2024 Summer Institute In Statistics for Clinical & Epidemiological Research

Module 3:

Design, Conduct, and Analysis of Randomized Clinical Trials with Time to Event Primary Endpoints

Lecture 19:

Introduction to Sequential Sampling

Scott S. Emerson, M.D., Ph.D.
Professor Emeritus of Biostatistics
University of Washington

1

Sequential Monitoring

Motivation

Where am I going?

- Ethical and efficiency issues related to RCT are often addressed through interim analyses of the data

2

2

Statistical Sampling Plan

- Ethical and efficiency concerns are addressed through sequential sampling
- During the conduct of the study, data are analyzed at periodic intervals and reviewed by the DMC
- Using interim estimates of treatment effect
  - Decide whether to continue the trial
  - If continuing, decide on any modifications to
    - scientific / statistical hypotheses and/or
    - sampling scheme

3

3

Ultimate Goal

- Modify the sample size accrued so that
  - Minimal number of subjects treated when
    - new treatment is harmful,
    - new treatment is minimally effective, or
    - new treatment is extremely effective
  - Only proceed to maximal sample size when
    - not yet certain of treatment benefit, and
    - potential remains that results of clinical trial will eventually lead to modifying standard practice

4

4

Question

- Under what conditions should we stop the study early?
  - Safety
  - Efficacy
  - Harm
  - Approximate equivalence
  - Futility

5

5

Statistical Criteria

- Extreme estimates of treatment effect
- Statistical significance (Frequentist)
  - At final analysis: Curtailment
  - Based on experimentwise error
    - Group sequential rule
    - Error spending function
- Statistical credibility (Bayesian)
- Probability of achieving statistical significance / credibility at final analysis
  - Condition on current data and presumed treatment effect

6

6

Sequential Sampling Issues

- Design stage
  - Choosing sampling plan which satisfies desired operating characteristics
    - E.g., type I error, power, sample size requirements
- Monitoring stage
  - Flexible implementation to account for assumptions made at design stage
    - E.g., adjust sample size to account for observed variance
- Analysis stage
  - Providing inference based on true sampling distribution of test statistics

7

7

Working Example

- Fixed sample two-sided tests
- Test of a two-sided alternative ($\theta_+ > \theta_0 > \theta_-$)
  - Upper Alternative: H_+: $\theta \geq \theta_+$ (superiority)
  - Null: H_0: $\theta = \theta_0$ (equivalence)
  - Lower Alternative: H_-: $\theta \leq \theta_-$ (inferiority)
- Decisions:
  - Reject H_0 , H_- (for H_+) $\Longleftrightarrow$ $T \geq c_U$
  - Reject H_+ , H_- (for H_0) $\Longleftrightarrow$ $c_L \leq T \leq c_U$
  - Reject H_+ , H_0 (for H_-) $\Longleftrightarrow$ $T \leq c_L$

8

8

Criteria for Monitoring Plans

- Choose a monitoring boundary that terminates with
  - Discrimination between scientifically relevant hypotheses
  - Statistically credible "discrimination"
- Possibility of early termination requires special statistical consideration
  - A "multiple comparison" issue

9

9

Sequential Monitoring

Statistical Issues

Where am I going?

- Sequential monitoring changes the sampling distribution for the data, and thus statistical analysis must account for this change.

