Choosing Interim Sample Sizes in Group Sequential Designs

Nancy Flournoy and Sergey Tarima

GMDS-CEN Satellite Meeting

December 3, 2020

Group Sequential Designs

- The number of patients isn't set in advance but is determined by a rule that is a function of cumulative data.
- Data is analyzed at pre-determined points during the experiment, and a decision is made to stop and draw a conclusion regarding the study's objective or continue collecting data.
- Typically, the decision rule at each interim analysis takes the form of a hypothesis test, and emphasis in the design is placed on maintaining the overall type I error rate in the presence of (potentially) multiple tests.
- The expected sample size may be substantially decreased the while controlling overall type I error and power.

Popular Uses

- Group sequential designs (GSDs) allow studies to stop earlier for efficacy or futility at pre-planned interim analysis times.

- GSDs can indicate studies should be stopped for unexpectedly high toxicity rates.

A simple illustrative example with a single interim analysis

Let Z_1 and Z_2 be independent N(θ,1),

but Z_2 is observed only when $Z_1 \leq 1.96$.

- **In study design,** the stopping decision is unknown and the design's characteristics should be evaluated based on the overall distribution of $\hat{\theta}$ where both stopping options are possible.
- **Data analysis** should be based on

$\widehat{\theta_1} =$ MLE given experiment stopped with Z_1; or

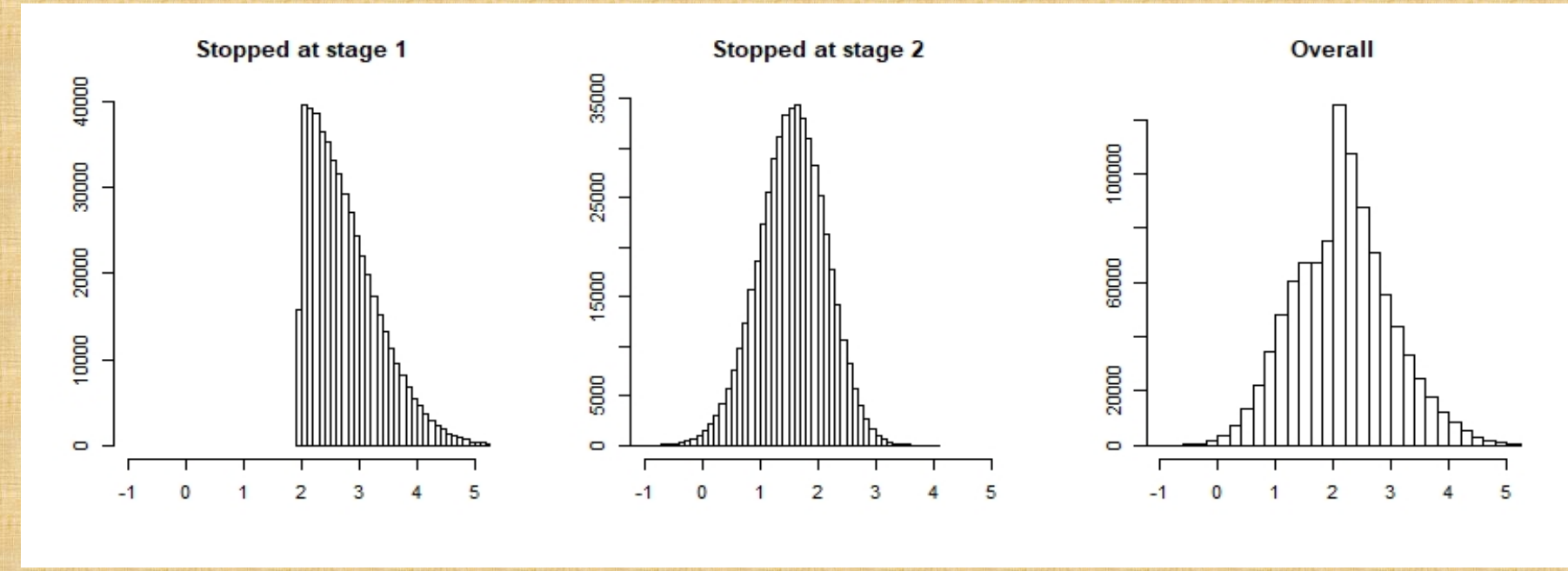
$\widehat{\theta_2} =$ MLE given experiment continued to include Z_2.

