ClinSimon

Yunhe Liu

Contents

Introduction 1
Overview of Simon's Two-Stage Design 2
Adaptive Threshold Simon Design (ATS Simon) 3
Adaptive Threshold and Sample Size Simon Design (ATSS Simon) 6
Post-Trial Inference for ATS and ATSS Simon Designs with Under- and Over-Enrollment 8
References 11

Introduction

The `ClinSimon` package serves three purposes, including:

- Providing an Adaptive Threshold Simon Design (ATS Simon) method for Simon's two-stage design in oncology trials when the realized sample sizes in the 1^{st} and/or 2^{nd} stage(s) are different from the planned sample sizes in the 1^{st} and/or 2^{nd} stage(s). The proposed ATS Simon design tries to follow sample sizes of the original design, to that end, this design updates the original thresholds of (r_1, r) in the 1^{st} and/or the 2^{nd} stage(s) to satisfy the type I error rate as the original planned design (note: power will decrease if the realized sample size is smaller than the original one).
- Providing an Adaptive Threshold and Sample Size Simon Design (ATSS Simon) method for Simon's two-stage design in oncology trials when the realized sample sizes in the 1^{st} and/or 2^{nd} stage(s) are different from the planned sample sizes in the 1^{st} and/or 2^{nd} stage(s). The proposed ATSS Simon method updates not only the original threshold of (r_1^*, r^*) but the original sample sizes of (n_1^*, n^*) to satisfy the type I error rate and power requirements as the original planned design (note: unlike the ATS Simon design, the sample size here will also be updated to satisfy the original power). In addition, the ATSS Simon design also satisfies the other criteria as in the originally planned design, such as minimizing the average sample size under the null hypothesis H_0.
- Providing comprehensive post-trial inference tools at the end of the trial for Simon's two-stage design when under- or over-enrollment occurs. This includes the computation of point estimates, confidence intervals, and p-values for the proposed ATS and ATSS Simon design methods.

This vignette introduces functionalities in {ClinSimon} tailored for
designing single-arm clinical trials using the proposed method. Examples featuring Simon's two-stage designs from {clinfun} demonstrate the application of these functions within our package.

To begin, install and load {ClinSimon} and {clinfun}:

```
library(clinfun)
library(ClinSimon)
```

Overview of Simon's Two-Stage Design

Simon's two-stage design is a statistical method used in clinical trials, particularly in Phase II studies, to evaluate the efficacy of a new treatment while minimizing the number of patients exposed to potentially ineffective treatments. It consists of two stages:

Stage 1:
-A predetermined number of patients (n_1) are enrolled and treated.
-If the number of responses (positive outcomes) in these patients meets or exceeds a certain threshold (r_1), the trial continues to the second stage.
-If the number of responses is equal to or below this threshold, the trial is stopped early for futility, concluding that the treatment is ineffective.

Stage 2:
-An additional number of patients (n_2) are enrolled and treated.
-The total number of responses from both stages is then evaluated against a final threshold (r).
-If the total number of responses exceeds the final threshold, the treatment is considered promising for further study.
-If not, the treatment is considered ineffective.

Simon's two-stage design aims to reduce the number of patients receiving ineffective treatment while ensuring that promising treatments are identified efficiently. It balances the need for early stopping in case of futility with the need for sufficient data to make a reliable decision about the treatment's efficacy (Simon, 1989).

From the above description, when designing a trial using Simon's two-stage design, there are four key design parameters that need to be specified:

- r_1: Threshold in stage 1
- r: Threshold in stage 2
- n_1: Number of patients in stage 1
- n: Number of patients in stages 1 and 2, e.g., $(n_1 + n_2)$

Design Parameters and Constraints

In order to determine the design parameters for Simon's two-stage design, the followings are required:

- p_0: Unacceptable efficacy rate
- p_1: Desirable efficacy rate
- α: Type I error rate
- β: Type II error rate

That is,

- H_0: The true treatment response rate is less than or equal to some unacceptable level ($p \leq p_0$)
- H_1: The true treatment response rate is greater than or equal to some desirable level ($p \geq p_1$)

Example 1 (Designing a Study Using Simon's Two-Stage Design)

One trial employed Simon's two-stage design, considering an overall response rate of 25% as unacceptable and 45% as desirable. The trial set a type I error rate at 10% and aimed for a target power of 90%, corresponding to a 10% type II error rate.

