Two-Stage Designs

Examples and Background

Helmut Schütz

August 2017

Example 1. Data of Potvin *et al.*, Example 2.
'Method B': GMR 0.95, target power 0.80, α_1 0.0294, α_2 0.0294.
Stage 1 MSE 0.032634, ln(T)–ln(R) 0.08396, n_1 12.
Final MSE 0.045896, ln(T)–ln(R) 0.014439, N 20.

```
Data for the interim analysis
-----------------------------------------
CV (MSE)             :  18.21% (0.032634)
PE (ln(T)-ln(R))     : 108.76% (0.08396)
Sample size          :  12

Data for the final (pooled) analysis
-----------------------------------------
CV (MSE)             :  21.67% (0.045896)
PE (ln(T)-ln(R))     : 101.45% (0.014439)
Total sample size    :  20

Study conditions and assessment of empiric Type I Error
-------------------------------------------------------------
Design               : 2×2×2 crossover
TSD Type             : 1 (Potvin et al. 2008, Method B)
Target power         : 0.80
GMR used             : 0.95 (fixed)
Interim power check: yes
Futility criterion : none
Minimum n2           : not specified
Maximum N            : not specified
Specified α 1, 2     : 0.0294, 0.0294
Specified CIs        : 94.12%, 94.12%
TIE for specified α: 0.04307 (≤0.05)
                       Applied adjustment is justified.

Interim analysis (specified α1 0.0294)
-----------------------------------------------------
94.12% CI: 92.93–127.28% (failed to demonstrate BE)
Power    : 0.5049 (approx. via shifted central t)
Second stage with 8 subjects (N=20) is justified.

Power based on interim data (specified α)
-----------------------------------------------------
Method               : approx. via shifted central t
Stage 1              : 0.5248
Both stages          : 0.8560
Studies in stage 2 : 44.2%
Expected total sample size (N)
  Average            : 17.5
  Median             : 12
  5, 95 percentiles: 12, 34

Final analysis of pooled data (specified α2 0.0294)
=====================================================
94.12% CI: 88.45–116.38% (BE concluded)
Post hoc power (irrelevant; for validation purposes)
Based on GMR         : 0.6324
Based on PE          : 0.7363


 Since no inflation of the Type I Error is expected,
 can accept the reported analysis.
```

In 'Type 1' Two-Stage Design (TSDs) BE is assessed with the adjusted α in the interim first and then power. Since the study failed to demonstrate BE (line 35) and power is lower than the target 0.8 (line 36), the second stage can be initiated (line 37). Otherwise, the study should have stopped already in the interim. The code estimates the sample size of the second stage (based on the GMR, target power and α_2). Lines 39–48 give the result of simulating power (argument `pa=TRUE`). The average (total) sample size (called ASN by some authors) is 17.5. With the default setting (`pa=FALSE`) this part is not shown.

In the final analysis BE is demonstrated. *Post hoc* power is only given to compare the result with the reference (with the default setting this part is not shown). The assessment is given in the box.

Example 2. Data from above but 'Method C': α_0 0.05, $\alpha_1 = \alpha_2$ 0.0294.

```
Study conditions and assessment of empiric Type I Error
──────────────────────────────────────────────────────────────
Design              : 2×2×2 crossover
TSD Type            : 2 (Potvin et al. 2008, Method C)
Target power        : 0.80
GMR used            : 0.95 (fixed)
Interim power check: yes
Futility criterion : none
Minimum n2          : not specified
Maximum N           : not specified
Specified α 1, 2    : 0.050|0.0294, 0.0294
Specified CIs       : 90.00%|94.12%, 94.12%
TIE for specified α: 0.05062 (>0.05)
                      Applied adjustment is not justified.

Interim analysis (specified α1 0.0294)
──────────────────────────────────────────────────────
94.12% CI: 92.93–127.28% (failed to demonstrate BE)
Power    : 0.6494 (approx. via shifted central t)
Second stage with 8 subjects (N=20) is justified.

Power based on interim data (specified α)
──────────────────────────────────────────────────────
Method              : approx. via shifted central t
Stage 1             : 0.5449
Both stages         : 0.8635
Studies in stage 2 : 40.6%
Expected total sample size (N)
  Average           : 17.4
  Median            : 12
  5, 95 percentiles: 12, 34

Final analysis of pooled data (specified α2 0.0294)
══════════════════════════════════════════════════════
94.12% CI: 88.45–116.38% (BE concluded)
Post hoc power (irrelevant; for validation purposes)
Based on GMR        : 0.6324
Based on PE         : 0.7363

α-optimization (objective function: TIE - 0.05 → 0)
──────────────────────────────────────────────────────────
Method              : approx. via shifted central t
Convergence         : 18 iterations (run-time 5.15 min)
Estimated precision: 5.07E-09
Adjusted α 1, 2     : 0.050|0.02858, 0.02858
Adjusted CIs        : 90.00%|94.28%, 94.28%
TIE for adjusted α : 0.04992 (n.s. >0.05)

Interim analysis (adjusted α1 0.02858)
──────────────────────────────────────────────────────
94.28% CI: 92.82–127.44% (failed to demonstrate BE)
Power    : 0.6494 (approx. via shifted central t)
Second stage with 8 subjects (N=20) is justified.

Power based on interim data (adjusted α)
──────────────────────────────────────────────────────
Method              : approx. via shifted central t
Stage 1             : 0.5387
Both stages         : 0.8639
Studies in stage 2 : 41.2%
Expected total sample size (N)
  Average           : 17.5
  Median            : 12
  5, 95 percentiles: 12, 34

Final analysis of pooled data (adjusted α2 0.02858)
══════════════════════════════════════════════════════
94.28% CI: 88.36–116.49% (BE concluded)
Post hoc power (irrelevant; for validation purposes)
Based on GMR        : 0.6261
Based on PE         : 0.7305


 Since conclusions of both analyses agree,
 can accept the original analysis.
```