The simple example continued

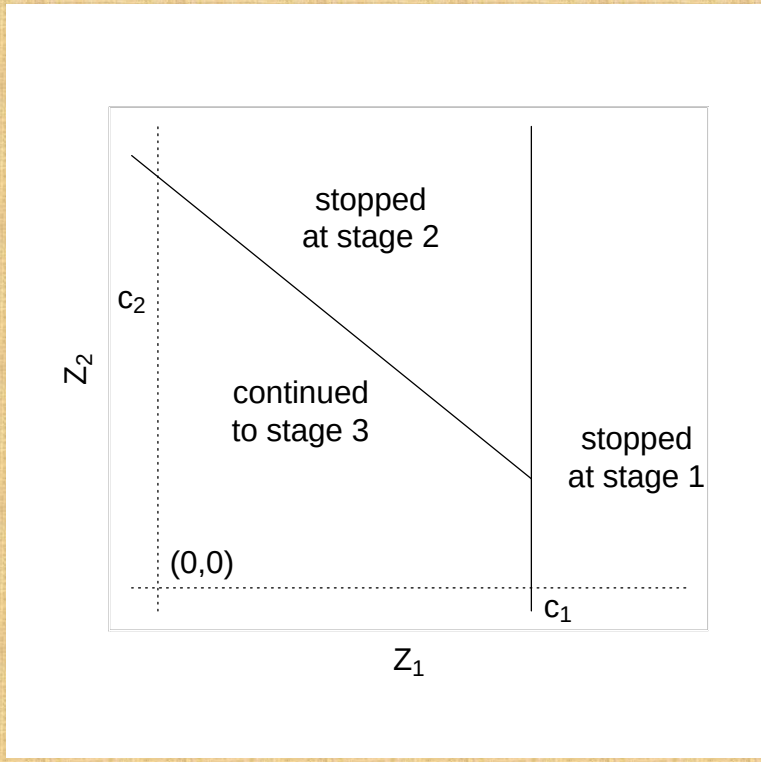
- If Z_1>1.96, the MLE of θ is $\widehat{\theta_1} = Z_1$,
- If $Z_1 \leq 1.96$, the MLE of θ is $\widehat{\theta_2} = (Z_1 + Z_2)/2$.
- Unconditionally, the MLE $\hat{\theta}$ of θ is

$$\hat{\theta} = I(Z_1 > 1.96)\widehat{\theta_1} + I(Z_1 \leq 1.96)\widehat{\theta_2}.$$

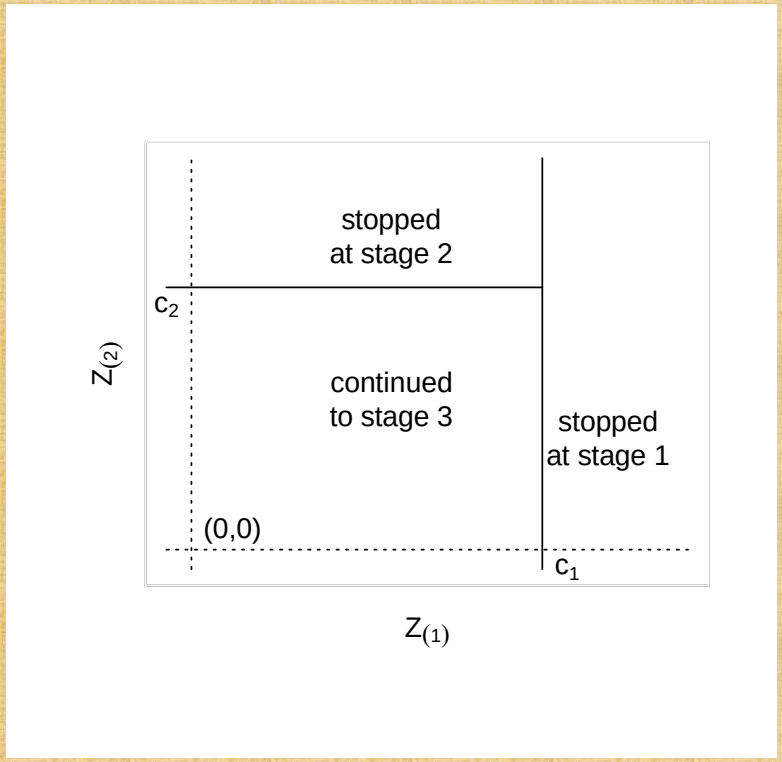
Distributions of the MLE



Support for A Three Stage GSD



Stage-Specific Test Statistics



Cumulative Test Statistics

Pocock's One-Sided 3-Stage GSD

- Suppose a new treatment needs to be tested against a historically established level.

- $Overall\ Testing\ Goal$: $H_0: \theta = 0$ *versus* $H_A: \theta = \theta_A = 1$

FN1

The alternative hypothesis is defined on a standardized scale (mean divided by standard deviation): $\theta = 1$. The use a standardized scale (effect size) eliminates a need to estimate a nuisance parameter (standard deviation) and separates this from the design problem.