10

10

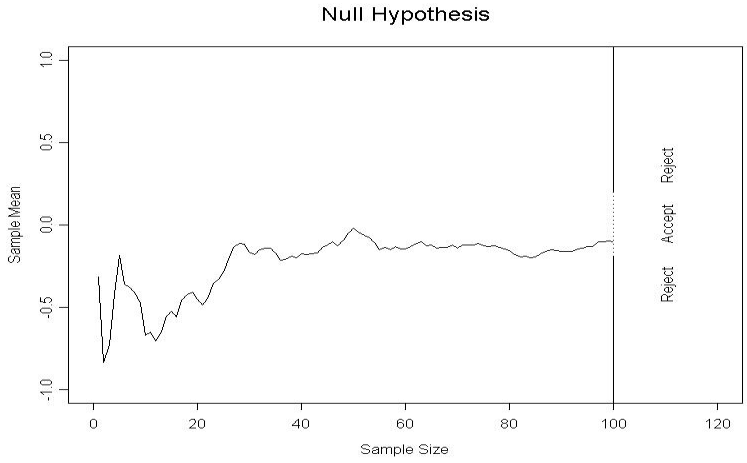
Sampling Plan: General Approach

- Perform analyses when sample sizes $N_1 \ldots N_J$
  - Can be randomly determined
- At each analysis choose stopping boundaries
  - $a_j < b_j < c_j < d_j$
- Compute test statistic $T_j = T(X_1 \ldots X_{Nj})$
  - Stop if $T_j < a_j$ (extremely low)
  - Stop if $b_j < T_j < c_j$ (approximate equivalence)
  - Stop if $T_j > d_j$ (extremely high)
  - Otherwise continue (maybe adaptive modification of analysis schedule, sample size, etc.)
    - Boundaries for modification of sampling plan

11

11

Sample Path for a Statistic

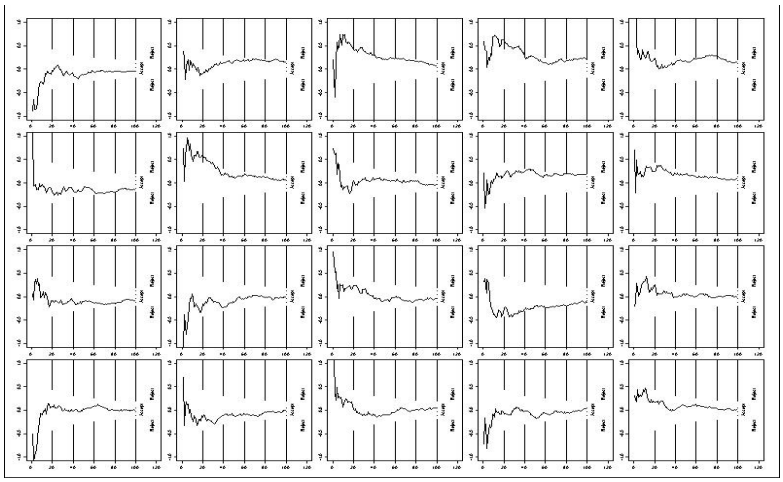


12

12

Fixed Sample Methods Wrong

- Simulated trials under null stop too often



13

13

Simulated Trials (Pocock)

- Three equally spaced level .05 analyses

Pattern of Significance	Proportion Significant 1st	2nd	3rd	Ever
1st only	.03046			.03046
1st, 2nd	.00807	.00807		.00807
1st, 3rd	.00317		.00317	.00317
1st, 2nd, 3rd	.00868	.00868	.00868	.00868
2nd only		.01921		.01921
2nd, 3rd		.01426	.01426	.01426
3rd only			.02445	.02445
Any pattern	.05038	.05022	.05056	.10830

14

14

Pocock Level 0.05

- Three equally spaced level .022 analyses

Pattern of Significance	Proportion Significant 1st	2nd	3rd	Ever
1st only	.01520			.01520
1st, 2nd	.00321	.00321		.00321
1st, 3rd	.00113		.00113	.00113
1st, 2nd, 3rd	.00280	.00280	.00280	.00280
2nd only		.01001		.01001
2nd, 3rd		.00614	.00614	.00614
3rd only			.01250	.01250
Any pattern	.02234	.02216	.02257	.05099

15

15

Unequally Spaced Analyses

- Level .022 analyses at 10%, 20%, 100% of data

Pattern of Significance	Proportion Significant 1st	2nd	3rd	Ever
1st only	.01509			.01509
1st, 2nd	.00521	.00521		.00521
1st, 3rd	.00068		.00068	.00068
1st, 2nd, 3rd	.00069	.00069	.00069	.00069
2nd only		.01473		.01473
2nd, 3rd		.00165	.00165	.00165
3rd only			.01855	.01855
Any pattern	.02167	.02228	.02157	.05660

16

16

Varying Critical Values (OBF)

- Level 0.10 O'Brien-Fleming (1979); equally spaced tests at .003, .036, .087

Pattern of Significance	Proportion Significant 1st	2nd	3rd	Ever
1st only	.00082			.00082
1st, 2nd	.00036	.00036		.00036
1st, 3rd	.00037		.00037	.00037
1st, 2nd, 3rd	.00127	.00127	.00127	.00127
2nd only		.01164		.01164
2nd, 3rd		.02306	.02306	.02306
3rd only			.06223	.01855
Any pattern	.00282	.03633	.08693	.09975