From the above, we have:

- p_0: 25% overall response rate
- p_1: 45% overall response rate
- α: 10% type I error rate
- β: 10% type II error rate

Thus, the hypotheses used for testing are:

- H_0: The true overall response rate is less than or equal to 25% ($p \leq 0.25$)
- H_1: The true overall response rate is greater than or equal to 45% ($p \geq 0.45$)

Now, using `ph2simon()` function from clinfun R package, specify these parameters and print the resulting object.

```
library(clinfun)
# Specify the parameters and constraints
trial = ph2simon(0.25, 0.45, 0.1, 0.1)
# Print
trial
#>
#>  Simon 2-stage Phase II design
#>
#> Unacceptable response rate:  0.25
#> Desirable response rate:  0.45
#> Error rates: alpha =  0.1 ; beta =  0.1
#>
#>            r1 n1  r  n EN(p0) PET(p0)   qLo   qHi
#> Minimax     5 23 13 39  31.50  0.4685 0.752 1.000
#> Admissible  3 15 13 40  28.47  0.4613 0.026 0.752
#> Optimal     3 14 14 44  28.36  0.5213 0.000 0.026
```

This output provides the computed design parameters (r_1, r, n_1, n) as well as $EN(p_0)$ and $PET(p_0)$ for three designs, optimal, minimax, and admissible designs.

As expected, the optimal design has the smallest expected sample size under the Null ($EN(p_0) = 28.36$), whereas the minimax design has the smallest maximum sample size ($n = 39$).

Adaptive Threshold Simon Design (ATS Simon)

In single arm phase II studies, under-enrollment or over-enrollment can be common issues when using Simon's two-stage design. For rare diseases, it can be particularly challenging to recruit additional patients quickly in cases of under-enrollment. Therefore, we propose an Adaptive Threshold Simon design to assist clinical

investigators in making decisions based on the actual sample size instead of the planned one while still adhering to the original design framework when under- or over-enrollment occurs.

A thorough literature review identified numerous clinical trials utilizing Simon's two-stage designs, where the proportion of patients classified as inevaluable for response surpassed thresholds of 20%, 30%, and even 40%. The factors contributing to inevaluability included patients failing to complete the requisite number of cycles for response evaluation, being deemed ineligible upon central review, early withdrawal due to noncompliance, among other considerations (Ji, 2022), leading to under-enrollment during at Go/NoGo decision-making timepoint in studies. Over-enrollment in multi-site trials often occurs due to variations in patient recruitment rates across different locations.

It is important to note that deviations our methods address, such as under-enrollment and over-enrollment, are *incidental* or *non-informative*, indicating they arise from unforeseen circumstances or factors rather than deliberate bias or systematic error.

We identify an updated 1^{st} stage threshold r_1^* based on the actual 1^{st} stage sample size n_1^* such that the updated design's probability of early termination (PET) under the null hypothesis is the closest to the original PET under the null hypothesis.

$$P\left(Y_1 \leq r_1^* \mid n_1^*, p_0\right) \approx P\left(Y_1 \leq r_1 \mid n_1, p_0\right)$$

The primary objective of Simon's two-stage design, as well as our proposed ATS or ATSS Simon design, in single arm phase II studies, is to identify and eliminate ineffective drugs as early as possible, Therefore, in cases of deviations in sample size, the updated type I error rate (α) of a new method (including the operating characteristic, like probability of early termination) should still be close to the original one.

In the field of group sequential design, the alpha-spending function is a natural way to reallocate the type I error rate (α). We have adopted this method in our proposed ATS Simon's design.

Based on the identified r_1^*, we then define the alpha-spending function based on the actual total sample size (n^*). Specifically, we use the Lan-DeMets spending function (Lan-DeMets et al., 1983).

$$\alpha(n^*) = \begin{cases} 2 - 2\Phi\left(\frac{z_{1-\alpha/2}}{(n^*/n)^{1/2}}\right) & \text{if } n^* \leq n \\ \alpha & \text{if } n^* > n \end{cases}$$

where α here is the type I error rate in the original Simon two-stage design.

