Design, Monitoring, and Analysis of Clinical Trials

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

October 16, 2009

Clinical Trials

- Experimentation in human volunteers
  - Investigates a new treatment/preventive agent
    - Safety:
      - » Are there adverse effects that clearly outweigh any potential benefit?
    - Efficacy:
      - » Can the treatment alter the disease process in a beneficial way?
    - Effectiveness:
      - » Would adoption of the treatment as a standard affect morbidity / mortality in the population?

2

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

3

Sample Size Calculation

- Traditional approach
  - Sample size to provide high power to “detect” a particular alternative
- Decision theoretic approach
  - Sample size to discriminate between hypotheses
    - “Discriminate” based on interval estimate
    - Standard for interval estimate: 95%
      - Equivalent to traditional approach with 97.5% power

4

Issues

- Summary measure
  - Mean, geometric mean, median, proportion, hazard…
- Structure of trial
  - One arm, two arms, k arms
  - Independent groups vs cross over
  - Cluster vs individual randomization
  - Randomization ratio
- Statistic
  - Parametric, semi-parametric, nonparametric
  - Adjustment for covariates

5

Refining Scientific Hypotheses

- Scientific hypotheses are typically refined into statistical hypotheses by identifying some parameter θ measuring difference in distribution of response
  - Difference/ratio of means
  - Ratio of geometric means
  - Difference/ratio of medians
  - Difference/ratio of proportions
  - Odds ratio
  - Hazard ratio

6

Inference

- Generalizations from sample to population
  - Estimation
    - Point estimates
    - Interval estimates
  - Decision analysis (testing)
    - Quantifying strength of evidence

7

Measures of Precision

- Estimators are less variable across studies
  - Standard errors are smaller
- Estimators typical of fewer hypotheses
  - Confidence intervals are narrower
- Able to statistically reject false hypotheses
  - Z statistic is higher under alternatives

8

Std Errors: Key to Precision

- Greater precision is achieved with smaller standard errors

Typically: $se(\hat{\theta}) = \sqrt{\frac{V}{n}}$

(V related to average "statistical information")

Width of CI: $2 \times (crit\ val) \times se(\hat{\theta})$

Test statistic: $Z = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})}$

9

Ex: One Sample Mean

$$iid\ Y_i \sim (\mu, \sigma^2),\ i = 1, \ldots, n$$

$$\theta = \mu \qquad \hat{\theta} = \overline{Y}$$

$$V = \sigma^2 \qquad se(\hat{\theta}) = \sqrt{\frac{\sigma^2}{n}}$$

10

Ex: Difference of Indep Means

$$ind\ Y_{ij} \sim (\mu_i, \sigma_i^2),\ i = 1,2;\ j = 1, \ldots, n_i$$

$$n = n_1 + n_2; \quad r = n_1 / n_2$$

$$\theta = \mu_1 - \mu_2 \qquad \hat{\theta} = \overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}$$

$$V = (r+1)\left[\sigma_1^2 / r + \sigma_2^2\right] \qquad se(\hat{\theta}) = \sqrt{\frac{V}{n}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

11

Ex: Difference of Paired Means

$$Y_{ij} \sim (\mu_i, \sigma_i^2),\ i = 1,2;\ j = 1, \ldots, n$$

$$corr(Y_{1j}, Y_{2j}) = \rho; \quad corr(Y_{ij}, Y_{mk}) = 0 \text{ if } j \neq k$$

$$\theta = \mu_1 - \mu_2 \qquad \hat{\theta} = \overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}$$

$$V = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2 \qquad se(\hat{\theta}) = \sqrt{\frac{V}{n}}$$

