A Comparison of Approaches for Unplanned Sample Size Changes in Phase II Clinical Trials

Molly Olson
molly.a.olson@vanderbilt.edu
Advisor: Tatsuki Koyama

Vanderbilt University
Department of Biostatistics

June 13, 2017

Outline

Phase II Trials

- Phase I: Evaluate safety and dose
- Phase II: Evaluate initial effect to determine phase III trial
- Phase III: Evaluate efficacy

- Phase II - Two-stage with futility stop
  - Mitigate the risk of exposure
  - Don't want to "waste" resources

Two-stage Phase II Trial

$H_0 : p \leq p_0$, $H_1 : p > p_0$, power set at $p_1 > p_0$

1. Stage 1:
   - $X_1 \sim \text{Binomial}(n_1, p) = \#$ of successes in first stage
2. If $x_1 \leq r_1$, trial stopped for futility
3. Otherwise, stage 2:
   - $X_2 \sim \text{Binomial}(n_2, p) = \#$ of successes in second stage
   - $X_t = X_1 + X_2$
   - $n_t = n_1 + n_2$
4. If $x_t \leq r_t$, lack of efficacy concluded
5. Otherwise efficacy concluded

- $n_1, n_t, r_1, r_t, \alpha, \beta$ are design parameters
- n_1, n_t, r_1, r_t are chosen so that type I error rate is less than α and the type II error rate is less than β.

Types of Two-Stage Designs

- Simon introduced Optimal and Minimax criteria for good designs
  - Optimal minimizes the expected sample size under H_0
  - Minimax minimizes the maximum sample size (n_t)
- Jung *et al.* introduced Admissible designs
  - Compromise between Optimal and Minimax

Two-Stage Designs

Suppose $H_0 : p_0 \leq 0.25$, $H_1 : p_1 > 0.4$, $\alpha = 0.05$, $\beta = 0.2$

Design	n_t	n_1	r_1	r_t	EN_0	PET_0
Optimal	71	20	5	23	39.5	0.617
Admissible	63	25	6	21	41.7	0.561
Minimax	60	51	16	20	52.0	0.886

Deviation from the Design

- Attain different enrollment than planned in first and/or second stage
- Why would we deviate?
  - Unanticipated recruitment speed
  - Unanticipated drop out rates
  - Delay in communication for multi-center trials
- Nice properties go out the window
- Currently, common practice is to treat attained sample size as planned
- Leads to inflated type I error
- Hypothesis testing (controlling type I error) is not straightforward

Setting the Scene

- Goal is to make a decision
- How do we do this if our attained sample size is different than planned?
- Consider prespecified “redesigns” - recalculating critical values

Deviation from Planned Sample Sizes in Second Stage

- Over-enrollment in first stage: perform interim analysis on the planned number of first stage, adjust testing procedure for attained second stage
- Under-enrollment: just wait
- Literature exists for point estimation, calculation of p-values when stage II differs (Review: Porcher *et al.*)
- Koyama and Chen, 2008 – calculate new stage II CV s.t. $P[X_2^* \geq r_t^* | X_1 = x_1] \leq P[X_2 \geq r_t | X_1 = x_1]$
- Zeng *et al*, 2015 – calculate new stage II CV s.t. unconditional power maximized while subject to type I error constraint
- Use normal approximation

Deviation from Planned Sample Sizes in First Stage

- Older methods exist (1990s)
- Green & Dahlberg and Chen & Ng of note
- Limitation: Unclear how to generalize
- Limitation: Lacks theoretical justification
- Newer methods introduced

Chang *et al.*

- Chang *et al.* Biometrics & Biostatistics, 2015
- Recall: $n_1, n_t, r_1, r_t, p_0, p_1, \alpha, \beta$
- Notation: n_1^{**}, n_2^{**} - attained sample sizes
- Notation: s_1, s_t - new critical values
- First define β-spending function

$$\beta(n_1^{**}) = \begin{cases} \beta_1 n_1^{**}/n_1 & \text{if } n_1^{**} \leq n_1 \\ \beta_1 + (\beta - \beta_1)(n_1^{**} - n_1)/n_2 & \text{if } n_1^{**} > n_1 \end{cases}$$

