Biost 524: Design of Medical Studies

Lecture 9:
Sequential Stopping Rules

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

May 11, 2011

1

Lecture Outline

- Sequential Monitoring
  - Choice of Stopping Rules
  - Evaluation of Designs
  - Adaptive Designs

2

Sequential Monitoring

Choice of Stopping Rules

Where am I going?

- A wide variety of stopping rules have been proposed for different RCT settings.
- Families of designs have been described on a variety of statistical scales.

3

Statistical Planning

- Satisfy collaborators as much as possible
  - Discriminate between relevant scientific hypotheses
    - Scientific and statistical credibility
  - Protect economic interests of sponsor
    - Efficient designs
    - Economically important estimates
  - Protect interests of patients on trial
    - Stop if unsafe or unethical
    - Stop when credible decision can be made
  - Promote rapid discovery of new beneficial treatments

4

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

5

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_-) $\Longleftrightarrow$ $T \leq c_L$

6

Sampling Plan: General Approach

- Perform analyses when sample sizes N1. . . NJ
  - Can be randomly determined
- At each analysis choose stopping boundaries
  - $a_j < b_j < c_j < d_j$
- Compute test statistic $T_j = T(X_1 . . . X_{Nj})$
  - 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

7

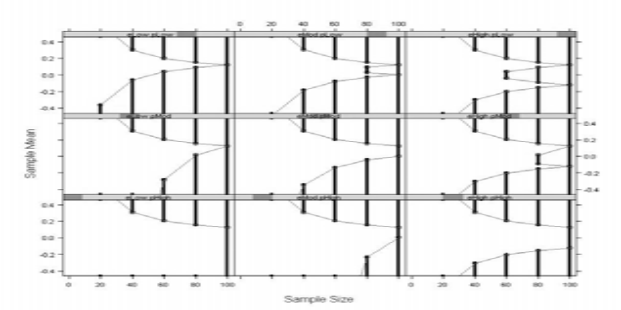
Choice of Stopping Rule

- The choice of stopping rule will vary according to the exact scientific and clinical setting for a clinical trial
  - Each clinical trial poses special problems
  - Wide variety of stopping rules needed to address the different situations
  - (One size does not fit all)

8

Unified Family: MLE Scale

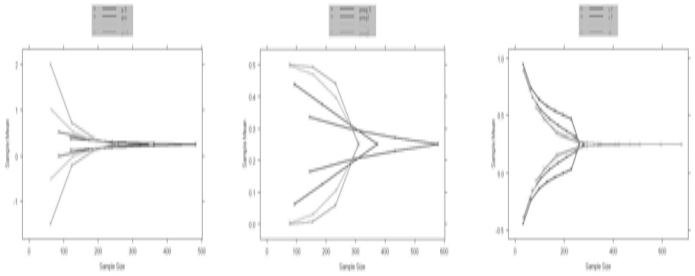
- Down columns: Early stopping vs no early stopping
- Across rows: One-sided vs two-sided decisions



9

Unified Family: MLE Scale

- All of the rules depicted have the same type I error and power to detect the design alternative



Impact on Sampling Density

- When using a stopping rule, the sampling density depends on exact stopping rule
  - This is obvious from what we have already seen.
  - A fixed sample test is merely a particular stopping rule:
    - Gather all N subjects' data and then stop

11

Compared to Fixed Sample

- The magnitude of the effect of the stopping rule on trial design operating characteristics and statistical inference can vary substantially
  - Rule of thumb:
    - The more conservative the stopping rule at interim analyses, the less impact on the operating characteristics and statistical inference when compared to fixed sample designs.

12

Reasons for Early Stopping

- Efficacy, Futility, Harm
- Ethical
  - Individual
    - Protect patients on study
    - Protect patients who might be accrued to study
  - Group
    - Promote rapid discovery of new treatments
- Economic
  - Avoid unnecessary costs of RCT
  - Facilitate earlier marketing

13

Role of Futility Boundaries

- When clinically relevant improvement has been convincingly ruled out and no further useful information to be gained
  - (Is further study of subgroups or other endpoints still in keeping with informed consent?)
- Futility boundaries usually do not indicate harm
- Because most RCT do not reject the null hypothesis, the major savings in early stopping are through a futility boundary
  - Also, not as much need for early conservatism

14

Potential Issue

- Compared to a stopping rule with no futility boundary
  - The critical value at the final analysis can be lower
    - Some of the trials stopped early for futility might have otherwise been type I errors at the final analysis
    - Depends on the early conservatism of the futility boundary