In 'Type 2' TSDs power in the interim is assessed first. If power is at least the target (here 0.8), this implies that the assumptions (CV, GMR) which lead to the sample size of the first stage seemingly are correct. According to the framework in this case no adjustment has to be done (BE can be assessed with α_0 0.05) since the study will stop in the interim (pass or fail). In the example power is less than the target (line 90) and therefore, BE must be assessed with α_1 0.0294. The study failed to demonstrate BE in the interim (line 89), and therefore, the second stage can be initiated (line 91).

Since an inflation of the TIE (0.05062) is expected, α is optimized (lines 111–118). With an α_2 of 0.02858 the TIE is controlled (0.04992). The interim with this α justifies a second stage as well (lines 120–124).

In the final analysis (lines 104–106) with the specified α 0.0294 BE is easily demonstrated (confidence limits far off the acceptance range). Repeating the final analysis with the adjusted α_2 0.02858 shows BE as well. The assessment is given in the box.

Example 3. Montague *et al.* Method D: GMR 0.90, target power 0.80, α_0 0.05, $\alpha_1 = \alpha_2$ 0.0280.
Stage 1 CV 20%, PE 0.92, n_1 12.
Final CV 23.315%, PE 0.88, N 45 (estimated 46; but one dropout in the second stage). Only part of the output is shown below.

```
Design              : 2×2×2 crossover
TSD Type            : 2 (Montague et al. 2011, Method D)
Target power        : 0.80
GMR used            : 0.90 (fixed)
Interim power check: yes
Futility criterion : none
Minimum n2          : not specified
Maximum N           : not specified
Specified α 1, 2    : 0.050|0.0280, 0.0280
Specified CIs       : 90.00%|94.40%, 94.40%
TIE for specified α: 0.05153 (>0.05)
                      Applied adjustment is not justified.

Interim analysis (specified α1 0.028)
──────────────────────────────────────────────────────
94.40% CI: 77.25–109.57% (failed to demonstrate BE)
Power     : 0.3407 (approx. via shifted central t)
Second stage with 34 subjects (N=46) is justified.

Final analysis of pooled data (specified α2 0.028)
══════════════════════════════════════════════════════
94.40% CI: 80.00–96.80% (BE concluded)

α-optimization (objective function: TIE - 0.05 → 0)
──────────────────────────────────────────────────────
Method              : approx. via shifted central t
Convergence         : 19 iterations (run-time 5.56 min)
Estimated precision: 5.18E-09
Adjusted α 1, 2     : 0.050|0.02709, 0.02709
Adjusted CIs        : 90.00%|94.58%, 94.58%
TIE for adjusted α : 0.04998 (n.s. >0.05)

Interim analysis (adjusted α1 0.02709)
──────────────────────────────────────────────────────
94.58% CI: 77.13–109.74% (failed to demonstrate BE)
Power     : 0.3407 (approx. via shifted central t)
Second stage with 34 subjects (N=46) is justified.

Final analysis of pooled data (adjusted α2 0.02709)
══════════════════════════════════════════════════════
94.58% CI: 79.94–96.88% (failed to demonstrate BE)
Post hoc power (irrelevant; for validation purposes)
Based on GMR        : 0.6623
Based on PE         : 0.4855
```

> Accepting the reported analysis could
> increase the relative consumer risk by ~3.1%.

This example represents one of the borderline cases; the reported lower confidence limit is at the acceptance range. CV 20% and n_1 12 is the location of the maximum inflation of the TIE of this method (both acc. to the authors' results and obtained by the `R`-package `Power2Stage`).

Conclusions in the final analyses do *not* agree (the study *passes* with the 94.40% CI but *fails* with the 94.58% CI). See lines 185, 204 and the assessment given in the box.

Note that this represents also an example where already rounding of the CI according to the BE guideline slightly inflates the TIE (even in fixed sample designs). The 94.40% CI is actually 79.99842–96.80191%. The study passes only due to rounding the lower CL up.