- β_1 is planned stage I type II error probability
- Choose s_1 s.t. P[type II error $|n_1^{**}] \approx \beta(n_1^{**})$
- Choose s_t s.t. type I error $\leq \alpha$

Olson and Koyama

- Probability of early termination (PET) – probability that trial stops in first stage, usually under H_0
- Select s_1 s.t. $\text{PET}_0^{**} \approx \text{PET}_0$
- Conservative
- Could have done $\text{PET}_0^{**} \geq \text{PET}_0$, but similar when small deviations

Background: Likelihood

- Law of likelihood: "If $H_1 \Rightarrow P(X = x) = P_1(X)$, $H_2 \Rightarrow P(X = x) = P_2(X)$, then the observation X=x is evidence supporting H_1 over H_2 iff $P_1(X) > P_2(X)$. Likelihood ratio measures strength of evidence."
- Likelihood function:

$$\begin{aligned} \mathrm{L}_n(p) &= P(X|p, n) \\ &= \binom{n}{x} p^x (1-p)^{n-x} \\ &\propto p^x (1-p)^{n-x} \end{aligned} \tag{1}$$

- Likelihood ratio:

$$\mathrm{LR}_n = \frac{\mathrm{L}_n(p_1)}{\mathrm{L}_n(p_2)} \tag{2}$$

Likelihood

- k is a benchmark for strength of evidence
- Can calculate probability of observing weak evidence, strong evidence, misleading evidence
- Universal bound of misleading evidence is $\leq 1/k$

Ayers and Blume

- Likelihood two-stage design
- Enroll n_1
  - $1/k < LR_{n_1} < k \rightarrow$ second stage
  - $LR_{n_1} < 1/k \rightarrow$ stop, conclude futility
  - $LR_{n_1} > k \rightarrow$ stop, conclude efficacy
- Enroll n_2, $LR_{n_t} = LR_{n_1} LR_{n_2}$
  - $1/k < LR_{n_t} < k \rightarrow$ conclude weak
  - $LR_{n_t} < 1/k \rightarrow$ conclude futility
  - $LR_{n_t} > k \rightarrow$ conclude efficacy
- Can add cohorts, easily generalized
- Likelihood unaffected by number of looks at data

Ayers and Blume

- Emulate conventional two-stage designs
  - One look (interim)
  - Two stages
  - Two evidential zones
  - Use of critical values for decision making
- Start with conventional two-stage design
- Conventional $\rightarrow$ Likelihood by redefining lower LR bound, set upper LR bound $= \infty$
- Likelihood $\rightarrow$ conventional by translating LR bounds to critical values

Ayers and Blume

- Can recalculate probability of weak, strong, and misleading evidence, $\mathrm{PET}_0, \mathrm{EN}_0$ under attained
- Minimizes average of error rates
- Type I error rate often below nominal rates
- Inference more straightforward because no concern for error rates or p-values

Comparison of Attained Methods

- Compare methods of Chang *et al.*, OK, Likelihood
- Start with conventional two-stage design
- Vary stage I sample size up to ± 10
- Compare each method using type I and type II error rates, PET, EN
- Keep original total sample size ($n_t^{**} = n_t, n_2^{**} = n_t - n_1^{**}$)
  - Could keep original second stage sample size ($n_t^{**} = n_1^{**} + n_2$)

Comparison of Attained Methods

- Two concrete examples comparing methods
- Simulation results
- Big picture results and takeaways (from situations that we've considered)
- Discussion points

Comparison of Attained Methods

Recall:

- Chang *et al.* method: used type II error β-spending function
- OK method: aimed to keep PET close to planned
- Likelihood: restricted likelihood

Scenerio 1: Low p_0

$$p_0 = 0.10, p_1 = 0.25$$

$$n_1 = 15, n_t = 41$$

$$r_1 = 1, r_t = 7$$

$$\text{PET}_0 = 55\%, \text{EN}_0 = 26.7$$

Method	n_1^{**}	s_1	s_t	α^{**}	$1-\beta^{**}$	PET_0^{**}	EN_0^{**}
Chang *et al.*	13	0	7	0.046	0.830	25%	27.8
OK	13	1	7	0.040	0.771	62%	23.1
Likelihood	13	0	7	0.046	0.830	25%	27.8