15

Nonbinding Futility

- Some clinical trialists believe that FDA requires that the futility rule be ignored when making inference
  - Such builds in conservatism
  - True type I error is smaller than nominal
  - True power is smaller than normal
- This is purposely using the wrong sampling density
  - Not good statistics—game theory must be motivation

16

Correct Inference

- The statistically correct, efficient approach is to base inference on the real futility boundary
  - Demands correct pre-specification of the futility boundary
  - Demands a clear paper trail of analyses performed

17

Boundary Scales

- Stopping rule for one test statistic is easily transformed to a rule for another statistic
  - "Group sequential stopping rules"
    - Sum of observations
    - Point estimate of treatment effect
    - Normalized (Z) statistic
    - Fixed sample P value
    - Error spending function
  - Bayesian posterior probability
  - Stochastic Curtailment
    - Conditional probability
    - Predictive probability

18

Correspondence Among Scales

- Choices for test statistic T_j
  - All of those choices for test statistics can be shown to be transformations of each other
  - Hence, a stopping rule for one test statistic is easily transformed to a stopping rule for a different test statistic
  - We regard these statistics as representing different scales for expressing the boundaries

19

Relative Advantages

- Which is the best scale to view a stopping rule?
  - Maximum likelihood estimate
  - Z score / fixed sample P value
  - Error spending scale
  - Stochastic curtailment
    - Conditional power
    - Predictive power

20

Statistics Used In Science

- "Scientific scales"
  - Summary measures of the effect
    - Means, medians, geometric means, proportions…
  - Interval estimates for those summary measures
    - (Probabilities merely used to characterize the definition of the interval)
- "Statistical scales"
  - The precision with which you know the true effect
    - Power, P values, posterior probabilities
  - Predictions of the sample you will obtain
    - Conditional power, predictive power

21

Example

- Pre-hospital emergency setting
  - Severe trauma
- Waiver of informed consent
  - Effectiveness studies
  - Impact on prisoners, minors, DOD
  - Notification of participants
- Treatment in field
  - Hospital care according to current local standards
  - Largely passive collection of hospital data

22

Hypertonic Resuscitation

- Hypertonic saline +/- dextran vs normal saline
  - Osmotic pressure to restore blood volume
  - Modulation of immune response during reperfusion
- Hypovolemic shock
  - SBP $\leq$ 70 OR SBP $\leq$ 90 and HR $\geq$ 108
  - Proportion alive at 28 days
    - 4.8% absolute improvement (69.4% vs 64.6%)

23

Sample Size

- Multiple comparison issue
  - HSD vs NS
  - HS vs NS
- Bonferroni adjustment
  - One-sided level 0.0125 tests
- Experimentwise power: 80%
  - Each comparison has 62.6% power
- Sample size: 3,726
  - 1 HSD : 1 HS : 1.414 NS

24

Noninferiority

- Department of Defense
  - 250 cc HS weighs less than 2,000 cc NS
  - Even if no benefit from HS, may want to use if not inferior to NS
- Proving noninferior
  - Define margin of "unacceptably inferior"
    - Absolute decrease of 3%
  - CI at end of trial must exclude the margin
    - 80% confidence interval

25

Okay, so far?

- 4.8% improvement in 28 day survival
  - 28 day survival clinically relevant?
  - 4.8% improvement clinically important?
  - Realistic based on prior knowledge?
- Experimentwise errors
  - HS and HSD clinically equivalent?
  - 0.025 type I error, 80% power statistically credible?

26

Okay, so far?

- Noninferiority
  - 3% decrease justified? In civilians?
  - 80% confidence interval reasonable standard?
  - Are we answering the DoD's questions?
    - (Additional fluids not restricted)
- Sample size of 3,726 without consent?

27

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

28

Protocol Stopping Rule

	N Accrue	Futility Boundary Z	Efficacy Boundary Z
First	621	-4.000	6.000
Second	1,242	-2.800	4.170
Third	1,863	-1.800	3.350
Fourth	2,484	-1.200	2.860
Fifth	3,105	-0.700	2.540
Sixth	3,726	-0.290	2.290