We can see, if the actual sample size $n^* > n$, the updated alpha $\alpha(n^*)$ will be at most α (the original one) and if the actual total sample size $n^* < n$, the $\alpha(n^*)$ may be smaller than α due to smaller sample size. Therefore, ATS Simon's design can control the type I error rate at or below the original level. Based on the above information, we identify a smallest integer r^* as the threshold at 2^{nd} stage such that

$$P\left(Y_1 > r_1^*, Y > r^* \mid n_1^*, n^*, p_0\right) \approx \alpha(n^*)$$

Here, n_1^* denotes as the actual sample size at the 1^{st} stage; n^* denotes as the actual sample size of the two stages.

Rationale of identifying a smallest integer r^* is that probability $P\left(Y_1 > r_1^*, Y > r^* \mid n_1^*, n^*, p\right)$ is a decreasing function of the threshold r^* in stage 2. So, if we find such a r^* under p_0 and based on the fact of $\alpha(n^*) \leq \alpha$, the type I error rate is well controlled. Since the trends of alpha and power are the same (i.e., if alpha increases, power also increases), if alpha is close to the original value, the power will also be close to the original value. This means we can simultaneously maintain the original power or minimize power loss. That is, if we select a larger r^* instead of the smallest one, we cannot minimize the power loss.

Example 2 (ATS Simon design for Addressing Under-or Over-Enrollment in Simon's Two-Stage Design based on Alpha Spending Function)

Example 2.1 (Under-enrollment in 1^{st} stage) Firstly, let's introduce the usage of `ATS_Design()`. Under-enrollment in the first stage is more critical based on our understanding, so we will focus on this

scenario for illustration purposes. In the optimal Simon's two-stage design described in Example 1, the planned sample size for the 1^{st} and 2^{nd} stages are 14 and 30, respectively.

Suppose we currently have outcome data for only 11 patients in the 1^{st} stage and assume the 2^{nd} stage sample size for evaluable patients remains as planned, We can use `ATS_Design()` to provide updated thresholds for the interim analysis, without needing to wait for the number of evaluable patients to reach 14. This means the input ***n1_star*** is 11 and ***n_star*** is 41 (=11+30 instead of the original 14+30 as the total sample size), as the original optimal design specifies a sample size of 30 for the second stage.

```
ATS_Design(n1=14,n=44,n1_star=11,n_star=41,r1=3,r=14,p0=0.25,p1=0.45,alpha=0.1)
#>                                  r1* r* n1* n* alpha(n*) Type I Power EN(p0)
#> Adaptive Threshold Simon Design   2 14  11 41     0.088   0.06 0.854 27.344
#>                                  PET(p0)
#> Adaptive Threshold Simon Design   0.455
```

The updated design parameters, (r_1^*, r^*), by ATS Simon design method are (2, 14). We also output (n_1^*, n^*) and they are just actual sample sizes of stage 1 and the total sample size. The type I error rate of this updated design is 0.06, which is below the original type I error constraint of 0.1. As we know, if alpha decreases, power will also decrease. Therefore, the current power of 85.4% is lower than the original design's 90% as expected, but still close to the original power. Additionally, the probability of early termination is 0.455, which is close to the original design's 0.521.

Now, suppose the number of patients responding to the new treatment in the first stage is > 2. In that case, we will make a "Go" decision and recruit an additional 30 patients for the second stage of the study.

Example 2.2 (Under-enrollment in both 1^{st} and 2^{nd} stages) We now consider an under-enrollment scenario in the 2^{nd} stage, in addition to the 1^{st} stage's under-enrollment.

For example, if the actual sample size in the 2^{nd} stage is 28 (the planned one is 30), indicating under-enrollment. Now the total sample size is 39, e.g., the input ***n_satr*** $= 39$.

Using the updated design parameters with sample sizes at stages 1 and 2 of $(n_1^*, n^*) = (11, 39)$, and the other original design parameters in the following code, we can determine the updated design parameters and operating characteristics.

```
ATS_Design(n1=14,n=44,n1_star=11,n_star=39,r1=3,r=14,p0=0.25,p1=0.45,alpha=0.1)
#>                                  r1* r* n1* n* alpha(n*) Type I Power EN(p0)
#> Adaptive Threshold Simon Design   2 13  11 39     0.081  0.077 0.864 26.254
#>                                  PET(p0)
#> Adaptive Threshold Simon Design   0.455
```

Our updated design parameters (r_1^*, r^*) are now (2, 13). We also output (n_1^*, n^*) and they are just actual sample sizes of stage 1 and the total sample size. The updated type I error rate is 0.077, which is below the constraint of 0.1.