- Set $\alpha = \Pr(\text{reject } H_0 \text{ at } 1,2, \text{or } 3) = 0.05$

Pocock's *α−spending function* has the same critical values at all interim analyses

This implies equal sample sizes at each stage:

$$m = m_1 = m_2 = m_3 = \frac{1}{3}n;$$

- $Define\ the\ cumulative\ test\ statistic\ Z_{(k)} = \frac{1}{\sqrt{2mk\sigma^2}}(\sum_{i=1}^{mk} X_i)$, $k = 1,2,3$

where X_i is the observed difference between the outcome and control of the ith subject.

Stop with Z_k if $z_k > C_p$=1.9922, $k = 1,2$

Slide 8

FN1 Flournoy, Nancy, 8/17/2020

Pocock Stopping Probabilities
(SAS Output)

θ	**Stage 1** $n_1 = 244$	**Stage 2** $n_2 = 488$	**Stage 3** $n_3 = 732$
0.0000	**0.02318**	**0.03866**	**0.05000**
0.5000	0.11289	0.22628	0.32918
1.0000	0.33343	0.62094	**0.80000**
1.5000	0.63698	0.91735	0.98455

First row gives **α**-spending function.

Sample sizes found to achieve 80% power at $H_A : \theta = 1$

Group Sequential Design Defined by Ordered Alternatives (SOA)

- Tarima and Flournoy (2019) suggested *simultaneously* targeting constant power against *multiple* ordered alternatives.

- E.g., with α=0.05, choose interim sample sizes to secure 80% statistical power at all three H_A: θ =0.3, θ =0.2, and θ =0.1 regardless of when stopping occurs.

- We used **equal** stage-specific rejection probabilities (0.0172) under H_0: θ =0.0.

- Sample sizes for interim analyses: n_1=98, n_2=196, n_3=772,
- stage-specific critical values: c_1=2.12, c_2=2.02 c_3=2.02.

SOA's Rejection Probabilities by Stage

and Expected Sample Sizes using $\alpha_1 = \alpha_2 = \alpha_3 = 0.0172$.

θ	Pr(Reject at k=1)	Pr(Reject at k≤2)	Pr(Reject at k≤3)	E(N)
0.0	0.0179	0.0337	**0.0509**	750.83
0.1	0.1320	0.3019	**0.7990**	585.18
0.2	0.4473	**0.7985**	0.9998	268.24
0.3	**0.8007**	0.9868	1.0000	125.17

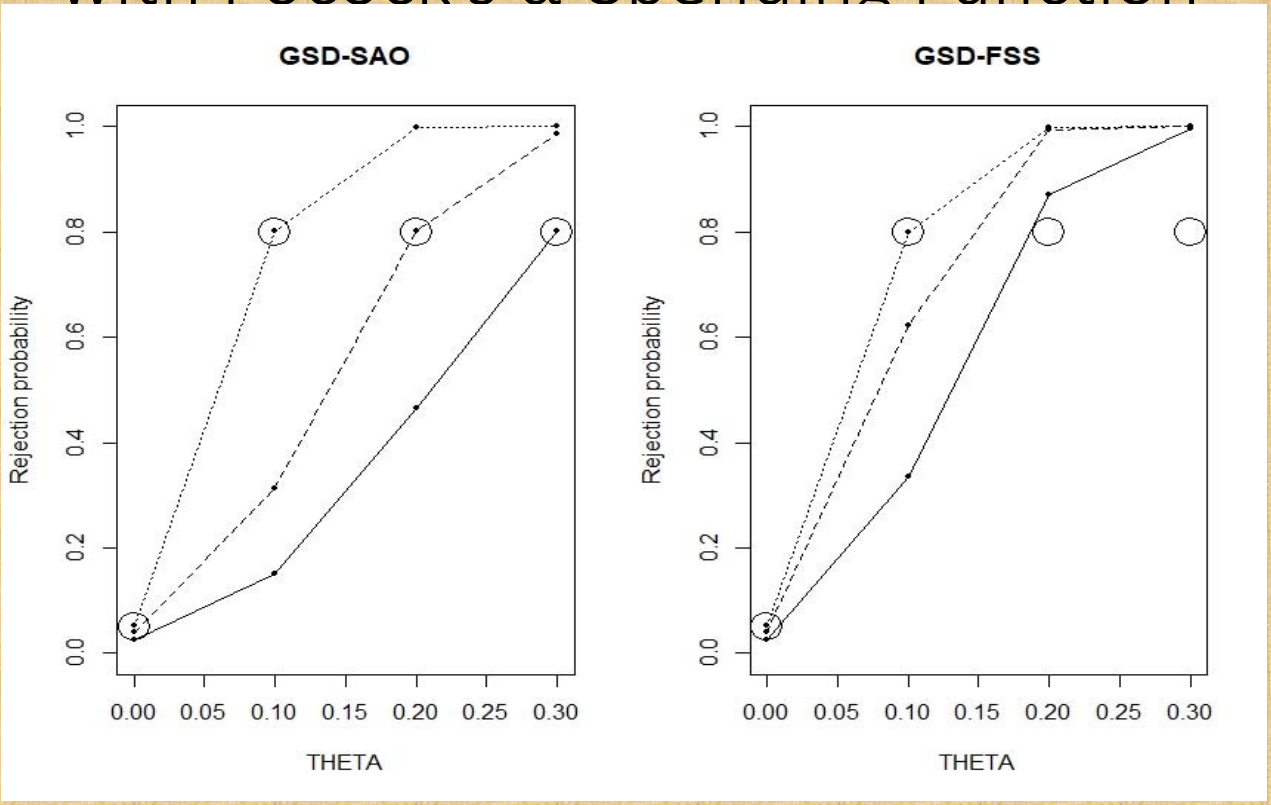
- The SOA design with Pocock's α-spending function has **80% power at each stage**