29

Efficacy Boundary

	N Accrue	Efficacy Boundary Z	Crude Diff	Est (95% CI; One-sided P)
First	621	6.000	0.272	0.263 (0.183, 0.329); P < 0.0001
Second	1,242	4.170	0.134	0.129 (0.070, 0.181); P < 0.0001
Third	1,863	3.350	0.088	0.082 (0.035, 0.129); P = 0.0004
Fourth	2,484	2.860	0.065	0.060 (0.019, 0.102); P = 0.0025
Fifth	3,105	2.540	0.052	0.048 (0.010, 0.085); P = 0.0070
Sixth	3,726	2.290	0.042	0.040 (0.005, 0.078); P = 0.0130

30

Futility Boundary

	N Accrue	Futility Boundary Z	Crude Diff	Est (95% CI; One-sided P)
First	621	-4.000	-0.181	-0.172 (-0.238, -0.092); P > 0.9999
Second	1,242	-2.800	-0.090	-0.084 (-0.137, -0.026); P = 0.9973
Third	1,863	-1.800	-0.047	-0.041 (-0.088, 0.006); P = 0.9581
Fourth	2,484	-1.200	-0.027	-0.022 (-0.064, 0.019); P = 0.8590
Fifth	3,105	-0.700	-0.014	-0.010 (-0.048, 0.028); P = 0.7090
Sixth	3,726	-0.290	-0.005	-0.003 (-0.041, 0.032); P = 0.5975

31

Sequential Monitoring

Evaluation of Designs

Where am I going?

- RCT design is most often an iterative process that involves
  - Defining an initial design,
  - Evaluating its operating characteristics, and
  - Modifying the design to better address constraints.

32

Evaluation of Designs

- Process of choosing a trial design
  - Define candidate design
    - Usually constrain two operating characteristics
      - Type I error, power at design alternative
      - Type I error, maximal sample size
  - Evaluate other operating characteristics
    - Different criteria of interest to different investigators
  - Modify design
  - Iterate

33

Collaboration of Disciplines

Discipline	Collaborators	Issues
Scientific	Epidemiologists Basic Scientists Clinical Scientists	Hypothesis generation Mechanisms Clinical benefit
Clinical	Experts in disease / treatment Experts in complications	Efficacy of treatment Adverse experiences
Ethical	Ethicists	Individual ethics Group ethics
Economic	Health services Sponsor management Sponsor marketers	Cost effectiveness Cost of trial / Profitability Marketing appeal
Governmental	Regulators	Safety Efficacy
Statistical	Biostatisticians	Estimates of treatment effect Precision of estimates
Operational	Study coordinators Data management	Collection of data Study burden Data integrity

34

Which Operating Characteristics

- The same regardless of the type of stopping rule
  - Frequentist power curve
    - Type I error (null) and power (design alternative)
  - Sample size requirements
    - Maximum, average, median, other quantiles
    - Stopping probabilities
  - Inference at study termination (at each boundary)
    - Frequentist or Bayesian (under spectrum of priors)
  - (Futility measures
    - Conditional power, predictive power)

35

At Design Stage

- In particular, at design stage we can know
  - Conditions under which trial will continue at each analysis
    - Estimates
      - » (Range of estimates leading to continuation)
    - Inference
      - » (Credibility of results if trial is stopped)
    - Conditional and predictive power
  - Tradeoffs between early stopping and loss in unconditional power

36

Operating Characteristics

- For any stopping rule, however, we can compute the correct sampling distribution with specialized software
  - From the computed sampling distributions we then compute
    - Bias adjusted estimates
    - Correct (adjusted) confidence intervals
    - Correct (adjusted) P values
  - Candidate designs are then compared with respect to their operating characteristics

37

Case Study: Clinical Trial In Gm- Sepsis

- Randomized, placebo controlled Phase III study of antibody to endotoxin
  - Intervention: Single administration
  - Endpoint: Difference in 28 day mortality rates
    - Placebo arm: estimate 30% mortality
    - Treatment arm: hope for 23% mortality
  - Analysis: Large sample test of binomial proportions
    - Frequentist based inference
    - Type I error: one-sided 0.025
    - Power: 90% to detect $\theta < -0.07$
    - Point estimate with low bias, MSE; 95% CI

38

Evaluation: Sample Size

- Number of subjects is a random variable
  - Quantify summary measures of sample size distribution as a function of treatment effect
    - maximum (feasibility of accrual) **(Sponsor)**
    - mean (Average Sample N- ASN) **(Sponsor, DMC)**
    - median, quartiles
  - Stopping probabilities **(Sponsor)**
    - Probability of stopping at each analysis as a function of treatment effect
    - Probability of each decision at each analysis