Scenerio 1

Accrual close to planned

Method	n_1^{**}	s_1	s_t	α^{**}	$1-\beta^{**}$	PET_0^{**}	EN_0^{**}
Chang *et al.*	13	0	7	0.046	0.830	25%	27.8
OK	13	1	7	0.040	0.771	62%	23.1
Likelihood	13	0	7	0.046	0.830	25%	27.8

Underaccrual

Method	n_1^{**}	s_1	s_t	α^{**}	$1-\beta^{**}$	PET_0^{**}	EN_0^{**}
Chang *et al.*	5	0	7	0.034	0.671	59%	19.7
OK	5	0	7	0.034	0.671	59%	19.7
Likelihood	5	0	7	0.034	0.671	59%	19.7

Overaccrual

Method	n_1^{**}	s_1	s_t	α^{**}	$1-\beta^{**}$	PET_0^{**}	EN_0^{**}
Chang *et al.*	23	3	7	0.040	0.785	80%	26.5
OK	23	2	7	0.046	0.827	59%	30.3
Likelihood	23	2	7	0.046	0.827	59%	30.3

Scenerio 2: High p_0

$$p_0 = 0.75, p_1 = 0.90$$

$$n_1 = 22, n_t = 39$$

$$r_1 = 17, r_t = 33$$

$$\text{PET}_0 = 68\%, \text{EN}_0 = 27.5$$

Method	n_1^{**}	s_1	s_t	α^{**}	$1-\beta^{**}$	PET_0^{**}	EN_0^{**}
Chang *et al.*	26	21	33	0.048	0.791	82%	28.4
OK	26	20	34	0.019	0.650	66%	30.4
Likelihood	26	20	33	0.051	0.810	66%	30.4

Scenerio 2

Accrual close to planned

Method	n_1^{**}	s_1	s_t	α^{**}	$1-\beta^{**}$	PET_0^{**}	EN_0^{**}
Chang *et al.*	20	15	34	0.019	0.650	59%	27.9
OK	20	15	34	0.019	0.659	59%	27.9
Likelihood	20	15	33	0.050	0.805	59%	27.9

Underaccrual Accrual close to planned

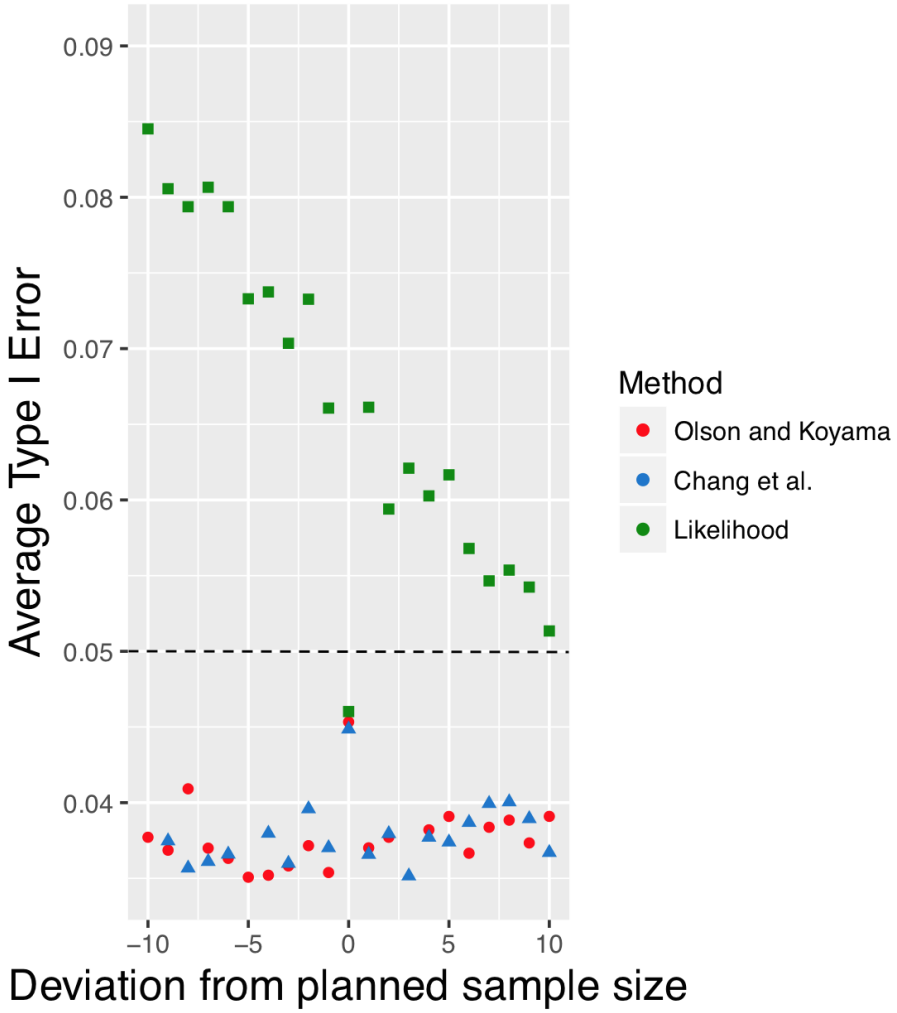
Method	n_1^{**}	s_1	s_t	α^{**}	$1-\beta^{**}$	PET_0^{**}	EN_0^{**}
Chang *et al.*	14	10	33	0.050	0.800	49%	29.5
OK	14	11	33	0.042	0.738	72%	21.0
Likelihood	14	10	33	0.050	0.800	49%	29.5

