figure 1 displays the coverage probability for the three approaches in an over-enrollment study with the second stage sample size $n'_2=20$ instead of $n_2 = 12$ from Simon's design that is $(n_1, n_2, r_1, r_2) = (10, 12, 2, 7)$ for the given design parameter $(\pi_0, \pi_1, \alpha, \beta) = (0.2, 0.5, 0.05, 0.1)$. When the KC approach is used, sample points that meet the following conditions ((1): $X_1 > r_1$, (2): $X'_2=0$, and (3): $X_1 \leq r_1$) do not have confidence intervals. As can be seen from the plots, the KC approach does not guarantee the coverage probability, with the majority of points having coverage probabilities below the nominal level, while the R approach and the P approach guarantee the coverage probability.

3 Numerical studies

We compare the performance of the KC approach, the R approach, and the proposed P approach with regard to simple AL and EL). These approaches are compared by using Simon's optimal design with under- or over-enrollment in the second stage. For each Simon's design with sample sizes n_1 and n_2 given study design parameter $(\pi_0, \pi_1, \alpha, \beta)$, seven different second stage sample sizes $n'_2 = \theta n_2$ are studied, where $\theta = 0.5, 0.7, 0.9, 1, 1.1, 1.3$, and 1.5 . When $\theta=1$, the studied design is Simon's design, and this design is used as reference. In the design parameters, α is set as 0.05 , and β equals either 0.2 or 0.1 representing 80% or 90% power. Eight response rates under null hypothesis are investigated:

$\pi_0 = 0.05, 0.1, 0.2, \dots, 0.7$, and $\pi_1 = \pi_0 + 0.2$. A total of $8 \times 2 \times 7 = 112$ configurations are studied. As aforementioned, the R approach and the P approach compute the exact lower limits for every sample point in the sample space, while the KC approach can only calculate the exact limit for part of the sample space. For this reason, we first compare the performance of the R and P approach, and then the better of the two will be further compared with the KC approach.

3.1 Comparison between the R approach and the P approach

The R approach and the P approach allow computation of the confidence limits for all sample points in the sample space. By their definition of confidence limits, the lower limits for sample points in set S are the same for both approaches. Therefore, these points are not included in the length calculation. Then, the simple AL of sample points in set G is defined as

$$AL = \frac{\sum_{(X_1, X'_2) \in G} [1 - L(X_1, X'_2)]}{N_G}$$

where N_G is the size of set G . The ALs for the one-sided lower intervals based on the R approach and the P approach are computed for all 112 configurations, and some typical results are presented in Table 1: seven cases when $(\pi_0 = 0.1, \pi_1 = 0.3, \beta = 0.1)$, and another seven cases when $(\pi_0 = 0.6, \pi_1 = 0.8, \beta = 0.2)$. As can be seen from the table, the P approach has a shorter AL than the R approach for all these 14 cases. Generally speaking, an interval with a short length is preferable. In addition to these special cases, we present the AL comparison between the two approaches for all 112 cases in Figure 2. The x-axis value is the AL of the R approach, and y-axis is the AL of the P approach. Each dot in the figure represents the ALs based on the R approach and the P approach for one case. When the dot is below the diagonal line, the P approach performs better than the R approach for that case. It is observed that all the dots are below the line, indicating that the P approach dominates the R approach with regard to the AL criteria.

All sample points in the AL calculation are assigned with equal weights. Alternatively, each sample point's length can be weighted by its probability. This is the second measurement to compare the performance of different intervals. It is called the EL, which is defined as

$$EL(\pi) = \sum_{(X_1, X'_2) \in G} [1 - L(X_1, X'_2)] b(X_1, n_1, \pi) b(X'_2, n_2, \pi)$$

where $b(X_1, n_1, \pi) b(X'_2, n_2, \pi)$ is the probability density function for the sample point (X_1, X'_2) , and π ranges from 0 to 1. Traditionally, the ELs between the two intervals can be compared by using the difference $EL_1(\pi) = EL_R(\pi) - EL_P(\pi)$, or the ratio $EL_2(\pi) = EL_R(\pi)/EL_P(\pi)$. When the difference is not that big for some π values, the ratio is a better