12

Ex: Mean of Clustered Data

$$Y_{ij} \sim (\mu, \sigma^2), \; i = 1, \ldots, n; \; j = 1, \ldots, m$$
$$corr(Y_{ij}, Y_{ik}) = \rho \text{ if } j \neq k; \quad corr(Y_{ij}, Y_{mk}) = 0 \text{ if } i \neq m$$
$$\theta = \mu_1 - \mu_2 \qquad \hat{\theta} = \overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}$$
$$V = \sigma^2 \left(\frac{1 + (m-1)\rho}{m} \right) \qquad se(\hat{\theta}) = \sqrt{\frac{V}{n}}$$

13

Ex: Independent Odds Ratios

$$ind \; Y_{ij} \sim B(1, p_i), \; i = 1,2; \; j = 1, \ldots, n_i$$
$$n = n_1 + n_2; \quad r = n_1 / n_2$$
$$\theta = \log\left(\frac{p_1 / (1 - p_1)}{p_2 / (1 - p_2)} \right) \qquad \hat{\theta} = \log\left(\frac{\hat{p}_1 / (1 - \hat{p}_1)}{\hat{p}_2 / (1 - \hat{p}_2)} \right)$$
$$\sigma_i^2 = \frac{1}{p_1 (1 - p_1)} = \frac{1}{p_1 q_1}$$
$$V = (r+1)\left[\sigma_1^2 / r + \sigma_2^2 \right] \qquad se(\hat{\theta}) = \sqrt{\frac{V}{n}} = \sqrt{\frac{1}{n_1 p_1 q_1} + \frac{1}{n_2 p_2 q_2}}$$

14

Ex: Hazard Ratios

$$ind \text{ censored time to event } (T_{ij}, \delta_{ij}),$$
$$i = 1,2; \; j = 1, \ldots, n_i; n = n_1 + n_2; \quad r = n_1 / n_2$$
$$\theta = \log(HR) \qquad \hat{\theta} = \hat{\beta} \text{ from PH regression}$$
$$V = \frac{(1+r)(1/r+1)}{\Pr[\delta_{ij} = 1]} \qquad se(\hat{\theta}) = \sqrt{\frac{V}{n}} = \sqrt{\frac{(1+r)(1/r+1)}{d}}$$

15

Ex: Linear Regression

$$ind \; Y_i \mid X_i \sim (\beta_0 + \beta_1 \times X_i, \sigma_{Y|X}^2), \; i = 1, \ldots, n$$
$$\theta = \beta_1 \qquad \hat{\theta} = \hat{\beta}_1 \text{ from LS regression}$$
$$V = \frac{\sigma_{Y|X}^2}{Var(X)} \qquad se(\hat{\theta}) = \sqrt{\frac{\sigma_{Y|X}^2}{nVar(X)}}$$

16

Controlling Variation

- In a two sample comparison of means, we might control some variable in order to decrease the within group variability
  - Restrict population sampled
  - Standardize ancillary treatments
  - Standardize measurement procedure

17

Adjusting for Covariates

- When comparing means using stratified analyses or linear regression, adjustment for precision variables decreases the within group standard deviation
  - Var (Y | X) vs Var (Y | X, W)

18

Ex: Linear Regression

$$ind\ Y_i \mid X_i, W_i \sim \left(\beta_0 + \beta_1 \times X_i + \beta_2 \times W_i, \sigma^2_{Y|X,W}\right), i = 1, \ldots, n$$

$$\theta = \beta_1 \qquad \hat{\theta} = \hat{\beta}_1 \text{ from LS regression}$$

$$V = \frac{\sigma^2_{Y|X,W}}{Var(X)\left(1 - r^2_{XW}\right)} \qquad se\left(\hat{\theta}\right) = \sqrt{\frac{\sigma^2_{Y|X,W}}{nVar(X)\left(1 - r^2_{XW}\right)}}$$

$$\sigma^2_{Y|X,W} = \sigma^2_{Y|X} - \beta_2^2 Var(W \mid X)$$

19

Precision with Proportions

- When analyzing proportions (means), the mean variance relationship is important
  - Precision is greatest when proportion is close to 0 or 1
  - Greater homogeneity of groups makes results more deterministic
    - (At least, I always hope for this)