Example 2.3 (Under-enrollment in 1^{st} and Over-enrollment in 2^{nd} stage) As another example, suppose the actual sample size in the 2^{nd} stage is now 31, indicating over-enrollment. Therefore, the input ***n_satr*** would be 42.

```
ATS_Design(n1=14,n=44,n1_star=11,n_star=42,r1=3,r=14,p0=0.25,p1=0.45,alpha=0.1)
#>                                  r1* r* n1* n* alpha(n*) Type I Power EN(p0)
#> Adaptive Threshold Simon Design   2 14  11 42     0.092  0.071 0.872 27.889
#>                                  PET(p0)
#> Adaptive Threshold Simon Design   0.455
```

The updated design parameters (r_1^*, r^*) are now (2, 14). We also output (n_1^*, n^*) and they are just actual sample sizes of stage 1 and the total sample size.The type I error for our ATS Simon design is 0.071, which is below the constraint of 0.1.

Adaptive Threshold and Sample Size Simon Design (ATSS Simon)

Rather than offering the above ATS design by merely adjusting thresholds based on the realized sample size, we present another option that follows the original Simon's two-stage design algorithm: the Adaptive Threshold and Sample Size Simon (ATSS Simon) method. This approach extends Simon's two-stage design by simultaneously adjusting both thresholds and sample size.

The overall strategy of this method is detailed as below:

• Scenario 1: When under-enrollment or over-enrollment occurs at the 1^{st} stage, we identify the design parameters (r_1^*, r^*, n^*) based on the actual sample size n_1^* at the 1^{st} to satisfy the significance level α and power $1-\beta$:

$$P\left(Y_1 > r_1^*, Y > r^* \mid n_1^*, n^*, p_0\right) \leq \alpha$$
$$P\left(Y_1 > r_1^*, Y > r^* \mid n_1^*, n^*, p_1\right) \geq 1-\beta$$

In addition, the design parameters (r_1^*, r^*, n^*) satisfies the same criteria as in the Optimal Simon's Two Stage: minimizing the average sample size under the null hypothesis.

• Scenario 2 (n^* changes again): When a Go decision has been made, the realized sample size n^{**} may be again different from n^*. Further adjustment of the threshold at the 2^{nd} stage is needed. So, we update again this threshold r^* such that we identify a smallest integer r^{**} given the design parameters (r_1^*, n_1^*, n^{**}) satisfying

$$P\left(Y_1 > r_1^*, Y > r^{**} \mid n_1^*, n^{**}, p_0\right) \leq \alpha$$

Here, n_1^* denotes as the actual sample size at the 1^{st} stage; n^* denotes as the actual total sample size of the two stages after the interim analysis based on n_1^* n^{**} denotes as the actual total sample size of the two stages if n^* is again updated.

Example 3 (Adaptive Threshold and Sample Size Simon Design (ATSS Simon) for Under- and/or Over-enrollment)

In the optimal Simon's two-stage design described in Example 1, the planned sample size for the 1^{st} and 2^{nd} stages are 14 and 30, respectively.

Example 3.1 (Under-enrollment at 1^{st} stage) Suppose we currently have outcome data for only 11 patients in the 1^{st} stage and assume the 2^{nd} stage sample size remains as planned. We can use `ATSS_Design_Stage1()` to implement the introduced ATSS Simon algorithm. Here, the parameter ***n1_satr*** is 11.

Note: Different from the previous ATS examples, we now should be noted that it is unnecessary to input ***n_star*** as a parameter, since the current ATSS method will compute an updated total sample size based on the sample size deviation of the stage 1 from the original design.

```
ATSS_Design_Stage1(p0=0.25,p1=0.45,n1_star=11,alpha=0.1,beta=0.1)
#>                    r1* r* n1* n* Type I Power EN(p0) PET(p0)
#> ATSS_Design_Stage1   2 15  11 47   0.09 0.901 30.613   0.455
```

The updated design parameters (r_1^*, r^*, n_1^*, n^*) are (2, 15, 11, 47). The type I error for our redesign is 0.09, which is controlled well and < 0.1. The updated power for our redesigned study is 0.901, surpassing the constraint of 0.9. This increase is due to the ATSS Simon design method, which now provides a larger total sample size of 47, compared to 44 in the original design.