39

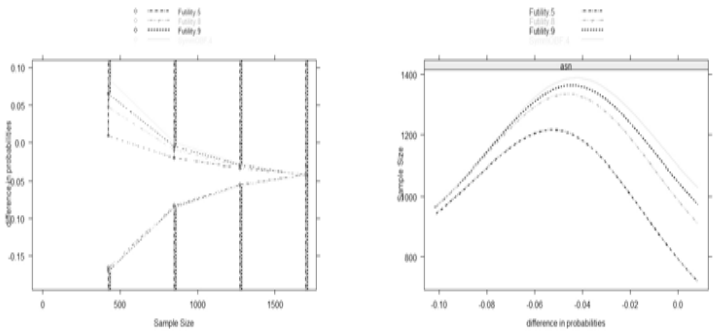
Sample Size

- What is the maximal sample size required?
  - Planning for trial costs
  - Regulatory requirements for minimal N treated
- What is the average sample size required?
  - Hopefully low when treatment does not work or is harmful
  - Acceptable to be high when uncertainty of benefit remains
  - Hopefully low when treatment is markedly effective
    - (But must consider burden of proof)

40

ASN Curve

- Expected sample size as function of true effect



Evaluation: Power Curve

- Probability of rejecting null for arbitrary alternatives
  - Level of significance (power under null) **(Regulatory)**
  - Power for specified alternative
  - Alternative rejected by design
    - Alternative for which study has high power
      - Interpretation of negative studies **(Scientists)**

42

Evaluation: Boundaries

- Decision boundary at each analysis: Value of test statistic leading to early stopping
  - On the scale of estimated treatment effect
    - Inform DMC of precision **(DMC, Statisticians)**
    - Assess ethics **(DMC)**
      - May have prior belief of unacceptable levels
    - Assess clinical importance **(Marketing)**
  - On the Z or fixed sample P value scales **(Often asked for, but of questionable relevance)**

43

Evaluation: Inference

- Inference on the boundary at each analysis
  - Frequentist
    - Adjusted point estimates
    - Adjusted confidence intervals
    - Adjusted P values

    (Scientists, Statisticians, Regulatory)
  - Bayesian
    - Posterior mean of parameter distribution
    - Credible intervals
    - Posterior probability of hypotheses
    - Sensitivity to prior distributions

    (Scientists, Statisticians, Regulatory)

44

Frequentist Inference

	O'Brien-Fleming				Pocock			
N	MLE	Bias Adj Estimate	95% CI	P val	MLE	Bias Adj Estimate	95% CI	P val
Efficacy								
425	-0.171	-0.163	(-0.224, -0.087)	0.000	-0.099	-0.089	(-0.152, -0.015)	0.010
850	-0.086	-0.080	(-0.130, -0.025)	0.002	-0.070	-0.065	(-0.114, -0.004)	0.018
1275	-0.057	-0.054	(-0.096, -0.007)	0.012	-0.057	-0.055	(-0.101, -0.001)	0.023
1700	-0.043	-0.043	(-0.086, 0.000)	0.025	-0.050	-0.050	(-0.099, 0.000)	0.025
Futility								
425	0.086	0.077	(0.001, 0.139)	0.977	0.000	-0.010	(-0.084, 0.053)	0.371
850	0.000	-0.006	(-0.061, 0.044)	0.401	-0.029	-0.035	(-0.095, 0.014)	0.078
1275	-0.029	-0.031	(-0.079, 0.010)	0.067	-0.042	-0.044	(-0.098, 0.002)	0.029
1700	-0.043	-0.043	(-0.086, 0.000)	0.025	-0.050	-0.050	(-0.099, 0.000)	0.025

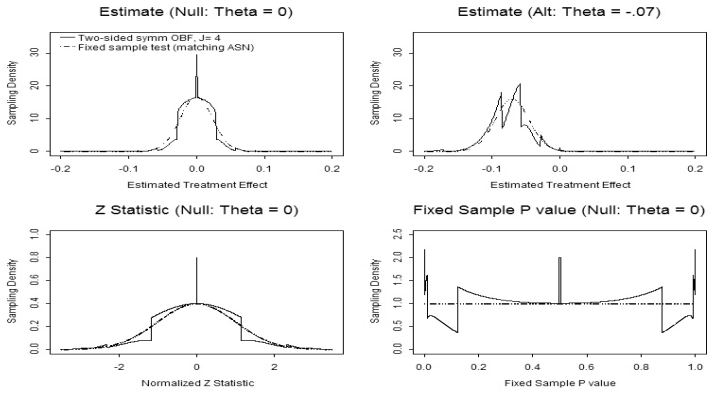
45

At Design Stage: Example

- With O'Brien-Fleming boundaries having 90% power to detect a 7% absolute decrease in mortality
  - Maximum sample size of 1700
  - Continue past 1275 if crude difference in 28 day mortality is between -2.9% and -5.7%
  - If we just barely stop for efficacy after 425 patients we will report
    - Estimated difference in mortality: -16.3%
    - 95% confidence interval: -8.7% to -22.4%
    - One-sided lower P < 0.0001