measurement to capture the relationship between the two intervals. We choose 2000 uniformly distributed π values from 0 to 1, and compute all the ratios $EL_2(\pi)$. Figure 3 displays some typical results for the comparison between the R approach and the P approach with regard to the EL by using the ratio as a function of π for the cases: $n_2' = 0.5n_2, 0.9n_2, 1.1n_2$, and $1.5n_2$. The original second stage sample size n_2 is computed for the following three scenarios: $(\pi_0, \pi_1, \beta) = (0.2, 0.4, 0.1), (0.4, 0.6, 0.1)$, and $(0.7, 0.9, 0.2)$. It can be seen that the performance of the R approach increases as π_0 increases and n_2' increases. However, in all these cases, the ratio is generally below 1, indicating that the P approach has a shorter EL than the R approach. Therefore, the P approach performs better than the R approach with regard to the EL criteria. For this reason, we will only use the P approach in the following comparison with the KC approach.

3.2 Comparison between the KC approach and the P approach

As mentioned in Section 2, the KC approach can only calculate the one-sided lower limit for sample points in set G which meet the following two constraints: $X_1 \geq r_t$ and $X_2' > 0$. This sub sample space is referred to as the Ω_{KC} . Suppose N_{KC} is the size of this sample space. Then, the AL is calculated as

$$AL = \frac{\sum_{(X_1, X_2') \in \Omega_{KC}} [1 - L(X_1, X_2')]}{N_{KC}}$$

Table 2 shows the ALs of the KC approach and the P approach for the 95% one-sided lower interval when $\pi_0 = 0.2$ and $\pi_1 = 0.4$ to attain 80% or 90% power. The P approach dominates the KC approach with a shorter AL for various over- and end-enrollment cases. Similar results are observed for other cases. This comparison may not be fair since the P approach guarantees the coverage probability while the KC approach does not. As seen in Figure 1, the KC approach is generally anti-conservative for the case, where the actual coverage probability is less than the nominal coverage level. To further compare the two intervals, we draw all the lengths for each sample point in set Ω_{KC} , $[1 - L(X_1, X_2')]$, in Figure 4, for the second case in Table 2 with $n_1 = 13$ and $n_2' = 21$. A total of 189 sample points are in the set Ω_{KC} . It can be seen from the figure that the two approaches have similar interval length for the majority of the sample points. The interval length of the P approach could be slightly longer than that of the R approach for a few sample points. But, the interval length of the R approach is significantly longer than that of the P approach for some sample points. These sample points often have a large X_1 value and a small X_2' value. In such cases, the lower limit of the KC approach is much less than that of the P approach. We also observe this trend from other configurations.

We also compare the performance between the KC approach and P approach with regard to the EL, which is calculated by using the sample space Ω_{KC} to make sure that the lower interval can be obtained by using the KC approach. The design configurations in Figure 3

are considered for the comparison between the KC approach and the P approach in Figure 5. Although the performance of the KC approach increases as n'_2 and π_0 increase, the P approach always performs better than the KC approach with a very high proportion of π values such that the P approach has a shorter EL when compared to the KC approach. The P approach outperforms the KC approach with regard to the AL criteria, the EL criteria, and the exactness of the interval.