20

Ex: Diff of Indep Proportions

$$ind\ Y_{ij} \sim B(1, p_i),\ i = 1,2;\ j = 1, \ldots, n_i$$

$$n = n_1 + n_2; \quad r = n_1 / n_2$$

$$\theta = p_1 - p_2 \qquad \hat{\theta} = \hat{p}_1 - \hat{p}_2 = \overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}$$

$$\sigma_i^2 = p_i\left(1 - p_i\right)$$

$$V = (r+1)\left[\sigma_1^2 / r + \sigma_2^2\right] \qquad se\left(\hat{\theta}\right) = \sqrt{\frac{V}{n}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

21

Precision with Odds

- When analyzing odds (a nonlinear function of the mean), adjusting for a precision variable results in more extreme estimates
  - odds = *p / (1-p)*
  - odds using average of stratum specific *p* is not the average of stratum specific odds
- Generally, little "precision" is gained due to the mean-variance relationship
  - Unless the precision variable is highly predictive

22

Precision with Hazards

- When analyzing hazards, adjusting for a precision variable results in more extreme estimates
- The standard error tends to still be related to the number of observed events
  - Higher hazard ratio with same standard error ➔ greater precision

23

Special Case: Baseline Adjustment

- Options
  - Final only (throw away baseline)
    - $V = 2\sigma^2$
  - Change (final – baseline)
    - $V = 4\sigma^2 (1 - \rho)$
  - ANCOVA (change or final adj for baseline
    - $V = 2\sigma^2 (1 - \rho^2)$

24

Ex: ANCOVA (Baseline Adjustment)

$$ind\ Y_{fi} \mid X_i \sim \left(\beta_0 + \beta_1 \times X_i + \beta_1 \times Y_{0i}, \sigma^2_{Y|X,Y_0}\right),$$
$$i = 1, \ldots, n \qquad \rho = corr\left(Y_{0i}, Y_{fi}\right)$$

$$\theta = \beta_1 \qquad \hat{\theta} = \hat{\beta}_1 \text{ from LS regression}$$

$$V = \frac{\sigma^2_{Y|X}\left(1-\rho^2\right)}{Var(X)} \qquad se\left(\hat{\theta}\right) = \sqrt{\frac{\sigma^2_{Y|X}}{nVar(X)}}$$

25

Criteria for Precision

- Standard error
- Width of confidence interval
- Statistical power
  - Probability of rejecting the null hypothesis
    - Select “design alternative”
    - Select desired power

26

Statistics to Address Variability

- At the end of the study:
  - Frequentist and/or Bayesian data analysis to assess the credibility of clinical trial results
    - Estimate of the treatment effect
      - Single best estimate
      - Precision of estimates
    - Decision for or against hypotheses
      - Binary decision
      - Quantification of strength of evidence

27

Sample Size Determination

- Based on sampling plan, statistical analysis plan, and estimates of variability, compute
  - Sample size that discriminates hypotheses with desired power, or
  - Hypothesis that is discriminated from null with desired power when sample size is as specified, or
  - Power to detect the specific alternative when sample size is as specified

28

Sample Size Computation

Standardized level α test (n = 1): $\delta_{\alpha\beta}$ detected with power β

Level of significance α when $\theta = \theta_0$

Design alternative $\theta = \theta_1$

Variability V within 1 sampling unit

Required sampling units: $n = \frac{\left(\delta_{\alpha\beta}\right)^2 V}{\left(\theta_1 - \theta_0\right)^2}$

(Fixed sample test: $\delta_{\alpha\beta} = z_{1-\alpha/2} + z_\beta$)

29

When Sample Size Constrained

- Often (usually?) logistical constraints impose a maximal sample size
  - Compute power to detect specified alternative

Find β such that $\delta_{\alpha\beta} = \sqrt{\frac{n}{V}}\left(\theta_1 - \theta_0\right)$