46

Z scale, Fixed Sample P value

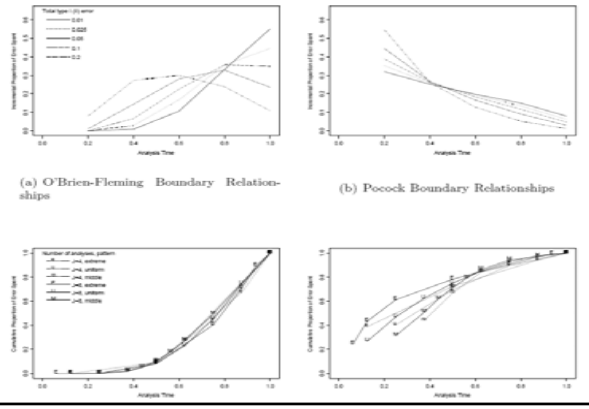


Error Spending Functions

- My view: Poorly understood even by the researchers who advocate them
  - There is no such thing as THE Pocock or O'Brien-Fleming error spending function
    - Depends on type I or type II error
    - Depends on number of analyses
    - Depends on spacing of analyses

48

OBF, Pocock Error Spending



(a) O'Brien-Fleming Boundary Relationships

(b) Pocock Boundary Relationships

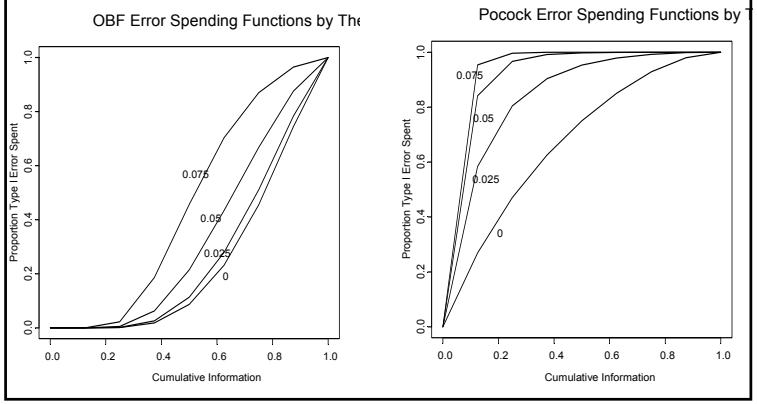
49

Function of Alternative

- Error spending functions depend on the alternative used to compute them
  - The same design has many error spending functions
- JSM 2009: Session on early stopping for harm in a noninferiority trial
  - Attempts to use error spending function approach
  - How to calibrate with functions used for lack of benefit?

50

Error Spent by Alternative



Evaluation: Futility

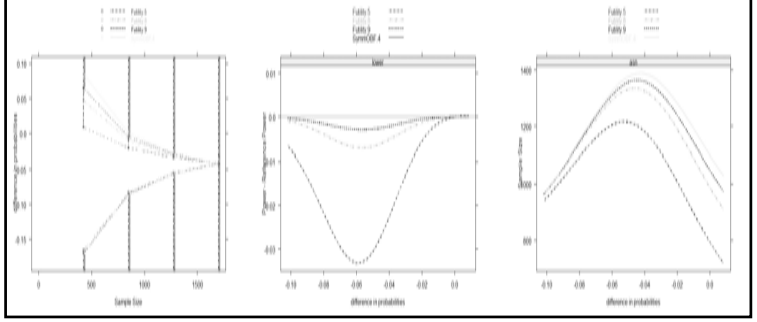
- Consider the probability that a different decision would result if trial continued **(Scientists, Sponsor)**
  - Compare unconditional power to fixed sample test with same sample size
  - Conditional power
    - Assume specific hypotheses
    - Assume current best estimate
  - Predictive power
    - Assume Bayesian prior distribution

(Often asked for, but of questionable relevance)

52

Efficiency / Unconditional Power

- Tradeoffs between early stopping and loss of power
- Boundaries Loss of Power Avg Sample Size



But What If?