3.3 Example

We illustrate the application the proposed P approach, the R approach, and the KC approach by using a Phase II clinical trial.⁹ The primary endpoint of this clinical trial is the response rate from patients with relapsed or refractory aggressive non-Hodgkin lymphoma (NHL) being treated with alisertib. Alisertib is a very highly selective small molecule inhibitor of Aurora A kinase.⁹ Complete response and partial response are considered as response in this study. The historical response rate is estimated as 5%, and the target response rate is at least 20%. Simon's optimal two-stage design is used. In order to achieve 90% power at the significance level of 0.05, Simon's two-stage design is calculated as

$$(n_1, n_2, r_1, r_t) = (21, 20, 1, 4)$$

At the end of the first stage, five responders were observed out of $n_1 = 21$ patients in that stage. Since the number of responses from the first stage $X_1 = 5$ is more than $r_1 = 1$, the trial goes to the second stage. At the end of the study, the total number of patients is 48 which is more than pre-planned sample size $n=41$. Out of the actual $n'_2=48-21=27$ patients in the second stage, eight responses were recorded. Friedberg et al.⁹ reported the confidence interval for the overall response rate without considering the nature of a two-stage design. The confidence interval based on the KC approach was available at that moment, but the KC approach cannot compute the confidence interval for the response rate when $X_1 > r_t$. The 95% exact one-sided lower limits based on the R approach and the P approach are 0.168 and 0.169, respectively. The lower limit based on the proposed P approach is slightly larger than that based on the R approach. Since π_0 is much less than the lower limits from these two approaches, the null hypothesis is rejected by using either of the two exact approaches.

4 Discussion

The hypothesis for these problems is often one-sided. Therefore, we propose an exact one-sided lower interval for the response rate to make proper and exact statistical inference. Meanwhile, the proposed exact approaches for lower limits can be easily extended to exact upper limits. From the numerical studies, we know that the commonly used approach by Koyama and Chen⁴ does not guarantee the coverage probability and can only compute a confidence interval for a limited number of sample points. The existing approach based on the number of responses, the R approach, is exact, but this approach is not as efficient as the proposed exact approach based on the p -value, the P approach. One reason is that the intervals based on the R approach have a substantial number of ties. The proposed P approach outperforms both the R approach and the KC approach with regard to AL and EL.

Therefore, it is recommended for use in practice. We are going to develop an R package based on the approaches developed in this article for public use.

To further improve the proposed P approach, one may consider the inductive order developed by Wang,¹⁰ who created a new ordering of the sample size based on the distance between the sample points. This approach has been successfully applied to many important statistical problems, the difference between two proportions^{10–13} and the relative risk and odds ratio.^{14,15} They all belong to a one-stage setting. We will consider extending his approach to a two-stage setting in the collaboration with Prof. Wang in the future.

Acknowledgments

The author is very grateful to the Editor, the Associate Editor, and two reviewers for their insightful comments that helped improve the manuscript.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is supported by National Institutes of Health grants 1P20GM109025, P20GM103440, and 5U54GM104944.

References

1. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials*. 1989; 10:1–10. [PubMed: 2702835]
2. Jung SHH, Kim M. On the estimation of the binomial probability in multistage clinical trials. *Stat Med*. 2004; 23:881–896. [PubMed: 15027078]
3. Raphaël P, Kristell D. What inference for two-stage phase II trials? *BMC Med Res Methodol*. 2012; 12:117. [PubMed: 22867439]
4. Koyama T, Chen H. Proper inference from Simon's two-stage designs. *Stat Med*. 2008; 27:3145–3154. [PubMed: 17960777]
5. Zhao J, Yu M, Xi-Ping PF. Statistical inference for extended or shortened phase II studies based on Simon's two-stage designs. *BMC Med Res Methodol*. 2015; 15:48. [PubMed: 26048655]
6. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. 1934; 26:404–413.
7. Wei-Yann YT, Yunchan C, Chia-Min MC. Interval estimation of binomial proportion in clinical trials with a two-stage design. *Stat Med*. 2008; 27:15–35. [PubMed: 17566141]
8. Donglin Z, Fei G, Kuolung H, et al. Hypothesis testing for two-stage designs with over or under enrollment. *Stat Med*. 2015; 34:2417–2426. [PubMed: 25809924]
9. Friedberg JW, Mahadevan D, Cebula E, et al. Bernstein. Phase II study of alisertib, a selective Aurora A kinase inhibitor, in relapsed and refractory aggressive B- and T-cell non-Hodgkin lymphomas. *J Clin Oncol*. 2014; 32:44–50. [PubMed: 24043741]
10. Wang W. On construction of the smallest one-sided confidence interval for the difference of two proportions. *Ann Stat*. 2010; 38:1227–1243.
11. Guogen S, Changxing M. Unconditional tests for comparing two ordered multinomials. *Stat Meth Med Res*. 2016; 25:241–254.
12. Guogen S, Weizhen W. ExactCIDiff: an R package for computing exact confidence intervals for the difference of two proportions. *R J*. 2013; 5:62–71.
13. Guogen S, Gregory EW, Alan DH, et al. Optimal adaptive two-stage designs for early phase II clinical trials. *Stat Med*. 2016; 35:1257–1266. [PubMed: 26526165]
14. Weizhen W, Guogen S. Exact confidence intervals for the relative risk and the odds ratio. *Biometrics*. 2015; 71:985–995. [PubMed: 26228945]