  - Compute alternative detected with high power

$$\theta_1 = \theta_0 + \delta_{\alpha\beta}\sqrt{\frac{V}{n}}$$

30

Increasing Precision

- Options
  - Increase sample size
  - Decrease V
  - (Decrease confidence level)

31

Comparison of Study Designs

– Single Arm: Mean; absolute reference	N= 25
– Single Arm: Mean; historical data	50
– Two Arms : Diff in Means	100
– Two Arms : Diff in Mean Change (r = 0.3)	140
– Two Arms : Diff in Mean Change (r = 0.8)	40
– Two Arms : ANCOVA (r = 0.3)	81
– Two Arms : ANCOVA (r = 0.8)	36
– Cross-over: Diff in Means (r = 0.3)	70
– Cross-over: Diff in Means (r = 0.8)	20

32

General Comments: Alternative

- What alternative to use?
  - Minimal clinically important difference (MCID)
    - To detect? (use in sample size formula)
    - To declare significant? (look at critical value)
  - Subterfuge: 80% or 90%

33

General Comments: Level

- What level of significance?
  - "Standard": one-sided 0.025, two-sided 0.05
  - "Pivotal": one-sided 0.005?
    - Do we want to be extremely confident of an effect, or confident of an extreme effect

34

General Comments: Power

- What power?
  - Science: 97.5%
    - Unless MCID for significance→ ~50%
  - Subterfuge: 80% or 90%

35

Role of Secondary Analyses

- We choose a primary outcome to avoid multiple comparison problems
  - That primary outcome may be a composite of several clinical outcomes, but there will only be one CI, test
- We select a few secondary outcomes to provide supporting evidence or confirmation of mechanisms
  - Those secondary outcomes may be
    - alternative clinical measures and/or
    - different summary measures of the primary clinical endpoint

36

Secondary Analysis Models

- Selection of statistical models for secondary analyses should generally adhere to same principles as for primary outcome, including intent to treat
- Some exceptions:
  - Exploratory analyses based on dose actually taken may be undertaken to generate hypotheses about dose response
  - Exploratory cause specific time to event analyses may be used to investigate hypothesized mechanisms

37

Subgroups

- Testing for effects in K subgroups
  - Does the treatment work in each subgroup?
  - Bonferroni correction: Test at α / K
    - No subgroups: N = 100
    - Two subgroups: N = 230
- Testing for interactions across subgroups
  - Does the treatment work differently in subgroups?
    - Two subgroups: N = 400

38

Safety Outcomes

- During the conduct of the trial, patients are monitored for adverse events (AEs) and serious adverse events (SAEs)
  - We do not typically demand statistical significance before we worry about the safety profile
    - We must consider the severity of the AE / SAE

39

Safety Outcomes: Conservatism

- If we perform statistical tests, it is imperative that we not use overly conservative procedures
  - When looking for rare events, Fisher's Exact Test is far too conservative
    - Safety criteria based on nonsignificance of FET is a license to kill
  - Unconditional exact tests provide much better power

40

Sample Size Considerations

- We can only choose one sample size
  - Secondary and safety outcomes may be under- or over-powered
- With safety outcomes in particular, we should consider our information about rare, devastating outcomes (e.g., fulminant liver failure in a generally healthy population)
  - The "three over N" rule pertains here
  - Ensure minimal number of treated individuals
    - Control groups are not as important here, if the event is truly rare

41

Sequential Sampling

42

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

43

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

44

Question

- Under what conditions should we stop the study early?