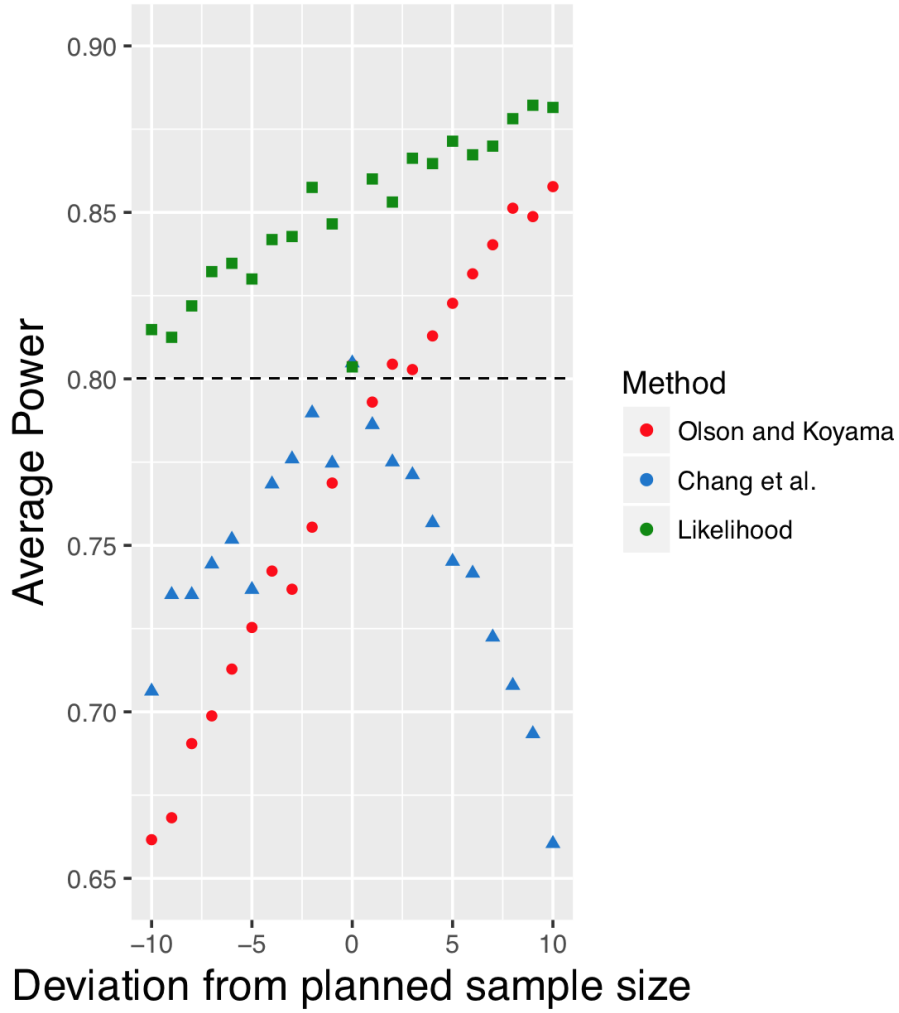
Overaccrual

Method	n_1^{**}	s_1	s_t	α^{**}	$1-\beta^{**}$	PET_0^{**}	EN_0^{**}
Chang *et al.*	32	26	34	0.019	0.650	86%	33.0
OK	32	25	34	0.019	0.650	72%	33.9