17

17

Error Spending: Pocock 0.05

Pattern of Significance	Proportion Significant 1st	2nd	3rd	Ever
1st only	.01520			.01520
1st, 2nd	.00321	.00321		.00321
1st, 3rd	.00113		.00113	.00113
1st, 2nd, 3rd	.00280	.00280	.00280	.00280
2nd only		.01001		.01001
2nd, 3rd		.00614	.00614	.00614
3rd only			.01250	.01250
Any pattern	.02234	.02216	.02257	.05099
Incremental error	.02234	.01615	.01250	
Cumulative error	.02234	.03849	.05099	

18

18

"Group Sequential Designs"

- At each analysis choose stopping boundaries
  - $a_j < b_j < c_j < d_j$
- "Boundary shape function" defines how conservative the threshold will be at the earliest analyses
  - "O'Brien – Fleming"
    - Very conservative early, like fixed sample late
  - "Triangular test"
    - More efficient for intermediate alternatives
  - "Pocock"
    - Tends toward most efficient for design hypothesis
- Choose critical values to achieve type I error, power

19

19

Role of Sampling Distribution

20

20

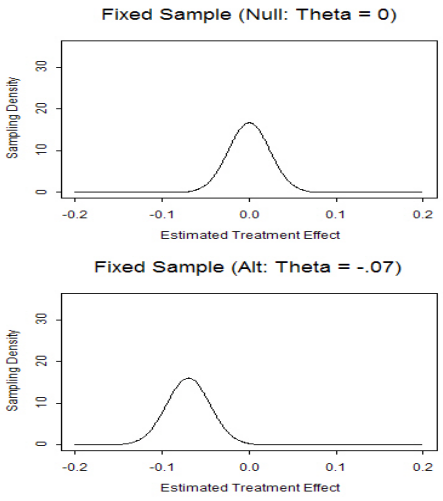
Major Issue

- Frequentist operating characteristics are based on the sampling distribution
- Stopping rules do affect the sampling distribution of the usual statistics
  - MLEs are not normally distributed
  - Z scores are not standard normal under the null
    - (1.96 is irrelevant)
  - The null distribution of fixed sample P values is not uniform
    - (They are not true P values)