45

Scientific Reasons

- Safety
- Efficacy
- Harm
- Approximate equivalence
- Futility

46

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

47

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

48

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

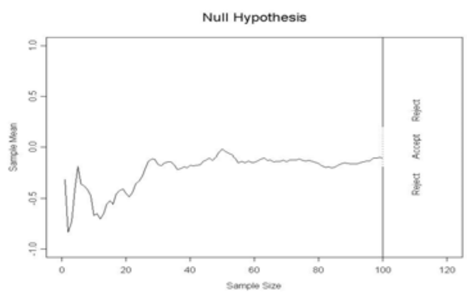
49

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 $T_j = T(X_1 \ldots X_{N_j})$
  - 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

50

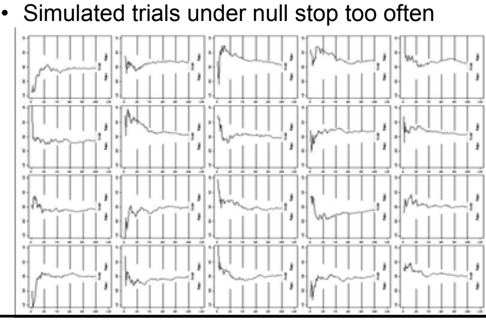
Sample Path for a Statistic



51

Fixed Sample Methods Wrong

- Simulated trials under null stop too often



52

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

53

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

54

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

55

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

56

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	

57

"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

58

Role of Sampling Distribution

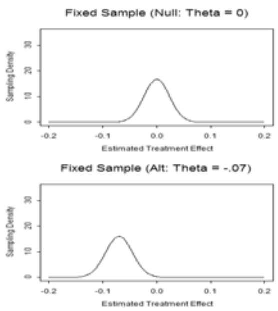
59

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)

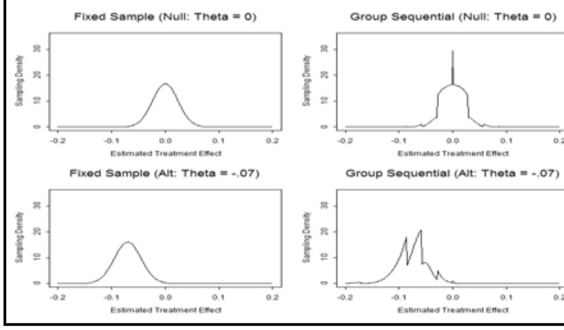
60

Sampling Distribution of MLE

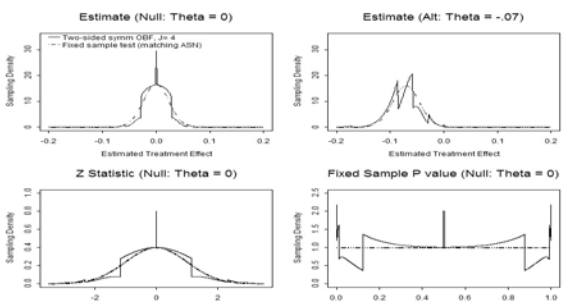


61

Sampling Distribution of MLE



Sampling Distributions



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.)

64

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
    - S-Plus S+SeqTrial
    - SAS PROC SEQDESIGN

65

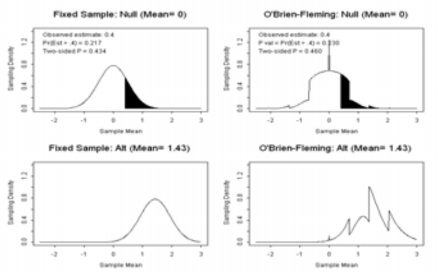
Familiarity and Contempt

- 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

66

Example: P Value

- Null sampling density tail



67

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

68

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)

69

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

70

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

71

Inferential Methods

Example

72

Stopping Boundaries

- The choice of stopping boundaries is typically governed by a wide variety of often competing goals.
  - The appropriateness of any particular boundary will need to be carefully evaluated
- For the present, however, we consider only the basic framework for a stopping rule as a "Sampling Plan".