- It is common for people to ask about the possibility of a reversed decision
  - But suppose we did not stop for futility. What would be the probability of getting a significant result if we continued to the maximal sample size
- This is easily computed conditional on the observed results IF we know the true treatment effect
  - Conditional power: Assume a particular effect
  - Predictive power: Use a Bayesian prior distribution

54

Stochastic Curtailment

- Stopping boundaries chosen based on predicting future <u>data</u>
- Probability of crossing final boundary
  - Frequentist: Conditional Power
    - A Bayesian prior with all mass on a single hypothesis
  - Bayesian: Predictive Power

55

Stochastic Curtailment

- Boundaries transformed to conditional or predictive power
  - Key issue: Computations are based on assumptions about the true treatment effect
    - Conditional power
      - "Design": based on hypotheses
      - "Estimate": based on current estimates
    - Predictive power
      - "Prior assumptions"

56

So What?

- Why not use stochastic curtailment?
  - What treatment effect should we presume?
    - Hypothesis rejected; current estimate?
  - What threshold should be used for a "low" probability
    - Choice of thresholds poorly understood
      - 10%, 20%, 50%, 80%?
      - How should it depend on sample size and treatment effect
  - Inefficient designs result
    - Conditional and predictive power do not correspond directly to unconditional power

57

Assumed Effect and Threshold

- Probability threshold should take into account the timing of the analysis and the presumed treatment effect
  - It is not uncommon for naïve users to condition on a treatment effect that has already been excluded

58

Predictive Power: Example 1

- Sepsis trial to detect difference in 28 day survival: Null 0.00 vs Alt -0.07 (90% power)
- Futility bounday at first of 4 analyses
  - Futile if observed diff > 0.0473 (so wrong direction)
  - Inference at boundary
    - Bias adjusted: 0.038 (95% CI -0.037 to 0.101)

59

Predictive Power: Example 1

- MLE: 0.0473 Bias Adj: 0.038 (CI: -0.037, 0.101)

Presumed True Effect	Predictive Power
-0.086	71.9%
-0.070	43.2%
-0.037	10.3%
Spons prior	2.8%
Flat prior	0.8%
0.047	<0.005%

60

Predictive Power: Ex 2 (OBF)

- Sepsis trial to detect difference in 28 day survival: Null 0.00 vs Alt -0.07 (90% power)
- Futility bounday at first of 4 analyses
  - Futile if observed diff > 0.0855 (so wrong direction)
  - Inference at boundary
    - Bias adjusted: 0.077 (95% CI 0.000 to 0.139)

61

Predictive Power: Ex 2 (OBF)

- MLE: 0.0855 Bias Adj: -0.077 (CI: 0.000, 0.139)

Presumed True Effect	Predictive Power
-0.086	50.0%
-0.070	26.5%
0.000	.03%
Spons prior	0.3%
Flat prior	0.03%
0.085	<0.005%

62

Key Issues

- Very different probabilities based on assumptions about the true treatment effect
  - Extremely conservative O'Brien-Fleming boundaries correspond to conditional power of 50% (!) under alternative rejected by the boundary
  - Resolution of apparent paradox: if the alternative were true, there is less than .003 probability of stopping for futility at the first analysis

63

Stopping Probs for θ = -0.07

Group Sequential test

		Efficacy	Futility	
N= 425	0.009	< - 0.170	> 0.047	0.003
N= 850	0.298	< - 0.085	> - 0.010	0.022
N= 1275	0.401	< - 0.057	> - 0.031	0.039
N= 1700	0.179	< - 0.042	> - 0.042	0.048
Total	0.888			0.112

64

Apples with Apples

- Can compare a group sequential rule to a fixed sample test providing
  - Same maximal sample size (N= 1700)
  - Same (worst case) average sample size (N= 1336)
  - Same power under the alternative (N= 1598)
- Consider probability of "discordant decisions"
  - Conditional probability (conditional power)
  - Unconditional probability (power)

65

Cond/Uncond Comparison

- Probability of achieving the opposite result at the final analysis
  - Conditional probability
    - Probability among all studies that would stop at that analysis
  - Unconditional probability
    - Change in power of the test due to early stopping

Group Sequential test

	Efficacy		Futility	
	Cond	Uncond	Cond	Uncond
N= 425	0.002	0.000	0.348	0.001
N= 850	0.003	0.001	0.263	0.006
N= 1275	0.009	0.004	0.172	0.007
N= 1700	0.094	0.017	0.182	0.009
Total	0.024	0.022	0.197	0.022