Example 3.2 (Under-enrollment at 1^{st} and 2^{nd} stages) If more than 2 patients respond to the new treatment in the first stage of our redesigned study, we will proceed to recruit 36 additional patients (47 total minus the 11 from the 1^{st} stage). However, if now the actual sample size in the second stage is 34, indicating under-enrollment, we can use the function `ATSS_Design_Stage2()` to update the design parameters based on the realized total sample size. This adjustment means the parameter ***n_double_star*** is now 45.

Note: In the following code, ***n1_star*** and ***n_double_star*** are fixed since the trial is complete. Essentially, the ATSS algorithm only updates the threshold for stage 2, ***r_star***. In this example, however, ***r_star*** remains unchanged compared to Example 3.1.

```
ATSS_Design_Stage2(p0=0.25,p1=0.45,r1_star=2,n1_star=11,n_double_star=45,alpha=0.1)
#>                     r1* r* n1* n** Type I Power EN(p0) PET(p0)
#> ATSS_Design_Stage2   2 15  11  45  0.066 0.878 29.523   0.455
```

Our updated design parameters (r_1^*, r^*, n_1^*, n^*) are (2, 15, 11, 45). And the updated type I error rate is 0.066 smaller than the original one of 0.1 and the updated power is 0.878, both are due to the realized smaller total sample size, though in this example, the updated ***r_star*** is still 15.

Example 3.3 (Under-enrollment at 1^{st} and Over-enrollment at 2^{nd} stage) Suppose more than 2 patients respond to the new treatment at the end of the 1^{st} stage of our redesigned study. In that case, we will make a "Go" decision to recruit 36 more patients into our study. However, if the actual sample size in the 2^{nd} stage is 37, indicating over-enrollment, we can use the function `ATSS_Design_Stage2()` to update the design parameters based on the actual total sample size. In the following code, this means the parameter ***n_double_star*** is now 48.

```
ATSS_Design_Stage2(p0=0.25,p1=0.45,r1_star=2,n1_star=11,n_double_star=48,alpha=0.1)
#>                     r1* r* n1* n** Type I Power EN(p0) PET(p0)
#> ATSS_Design_Stage2   2 16  11  48  0.061 0.884 31.158   0.455
```

Our updated design parameters (r_1^*, r^*, n_1^*, n^*) are (2, 16, 11, 48). And updated type I error rate is 0.061 and updated power is 0.884.

Note: Compared to Example 3.2, although the current total sample size is larger (48 vs. 45), we expect to see an increase in power and type I error rate larger than those in Example 3.2. However, while the power is as expected, the type I error rate is not. This is because the optimally searched threshold at stage 2, ***r_star***, is 16, compared to 15 in Example 3.2. Our algorithm dictates controlling the type I error rate and could only identify 16 as the threshold.

```
## Our algorithm searching process
result <- data.frame(
  r_star = round(c(2.0000000, 3.0000000, 4.0000000, 5.0000000, 6.0000000,
                   7.0000000, 8.0000000, 9.0000000, 10.0000000, 11.0000000,
                   12.0000000, 13.0000000, 14.0000000, 15.0000000, 16.0000000)),
  alpha = round(c(0.5447991, 0.5447929, 0.5447130, 0.5442052, 0.5421068,
                  0.5357601, 0.5207790, 0.4920445, 0.4460005, 0.3831060,
                  0.3087428, 0.2317242, 0.1611787, 0.1035884, 0.0614173), 3),
  power = round(c(0.9347765, 0.9347765, 0.9347765, 0.9347765, 0.9347763,
                  0.9347752, 0.9347684, 0.9347366, 0.9346115, 0.9341919,
                  0.9329744, 0.9298792, 0.9229204, 0.9089765, 0.8839142), 3)
)
print(result)
#>    r_star alpha power
#> 1       2 0.545 0.935
```

```
#> 2       3 0.545 0.935
#> 3       4 0.545 0.935
#> 4       5 0.544 0.935
#> 5       6 0.542 0.935
#> 6       7 0.536 0.935
#> 7       8 0.521 0.935
#> 8       9 0.492 0.935
#> 9      10 0.446 0.935
#> 10     11 0.383 0.934
#> 11     12 0.309 0.933
#> 12     13 0.232 0.930
#> 13     14 0.161 0.923
#> 14     15 0.104 0.909
#> 15     16 0.061 0.884
```

Summary

Both the proposed ATS and ATSS Simon design methods can address under- and over-enrollment. The ATS algorithm updates the thresholds based on the actual sample sizes, maintaining control over the type I error rate α but potentially resulting in lower power than the original design during under-enrollment. In contrast, the ATSS algorithm updates both the thresholds and sample sizes, thereby controlling the α while maintaining power similar to the original design, though it may require a larger sample size.