21

21

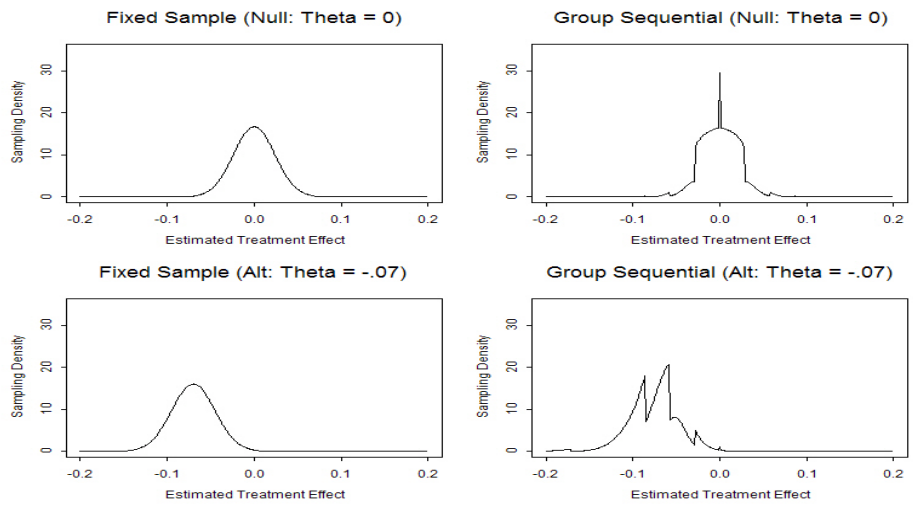
Sampling Distribution of MLE



22

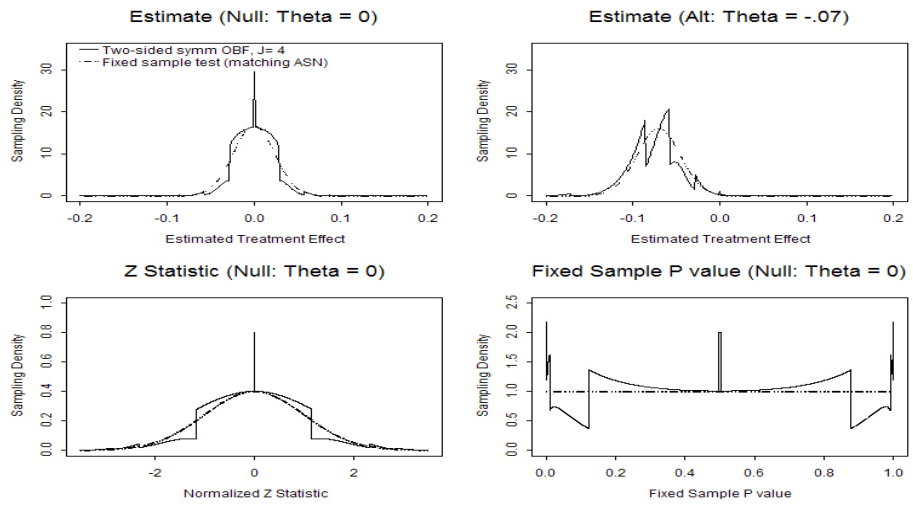
22

Sampling Distribution of MLE



23

Sampling Distributions



24

Sequential Sampling: The Price

- It is only through full knowledge of the sampling plan that we can assess the full complement of frequentist operating characteristics
- In order to obtain inference with maximal precision and minimal bias, the sampling plan must be well quantified
- (Note that adaptive designs using ancillary statistics pose no special problems if we condition on those ancillary statistics.)

25

25

Familiarity and Contempt

- For any known stopping rule, however, we can compute the correct sampling distribution with specialized software
  - Standalone programs
    - PEST (some integration with SAS)
    - EaSt
  - Within statistical packages
    - RCTdesign, gsdesign in R (S-Plus S+SeqTrial)
    - SAS PROC SEQDESIGN

26

26

Familiarity and Contempt

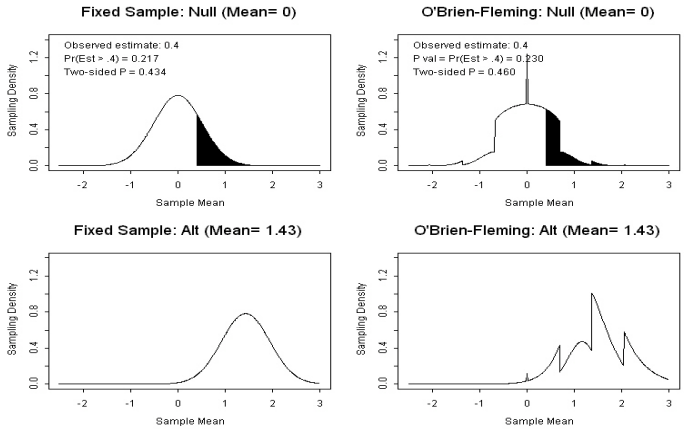
- From the computed sampling distributions we then compute
  - Bias adjusted estimates
  - Correct (adjusted) confidence intervals
  - Correct (adjusted) P values
- Candidate designs can then be compared with respect to their operating characteristics

27

27

Example: P Value

- Null sampling density tail



28

28

Inferential Methods

- Just extensions of methods that also work in fixed samples
- But in fixed samples, many methods converge on the same estimate, unlike in sequential designs

29

29

Point Estimates

- Bias adjusted (Whitehead, 1986)
  - Assume you observed the mean of the sampling distribution
- Median unbiased (Whitehead, 1983)
  - Assume you observed the median of the sampling distribution
- Truncation adapted UMVUE (Emerson & Fleming, 1990)
- (MLE is the naïve estimator: Biased with high MSE)

30

30

Interval Estimates

- Quantile unbiased estimates
  - Assume you observed the 2.5th or 97.5th percentile
- Orderings of the outcome space
  - Analysis time or Stagewise
    - Tend toward wider CI, but do not need entire sampling distribution
  - Sample mean
    - Tend toward narrower CI
  - Likelihood ratio
    - Tend toward narrower CI, but less implemented

31

31

P values

- Orderings of the outcome space
- Analysis time ordering
  - Lower probability of low p-values
  - Insensitive to late occurring treatment effects
- Sample mean
  - High probability of lower p-values
- Likelihood ratio
  - Highest probability of low p-values

32

32