Likelihood	32	25	33	0.051	0.810	72%	33.9

Monte Carlo Simulation

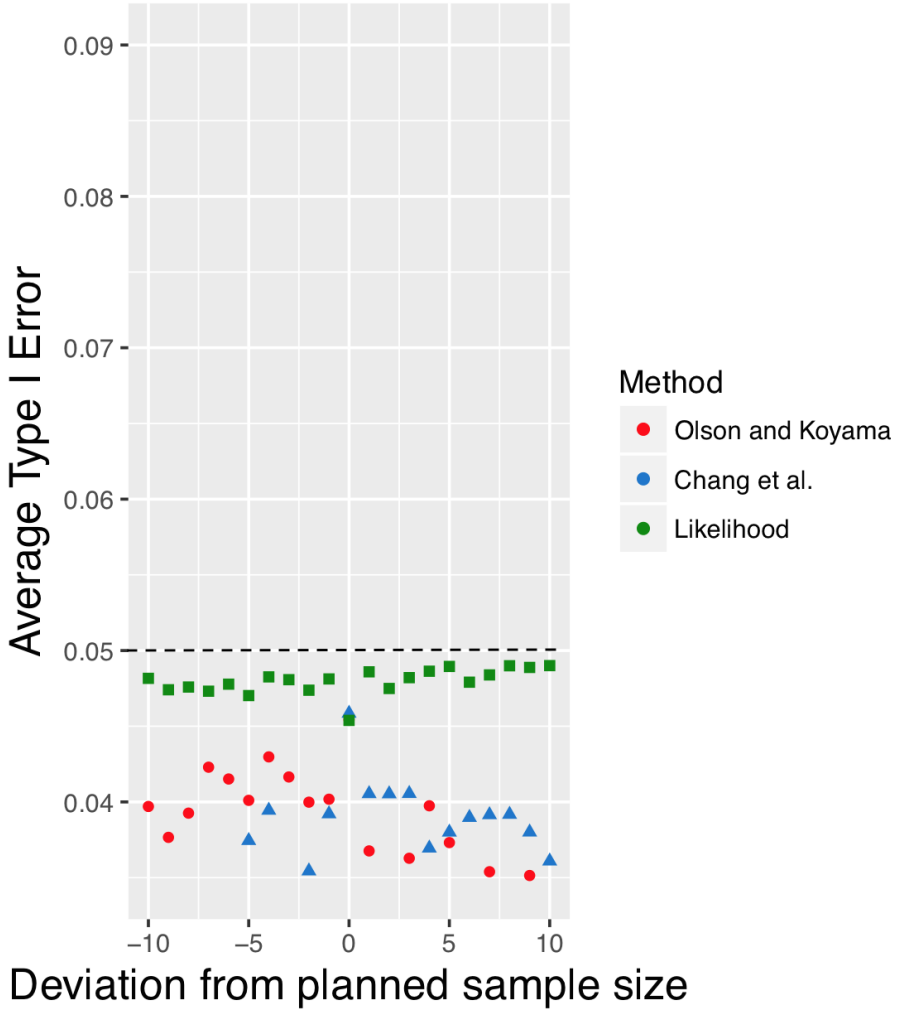
Average error rates of 20 two-stage designs when $n_t^{**} = n_1^{**} + n_2$