Post-Trial Inference for ATS and ATSS Simon Designs with Under- and Over-Enrollment

Key design details of two-stage single-arm trials are frequently left unreported, and their statistical inference is often not conducted in a way that mitigates the bias introduced by interim analyses. This issue is particularly concerning given the increasing reliance on non-randomized trials, which now represent a significant portion of the evidence on treatment effectiveness in rare, biomarker-defined patient subgroups (Grayling, 2021). Motivated by this problem, we inherit some widely used post-trial inference methods for ATS and ATSS Simon Designs with Under- and Over-Enrollment.

Point Estimate

When a multistage trial is ended, we also want to estimate the true response probability π_p of the new therapy. The most commonly used estimator is the sample response rate, i.e. the maximum likelihood estimator (MLE):

$$\hat{\pi}_p = \frac{S}{N}$$

However, in multi-stage designs like Simon's Two-Stage design, we observe only extreme cases by crossing the threshold in the first stage, and hence the MLE is biased. In other words, the MLE is biased due to the sequential nature of the trial. This is known as the *optional sampling effect*. Here, we adopt Jung's method for the estimation of the binomial probability in multistage clinical trials (Jung, 2004). Based on the Rao-Blackwell theorem, they derived the uniformly minimum variance unbiased estimator (UMVUE) as the conditional expectation of an unbiased estimator, which in this case is simply the maximum likelihood estimator based only on the first stage data, given the sufficient statistic. Let M denote the stopping stage (2^{nd} stage in our context) and let $S = S_M$ denote the total number of responders accumulated up to the

stopping stage. For observation (m, s), the UMVUE of the response rate p is given by:

$$\hat{\pi}_p = \begin{cases} \frac{S}{n_1} & \text{if } m = 1 \\ \frac{\sum_{x_1=(r_1+1)\vee(S-n_2)}^{S\wedge n_1} \binom{n_1}{x_1}\binom{n_2-1}{S-x_1-1}}{\sum_{x_1=(r_1+1)\vee(S-n_2)}^{S\wedge n_1} \binom{n_1}{x_1}\binom{n_2}{S-x_1}} & \text{if } m = 2 \end{cases}$$

At stage m, we may accrue slightly more (or possibly less) patients than planned sample size as we introduced previously, especially in multicenter trials. UMVUE also provides an unbiased estimator for the conditional expectation by using all realized sample size.

Confidence intervals:

A conventional approach for constructing confidence intervals is to use the Clopper-Pearson exact confidence interval, disregarding the group sequential nature of the trial (Clopper et al., 1934).

$$P(Y \geq y \mid p_L) = \sum_{k=y}^{n} \binom{n}{k} p_L^k (1-p_L)^{n-k} = \frac{\alpha}{2}$$

$$P(Y \leq y \mid p_U) = \sum_{k=y}^{n} \binom{n}{k} p_U^k (1-p_U)^{n-k} = \frac{\alpha}{2}$$

We now focus on constructing confidence intervals by considering deviations in sample sizes from the planned ones in the 1^{st} and/or 2^{nd} stages. Jung (2004) proposed a method especially when the treatment successfully goes to the second stage. Let M denote the stage at which a trial is terminated, and S denote the number of responders at stage M. This method constructs confidence intervals based on the stochastic ordering of the distribution of (M, S) with respect to the response rate p.

$$P(\hat{p_u}(M,S) \geq \hat{p_u}(m,s) \mid p_L) = \frac{\alpha}{2}$$
$$P(\hat{p_u}(M,S) \leq \hat{p_u}(m,s) \mid p_U) = \frac{\alpha}{2}$$

However, the Clopper-Pearson confidence interval is known to be conservative, with the actual confidence level being bounded below by $(1-\alpha)$. Jung's method also inherits this conservatism (Porcher et al., 2012).