In the references the power approximation by the shifted central *t*-distribution was used for speed reasons (tenths of millions of studies had to be simulated). In actual studies likely the approximation by the noncentral *t*-distribution or even the exact method (Owen's Q) will be used. Both algorithms are available in commercial (`SAS`, `NQuery`) and open-source (`R` package `PowerTOST`) software. The former is also implemented in `PASS`. In the following a comparison of results:

example	TIE by method			adjusted α by			agreement†		
	a	b	c	a	b	c	a	b	c
1	0.04307	0.04269	0.04287	not necessary					
2	0.05062	0.05083	0.05087	0.02858	0.02856	0.02856	yes	yes	yes
3	0.05153	0.05180	0.05180	0.02709	0.02704	0.02704	no	no	no

† agreement in conclusions: analyses by optimized α (if necessary) *vs.* the pre-specified α.

a shifted central *t*-approximation

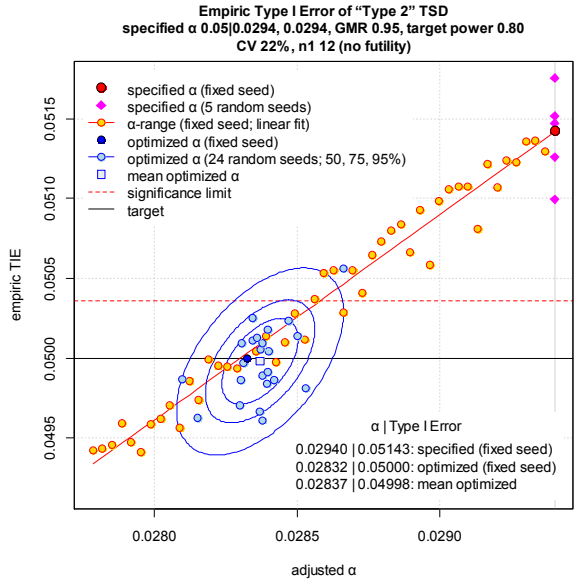
b noncentral *t*-approximation

c exact

Note: If an inflated TIE is detected with the exact method, adjusting α can take some [*sic*] hours.

It should be noted that in simulations of the references always *exactly* the re-estimated stage 2 sample size n_2 was used. Naturally, if in a study more subjects are dosed in the second stage (based on an assumed dropout-rate) and at the end of the study more than n_2 subjects are eligible, the chance to demonstrate BE increases and thus, potentially the TIE. Therefore, especially in such cases assessing the TIE is recommended.

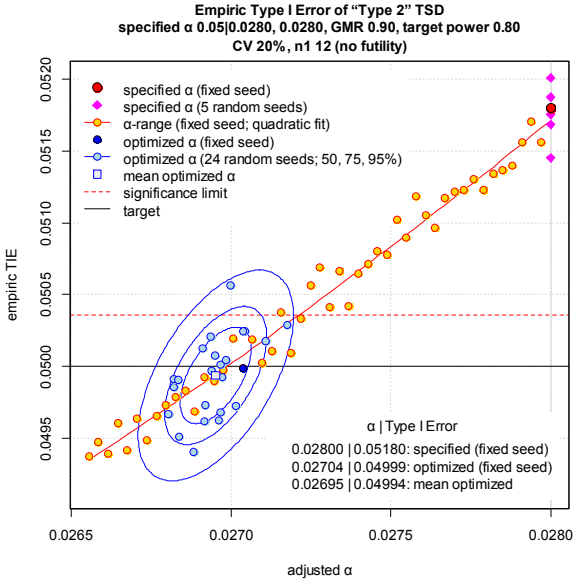
Can we expect that the Type I Error after optimization will *always* be ≤0.05 in simulations? Only if we use a *fixed* seed of the (pseudo) random generator – which is generally recommended in simulations for reproducibility. Let us explore the location of the maximum TIE for Potvin's 'Method C' (CV 22%, n_1 12) with power and sample size estimation by the noncentral *t*-approximation.



With the specified α 0.0294 we obtain an empiric TIE of 0.05143 (red circle). If we repeat the estimation with *random* seeds we get the magenta rhombi. If we assess lower alphas (*i.e.*, would adjust more), naturally the TIE decreases (yellow circles). With a fixed seed (blue circle) we get an optimized α of 0.02832 (TIE 0.05000) which is far below the significance limit for one million simulations (0.05036, binomial test) and hopefully accepted by the EMA's Biostatistics Working Party.

Now we repeat the optimization with random seeds and get the cluster of lightblue circles. The blue lines give the 50, 75, and 95 percentile ellipses (based on the bivariate normal distribution). Of course, one could use their mean (the square) as the 'best' estimate (here 0.02832), but would that really help? First, it is not reproducible any more (for every run of the code one will get another value) and second there is no guarantee that the TIE will be always ≤0.05. More about it in the next example. The run-time is demanding (almost three hours on my machine) and I do not think that one gets a substantial gain.

This example assesses the maximum TIE of 0.0518 reported by Montague *et al.* for 'Method D' (CV 20%, n_1 12).



With the specified α 0.0280 we obtain an empiric TIE of 0.05180. Note that this 'exact' match is due to chance since the seed is not given in the reference. As in the previous example magenta rhombi show results with random seeds. The mean of optimized alphas is slightly lower (0.02695, TIE 0.04994) than the first estimate (0.02704, TIE 0.04999), but the mean TIE can be >0.05 as shown in yet another run.