- Cumulative rejection probabilities 0.0232, 0.0387, 0.0500 are associated with stage-specific rejection probabilities α_1= 0.0232, α_2= 0.0158, and α_3 = 0.0118.

- Interim analyses will be at n_1=90, n_2=205, and n_3=86.
- stage-specific critical values are

  c_1= 1.9921, c_2=2.0216, and c_3=2.1812.
- This leads to **smaller overall expected sample size**

SAO's and Pocock's Cumulative Rejection Probabilities with Pocock's α-Spending Function



α_1= 0.0232
α_2= 0.0158
α_3 = 0.0118

solid line: stage 1
dashed line: stage 2
dotted line: stage 3

Conclusions

1. A possibility of early stopping changes the distribution the MLEs:

   it becomes non-normal.

2. A new group sequential design which controls an overall type 1 error while controlling statistical power at multiple alternatives is proposed.

References

1. P. Armitage, C. McPherson, and B. Rowe. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A* **132** 235–244, 1969.
2. C. Jennison, and B.W. Turnbull. Group Sequential Methods with Applications to Clinical Trials, *Chapman & Hall/CRC Interdisciplinary Statistics*, *CRC Pres*, New York, 1999.
3. M.A. Proschan, K. K. G. Lan, and J. T. Wittes. Statistical Monitoring of Clinical Trials: A Unified Approach, *Springer*, 2006.
4. S.S. Tarima and N. Flournoy. Effect of interim adaptations in group sequential designs, https://arxiv.org/pdf/1908.01411.pdf, 2019, 1-30.
5. S.S. Tarima and N. Flournoy. Asymptotic properties of maximum likelihood estimators with sample size recalculation. *Statistical Papers* **60** (2019), 373–394.
6. S.S. Tarima and N. Flournoy. Distribution theory following blinded and unblinded sample size re-estimation under parametric models. *Communications in Statistics-Simulation and Computation,* 2019, 1-12.
7. S.S. Tarima and N. Flournoy. Choosing interim sample sizes in group sequential designs, in press.
8. S.S. Tarima and N. Flournoy. Most powerful test sequences with early stopping options. Submitted, 2020.

References continued

8. A. Ivanova, W.F. Rosenberger, S.D. Durham and N. Flournoy. A birth and death urn for randomized clinical trials: asymptotic methods. *Sankhya, Series B* **62** 104-118, 2000. A. Ivanova and N. Flournoy A birth and death urn for ternary outcomes: stochastic processes applied to urn models. *Probability and Statistical Models with Applications* 583-600, 2001.
9. C. May and N. Flournoy. Asymptotics in response-adaptive designs generated by a two-color, randomly reinforced urn. *The Annals of Statistics* **37** 1058-1078, 2009.
10. N. Flournoy, C. May and C. Tommasi. The effects of adaptation on inference for non-linear regression models with normal errors. *arXiv preprint arXiv:1812.03970,* 2019.
11. K. Togo and M. Iwasaki. Optimal timing for interim analyses in clinical trials. *Journal of Biopharmaceutical Statistics* **23(5)** 1067-80, 2013.
12. S. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64(2)** 191-199, 1977.
13. T. Koyama and H. Chen. Proper inference from Simon's two-stage designs. Statistics in Medicine **27**(16) 3145-54, 2007.
14. J. Cohen. Statistical power analysis for the behavioral sciences. Academic Press, 2013.