To correct for this conservative nature, Porcher extended the Jung's confidence interval with a mid-p approach (Porcher et al., 2012). This method is our recommended approach, even though all the above methods have also been integrated into the {ClinSimon} R package.

$$P(\hat{p_u}(M,S) > \hat{p_u}(m,s) \mid p_L) + \frac{1}{2}P(P(\hat{p_u}(M,S) = \hat{p_u}(m,s) \mid p_L)) = \frac{\alpha}{2}$$
$$P(\hat{p_u}(M,S) < \hat{p_u}(m,s) \mid p_U) + \frac{1}{2}P(P(\hat{p_u}(M,S) = \hat{p_u}(m,s) \mid p_U)) = \frac{\alpha}{2}$$

p-Value:

Kunzmann et al. (2024) raised a question whether a frequentist inferential framework for two-stage designs has a consistent test decision among p-values, point estimates, and confidence intervals. In practice, however, the consistency in those test decisions is often presumed by the non-statistical readership of published trial results and ambiguous situation can be avoided by using a consistent framework.

So here, we used the p-value defined as the probability of obtaining more extreme estimates toward H_1 than the observed one when H_0 is true based on the UMVUE ordering (Jung, 2006). Hence, for testing $H_0 : p = p_0$ against $H_0 : p = p_0 (p_0 < p_1)$, the p-value for an estimate $\hat{p}(m, s)$ will be given as

$$p_s = \begin{cases} 1 - \sum_{(i,j):\hat{p_u}(i,j)<\hat{p_u}(m,s)} f_{p_0}(i,j) & \text{if } m = 1 \\ \sum_{(i,j):\hat{p_u}(i,j)\geq\hat{p_u}(m,s)} f_{p_0}(i,j) & \text{if } m = 2 \end{cases}$$

It can be rewritten as

$$p_s = \begin{cases} P(Y_1 \geq s \mid p_0) & \text{if } m = 1 \\ \sum_{y_1=r_1+1}^{n_1} P(Y_1 = y_1 \mid p_0) P(Y_2 \leq s - y_1 \mid p_0) & \text{if } m = 2 \end{cases}$$

Example 4 (Post-Trial Inference usage for ATS and ATSS Simon Designs)

Example 4.1 (Post-Trial Inference for ATS Simon Design with Under-Enrollment) Suppose we use the ATS Simon's design method to address the problem of under-enrollment. The original Simon's Two-Stage design is $(r_1, r, n_1, n) = (3, 14, 14, 44)$. Considering the under-enrollment problem only at the 1^{st} stage like Example 2.1, the new design is $(r_1^*, r^*, n_1^*, n^*) = (2, 14, 11, 41)$. Especially, it should be noted that the input ***alpha*** here is the updated type I error constraint $\alpha(n^*)$ if we used the ATS Simon's design. In Example 2.1, the updated type I error constraint $\alpha(n^*)$ is 0.088. So it means the input '***alpha*** here is 0.088. If the new treatment successfully enters the second stage and 20 patients respond to it, this means the parameter ***m*** is 2 and ***s*** is 20 under this context.

We now compute the point estimate (UMVUE), confidence interval and p-value based on the above setting. Note here, the option ***"CP"*** means Clopper-Pearson exact confidence interval option ***"Jung"*** means Jung's confidence interval and option ***"MIDp"*** means mid-p approach confidence interval.

```
SimonAnalysis(m=2, s=20, n1=11, n2=30, r1=2, r=14, alpha=0.088, quantile=c(0.025,0.975),
              CI_option = "CP", p0=0.25)
#>                       UMVUE CI(lower) CI(upper) p_Val
#> Post-Trial Inference 0.494     0.347      0.63 0.001
```

```
SimonAnalysis(m=2, s=20, n1=11, n2=30, r1=2, r=14, alpha=0.088, quantile=c(0.025,0.975),
              CI_option = "Jung", p0=0.25)
#>                       UMVUE CI(lower) CI(upper) p_Val
#> Post-Trial Inference 0.494     0.329     0.629 0.001
```

```
SimonAnalysis(m=2, s=20, n1=11, n2=30, r1=2, r=14, alpha=0.088, quantile=c(0.025,0.975),
              CI_option = "MIDp", p0=0.25)
#>                       UMVUE CI(lower) CI(upper) p_Val
#> Post-Trial Inference 0.494     0.339     0.641 0.001
```

Note: the computed UMVUEs and p values are the same since the above introduced same UMVUE method and approach for computing p value applied to all the three ways of computing the CIs. And the input ***alpha*** here is the updated type I error constraint $\alpha(n^*)$ if we used the ATS Simon's design.