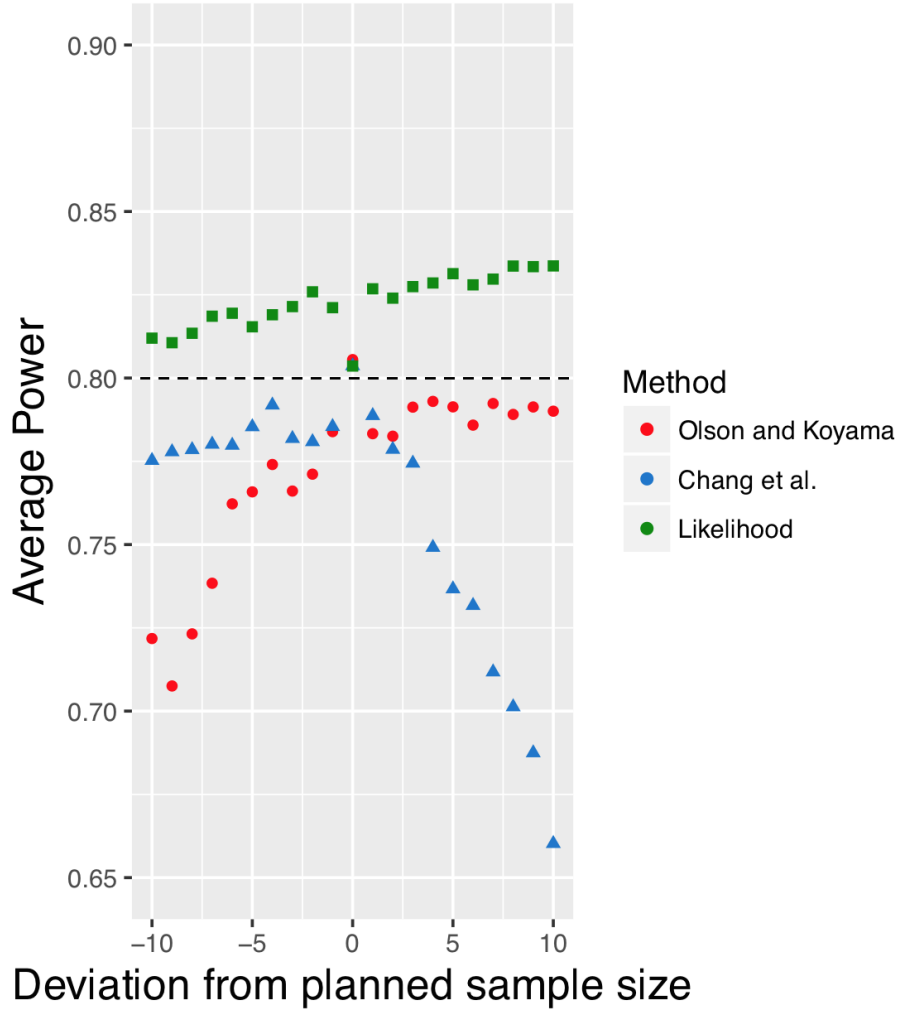




Monte Carlo Simulation

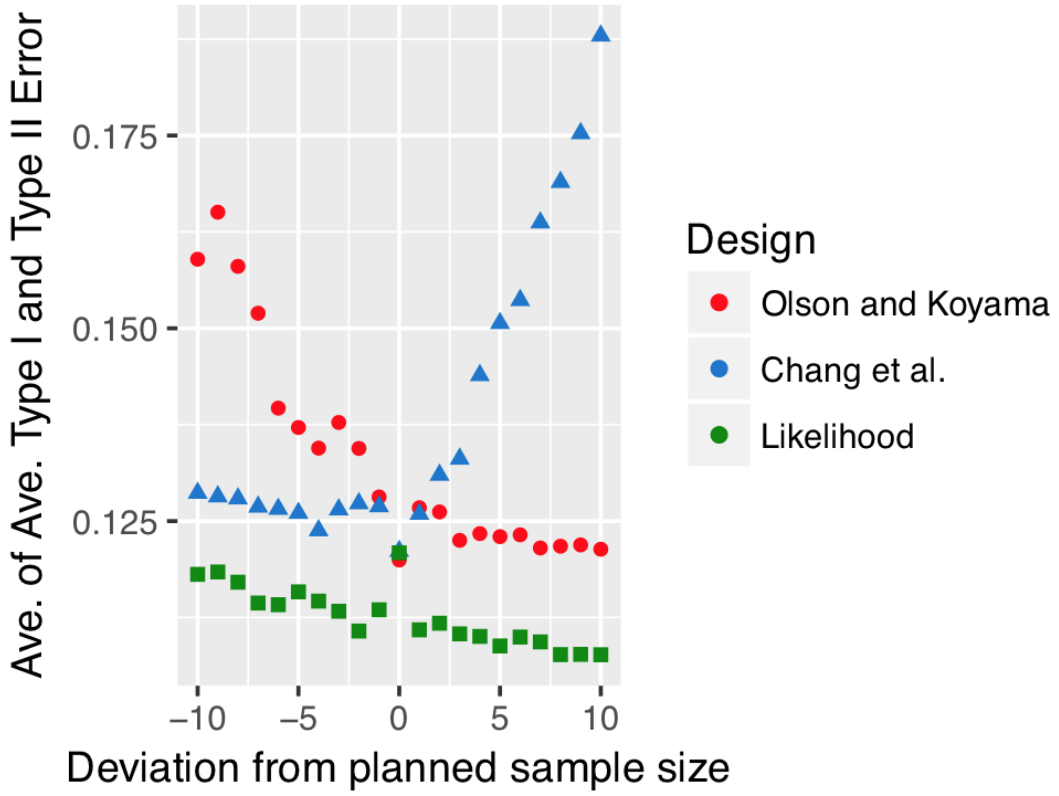
Average error rates of 20 two-stage designs when $n_t^{**} = n_t$





Monte Carlo Simulation

Average of average error rates of 20 two-stage designs when $n_t^{**} = n_t$



Big Picture Results

- Chang *et al.* and OK mostly differ when there are extreme deviations
- OK design and Likelihood are competative when PET above 50%
- One may be more favorable over another depending on hypotheses
- If s_1 is inconsistent between designs for a given deviation, usually within $\pm$ 1 of each other in cases considered

Recommending a Method

- Depends on statistical approach
- Do you want to abandon hypothesis testing?
  - If yes, Likelihood method
  - If no, OK method
- Do you want a conventional two-stage method?
  - Can't go wrong with either Likelihood or OK
- Do you want flexibility?
  - Likelihood

Discussion

- Calculation of p-values is hard – ordering of sample space is not straightforward
- Could calculate ignoring sample path – may have a different decision

Discussion

- Calculation of p-values is hard – ordering of sample space is not straightforward
- Could calculate ignoring sample path – may have a different decision
- Why can't we just wait for stage I sample size?
- Why do we want to keep the original total sample size the same?
  - Resources
  - Simulation results

Discussion