15. Weizhen W, Zhongzhan Z. Asymptotic infimum coverage probability for interval estimation of proportions. *Metrika*. 2014; 77:635–646.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.

Coverage probability comparison among the three approaches for the one-sided lower 95% limits for an over-enrollment two-stage design $(r_1, r_t, n_1, n_2') = (2, 7, 10, 20)$.

Figure 2.

AL comparison between the R approach and the P approach for all the 112 over- or under-enrollment studies. When a dot is below the diagonal line, the P approach has a shorter average length than the R approach for that configuration.

Figure 3.

Expected length ratio between the P approach and the R approach, EL_P/EL_R for over- or under-enrollment two-stage designs at $\alpha = 0.05$. The P approach has a shorter expected length than the R approach when the ratio is less than 1.

Figure 4.

Interval length comparison between the KC approach and the P approach for a study with sample sizes $n_1 = 13$ and $n'_2 = 21$ that has a total of 189 sample points. When a dot is below the diagonal line, the P approach has a shorter length than the KC approach for that sample point.

Figure 5.

Expected length ratio between the P approach and the KC approach, EL_P/EL_{KC} for over- or under-enrollment two-stage designs at $\alpha = 0.05$. The P approach has a shorter expected length than the KC approach when the ratio is less than 1.

Table 1

Average length comparison between the R approach and the P approach for over- or under-enrollment two-stage designs at $\alpha = 0.05$.

n_1	n_2'	AL_R	AL_P
$\pi_0 = 0.1, \pi_1 = 0.3, \text{ and } \beta = 0.1$			
18	8	0.605	0.577
18	12	0.601	0.571
18	15	0.598	0.567
18	17 ^a	0.596	0.564
18	19	0.594	0.562
18	22	0.591	0.559
18	26	0.588	0.556
$\pi_0 = 0.6, \pi_1 = 0.8, \text{ and } \beta = 0.2$			
11	16	0.445	0.379
11	22	0.442	0.378
11	29	0.439	0.375
11	32 ^a	0.438	0.375
11	35	0.437	0.374
11	42	0.435	0.372
11	48	0.434	0.372

^aSimon's design $n_2' = n_2$.

Table 2

Average length comparison between the P approach and the KC approach for over- or under-enrollment in the second stage for Simon's optimal two-stage designs when $\pi_0 = 0.2$ and $\pi_1 = 0.4$ at $\alpha = 0.05$.

n_1	n_2'	AL_{KC}	AL_P
80% power			
13	15	0.589	0.556
13	21	0.582	0.548
13	27	0.577	0.542
13	30 ^a	0.574	0.539
13	33	0.573	0.537
13	39	0.569	0.533
13	45	0.566	0.530
90% power			
19	18	0.618	0.589
19	24	0.612	0.578
19	32	0.607	0.568
19	35 ^a	0.605	0.564
19	38	0.604	0.561
19	46	0.600	0.554
19	52	0.598	0.550

^aSimon's design $n_2' = n_2$.