We can see the point estimate (UMVUE) for the response rate is 0.494; Clopper-Pearson exact confidence interval is (0.347, 0.630), Jung's confidence interval is (0.329, 0.629), mid-p approach confidence interval is (0.339, 0.641). Also, we see that p-value is 0.001 smaller than the updated type I error constraint $\alpha(n^*)$ of 0.088. So we can reject the null hypothesis H_0: The true treatment response rate is less than or equal to some unacceptable level ($p \leq p_0 = 0.25$).

Example 4.2 (Post-Trial Inference for ATSS Simon Design with Under-Enrollment) Suppose we use the ATSS Simon's design method to address the problem of under-enrollment. The original Simon's Two-Stage design is $(r_1, r, n_1, n) = (3, 14, 14, 44)$. Considering the under-enrollment problem only at the 1^{st} stage like Example 3.1, the new design is $(r_1^*, r^*, n_1^*, n^*) = (2, 15, 11, 47)$. If the new treatment successfully enters the second stage and 22 patients respond to it, this means the parameter ***m*** is 2 and ***s*** is 22 under this context.

We now compute the point estimate (UMVUE), confidence interval and p-value based on the above setting. Note here, the option *"CP"* means Clopper-Pearson exact confidence interval option *"Jung"* means Jung's confidence interval and option *"MIDp"* means mid-p approach confidence interval.

```
SimonAnalysis(m=2, s=22, n1=11, n2=36, r1=2, r=15, alpha=0.1, quantile=c(0.025,0.975),
              CI_option = "CP", p0=0.25)
#>                       UMVUE CI(lower) CI(upper) p_Val
#> Post-Trial Inference 0.478     0.342     0.597 0.001
```

```
SimonAnalysis(m=2, s=22, n1=11, n2=36, r1=2, r=15, alpha=0.1, quantile=c(0.025,0.975),
              CI_option = "Jung", p0=0.25)
#>                       UMVUE CI(lower) CI(upper) p_Val
#> Post-Trial Inference 0.478     0.322     0.604 0.001
```

```
SimonAnalysis(m=2, s=22, n1=11, n2=36, r1=2, r=15, alpha=0.1, quantile=c(0.025,0.975),
              CI_option = "MIDp", p0=0.25)
#>                       UMVUE CI(lower) CI(upper) p_Val
#> Post-Trial Inference 0.478      0.33     0.615 0.001
```

We can see the point estimate (UMVUE) for the response rate is 0.478; Clopper-Pearson exact confidence interval is (0.342, 0.597), Jung's confidence interval is (0.322, 0.604), mid-p approach confidence interval is (0.330, 0.615). Also, we see that p-value is 0.001 smaller than the type I error constraint α 0.01. So we can reject the null hypothesis H_0: The true treatment response rate is less than or equal to some unacceptable level ($p \leq p_0 = 0.25$).

References

Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. Controlled clinical trials, 10(1), 1-10. doi:10.1016/0197-2456(89)90015-9

Ji, L., Whangbo, J., Levine, J. E., & Alonzo, T. A. (2022). Inefficiency of two-stage designs in phase II oncology clinical trials with high proportion of inevaluable patients. Contemporary clinical trials, 120, 106849. doi:10.1016/j.cct.2022.106849

Gordon Lan, K. K., & DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. Biometrika, 70(3), 659-663. doi:10.1093/biomet/70.3.659

Grayling, M. J., & Mander, A. P. (2021). Two-stage single-arm trials are rarely analyzed effectively or reported adequately. JCO Precision Oncology, 5, 1813-1820. https://doi.org/10.1200/PO.21.00276

Jung, S. H., & Kim, K. M. (2004). On the estimation of the binomial probability in multistage clinical trials. Statistics in medicine, 23(6), 881-896. https://doi.org/10.1002/sim.1653

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika, 26(4), 404-413. https://doi.org/10.2307/2331986

Porcher, R., & Desseaux, K. (2012). What inference for two-stage phase II trials?. BMC medical research methodology, 12, 1-13. https://doi.org/10.1186/1471-2288-12-117

Kunzmann, K. (2024). Optimal Adaptive Designs for Early Phase II Trials in Clinical Oncology (Doctoral dissertation). https://archiv.ub.uni-heidelberg.de/volltextserver/34225/

Jung, S. H., Owzar, K., George, S. L., & Lee, T. (2006). P-value calculation for multistage phase II cancer clinical trials. Journal of Biopharmaceutical Statistics, 16(6), 765-775. https://doi.org/10.1080/10543400600825645