- Calculation of p-values is hard – ordering of sample space is not straightforward
- Could calculate ignoring sample path – may have a different decision
- Why can't we just wait for stage I sample size?
- Why do we want to keep the original total sample size the same?
  - Resources
  - Simulation results
- Attained designs able to accomodate shifts in stage II if needed
- These methods provide solutions to unplanned sample sizes... don't take advantage of them
- May want to use more conservative approach

Future Directions

- More conservative approach to OK design
- Investigate p-value calculation
- Consider a Bayesian approach

Acknowledgements

My deepest appreciation goes to:

- Advisor: Tatsuki Koyama
- DGS and committee member: Jeffrey Blume
- My family
- Faculty
- Fellow graduate students

Questions?

Supplemental slide... More Results: $\alpha = \beta = 0.10$

- $\alpha = \beta = 0.1$
- Stage I sample size low $(p_0 = 0.05, p_1 = 0.20)$
  - Under-accrual, drop in power and type I error
  - Attained $n_1^{**} = 1$, $\text{PET}_0^{**} \approx 1$
- Other two cases, similar results to $\alpha = 0.05, \beta = 0.20$

"Hypothesis testing procedures do not place any interpretation on the numerical value of the LR. The extremeness of an obsevation is measured, not by the magnitude of the LR, but by the probability of observing a likelihood ratio that large or larger. It's the tail area, not the likelihood ratio, that is meaningful quantity in hypothesis testing," Blume, 2002