66

Ordering of the Outcome Space

- Choosing a threshold based on conditional power can lead to nonsensical orderings based on unconditional power
  - Decisions based on 35% conditional power may be more conservative than decisions based on 18% conditional power
  - Can result in substantial inefficiency (loss of power)

67

Further Comments

- Neither conditional power nor predictive power have good foundational motivation
  - Frequentists should use Neyman-Pearson paradigm and consider optimal unconditional power across alternatives
    - And conditional/predictive power is not a good indicator in loss of unconditional power
  - Bayesians should use posterior distributions for decisions

68

Evaluation: Marketable Results

- Probability of obtaining estimates of treatment effect with clinical or marketing appeal
  - Modified power curve **(Marketing, Clinicians)**
    - Unconditional
    - Conditional at each analysis
  - Predictive probabilities at each analysis

69

Sequential Monitoring

Adaptive Designs

Where am I going?

- There has been much recent interest in the ability to modify an RCT design in the middle of the trial
- My view: You can always sell perpetual motion machines to the public
  - Many of the "adaptive designs" are strikingly ill-advised on scientific grounds as well as being statistically inefficient

70

Sequential Sampling Strategies

- Two broad categories of sequential sampling
  - Prespecified stopping guidelines
  - Adaptive procedures

71

Adaptive Sampling Plans

- At each interim analysis, possibly modify
  - Scientific and statistical hypotheses of interest
  - Statistical criteria for credible evidence
  - Maximal statistical information
  - Randomization ratios
  - Schedule of analyses
  - Conditions for early stopping

72

Adaptive Sampling: Examples

- Prespecified on the scale of statistical information
  - E.g., Modify sample size to account for estimated information (variance or baseline rates)
    - No effect on type I error IF
      - Estimated information independent of estimate of treatment effect
        - » Proportional hazards,
        - » Normal data, and/or
        - » Carefully phrased alternatives
      - And willing to use conditional inference
        - » Carefully phrased alternatives

73

Estimate Alternative

- If maximal sample size is maintained, the study discriminates between null hypothesis and an alternative measured in units of statistical information

$$n = \frac{\delta_1^{\,2} V}{(\Delta_1 - \Delta_0)^2} \qquad n = \frac{\delta_1^{\,2}}{\left(\frac{(\Delta_1 - \Delta_0)^2}{V}\right)}$$

74

Estimate Sample Size

- If statistical power is maintained, the study sample size is measured in units of statistical information

$$n = \frac{\delta_1^{\,2} V}{(\Delta_1 - \Delta_0)^2} \qquad \frac{n}{V} = \frac{\delta_1^{\,2}}{(\Delta_1 - \Delta_0)^2}$$

75

Adaptive Sampling: Examples

- E.g., Proschan & Hunsberger (1995)
  - Modify ultimate sample size based on conditional power
    - Computed under current best estimate (if high enough)
  - Make adjustment to inference to maintain Type I error

76

Incremental Statistics

- Statistic at the j-th analysis a weighted average of data accrued between analyses

$$\hat{\theta}_j = \frac{\sum_{k=1}^{j} N_k^* \hat{\theta}_k^*}{N_j} \qquad Z_j = \frac{\sum_{k=1}^{j} \sqrt{N_k^*}\, Z_k^*}{\sqrt{N_j}}.$$

77

Conditional Distribution

$$\hat{\theta}_j^* \mid N_j^* \sim N\left(\theta, \frac{V}{N_j^*}\right)$$

$$Z_j^* \mid N_j^* \sim N\left(\frac{\theta - \theta_0}{\sqrt{V / N_j^*}}, 1\right)$$

$$H_0$$

$$P_j^* \mid N_j^* \sim U(0,1).$$

78

Unconditional Distribution

- A mixture of normals, rather than a normal distribution

$$\Pr\left(Z_j^* \le z\right) = \sum_{n=0}^{\infty} \Pr\left(Z_j^* \le z \mid N_j^*\right) \Pr\left(N_j^* = n\right).$$

79

Two Stage Design

- Proschan & Hunsberger consider worst case
  - At first stage, choose sample size of second stage
    - $N_2 = N_2(Z_1)$ to maximize type I error
  - At second stage, reject if $Z_2 > a_2$
- Worst case type I error of two stage design

$$\alpha_{worst} = 1 - \Phi\left(a_2^{(Z)}\right) + \frac{\exp\left(-\left(a_2^{(Z)}\right)^2 / 2\right)}{4},$$

  - Can be more than two times the nominal
    - $a_2 = 1.96$ gives type I error of 0.0616
    - (Compare to Bonferroni results)

80

Better Approaches

- Proschan and Hunsberger describe adaptations using restricted procedures to maintain experimentwise type I error
  - Must prespecify a conditional error function which would maintain type I error
    - Then find appropriate a_2 for second stage based on N_2 which can be chosen arbitrarily
  - But still have loss of power

81

Other Approaches

- Bauer and Kohne:
  - Use R.A. Fisher's method for combining independent P values
- L. Fisher:
  - Variance spending function using prespecified weights at each stage
- Muller and Schafer:
  - Maintain conditional power function from some prespecified fixed sample test

82

Disadvantage Common to All

- Nonintuitive weighting of information from the different stages
  - Stages are not necessarily assigned weights proportional to the statistical information (sample size)
- Violation of the sufficiency principle
  - Inference depends on more information than is available in the minimal sufficient statistic

83

Motivation for Adaptive Designs

- Scientific and statistical hypotheses of interest
  - Modify target population, intervention, measurement of outcome, alternative hypotheses of interest
  - Possible justification
    - Changing conditions in medical environment
      - Approval/withdrawal of competing/ancillary treatments
      - Diagnostic procedures
    - New knowledge from other trials about similar treatments
    - Evidence from ongoing trial
      - Toxicity profile (therapeutic index)
      - Subgroup effects

84

Motivation for Adaptive Designs

- Modification of other design parameters may have great impact on the hypotheses considered
  - Statistical criteria for credible evidence
  - Maximal statistical information
  - Randomization ratios
  - Schedule of analyses
  - Conditions for early stopping

85

Cost of Planning Not to Plan

- Major issues with use of adaptive designs
  - What do we truly gain?
    - Can proper evaluation of trial designs obviate need?
  - What can we lose?
    - Efficiency? (and how should it be measured?)
    - Scientific inference?
      - Science vs Statistics vs Game theory
      - Definition of scientific/statistical hypotheses
      - Quantifying precision of inference

86

Prespecified Modification Rules

- Adaptive sampling plans exact a price in statistical efficiency
  - Tsiatis & Mehta (2002)
    - A classic prespecified group sequential stopping rule can be found that is more efficient than a given adaptive design
  - Shi & Emerson (2003)
    - Fisher's test statistic in the self-designing trial provides markedly less precise inference than that based on the MLE
      - To compute the sampling distribution of the latter, the sampling plan must be known

87

Conditional/Predictive Power

- Additional issues with maintaining conditional or predictive power
  - Modification of sample size may allow precise knowledge of interim treatment effect
    - Interim estimates may cause change in study population
      - Time trends due to investigators gaining or losing enthusiasm
    - In extreme cases, potential for unblinding of individual patients
      - Effect of outliers on test statistics

88

Final Comments

- Adaptive designs versus prespecified stopping rules
  - Adaptive designs come at a price of efficiency and (sometimes) scientific interpretation
  - With adequate tools for careful evaluation of designs, there is little need for adaptive designs

89

Sequential Monitoring

Documentation of Designs

Where am I going?

- Prespecification of the RCT design, monitoring plan, and analysis plan is of utmost importance

90

Specify Stopping Rule

- Null, design alternative hypotheses
- One-sided, two-sided hypotheses
- Type I error, Power to detect design alternative
- For each boundary
  - Hypothesis rejected
  - Error
  - Boundary scale
  - Boundary shape function parameters
- Constraints (minimum, maximum, exact)

91

Documentation of Rule

- Specification of stopping rule
- Estimation of sample size requirements
- Example of stopping boundaries under estimated schedule of analyses
  - sample mean scale, others?
- Inference at the boundaries
- Power under specific alternatives
- Behavior under possible scenarios
  - Alternative baseline rates, variability

92

Implementation

- Method for determining analysis times
- Operating characteristics to be maintained
  - Power (up to some maximum N?)
  - Maximal sample size
- Method for measuring study time
- Boundary scale for making decisions
- Boundary scale for constraining boundaries at previously conducted analyses
- (Conditions stopping rule might be modified)

93

Analysis Plan

- Stopping rule for inference
  - Nonbinding futility?
- Method for determining P values
- Method for point estimation
- Method for confidence intervals
- Handling additional data that accrues after decision to stop

94