73

Example

- Two-sided level .05 test of a normal mean (1 sample)
  - Fixed sample design
    - Null: Mean = 0; Alt : Mean = 2
    - Maximal sample size: 100 subjects
  - Early stopping for harm, equivalence, efficacy according to value of sample mean
    - A two-sided symmetric design (Pampallona & Tsiatis, 1994) with a maximum of four analyses and O'Brien-Fleming (1979) boundary shapes

74

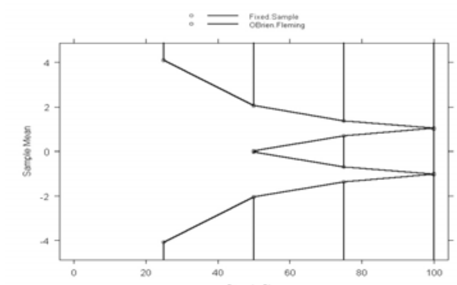
Example

- "O'Brien-Fleming" stopping rule
  - At each analysis, stop early if sample mean is indicated range

N	Harm	Equiv	Efficacy
25	< -4.09	----	> 4.09
50	< -2.05	(-0.006,0.006)	> 2.05
75	< -1.36	(-0.684,0.684)	> 1.36

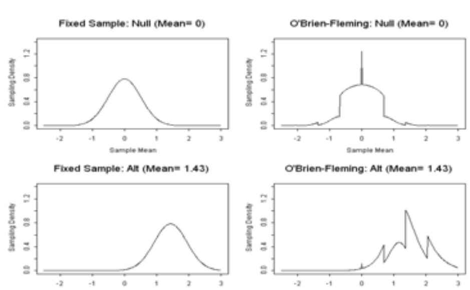
75

Stopping Rule



76

Sampling Densities



77

Statistical Issues

- Estimate of the treatment effect is no longer normally distributed The standardization to a Z statistic does not produce a standard normal
    - The number 1.96 is now irrelevant
  - Converting that Z statistic to a fixed sample P value does not produce a uniform random variable under the null
    - We cannot compare that fixed sample P value to 0.025

78

Statistical Issues: Type I, II Error

- Computation of design operating characteristics needs to use correct sampling density.
  - Type 1 error (size of test)
    - Probability of incorrectly rejecting the null hypothesis
  - Power (1 - type II error)
    - Probability of rejecting the null hypothesis
    - Varies with the true value of the measure of treatment effect

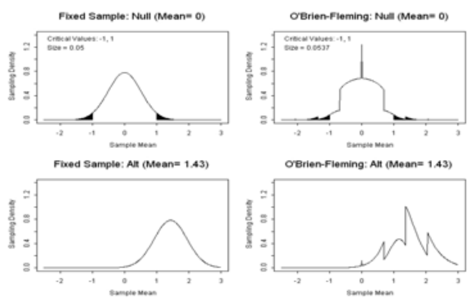
79

Type I Error

- Null sampling density tails beyond crit value
  - Fixed sample test: Mean 0, variance 26.02, N 100
    - Prob that sample mean is greater than 1 is 0.025
    - Prob that sample mean is less than -1 is 0.025
    - Two-sided type I error (size) is 0.05
  - O'Brien-Fleming stopping rule: Mean 0, variance 26.02, max N 100
    - Prob that sample mean is greater than 1 is 0.0268
    - Prob that sample mean is less than -1 is 0.0268
    - Two-sided type I error (size) is 0.0537

80

Type I Error: Area Under Tails



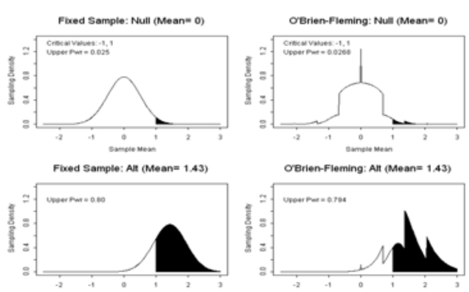
81

Power

- Alternative sampling density beyond crit value
  - Fixed sample test: variance 26.02, N 100
    - Mean 0.00: Prob that sample mean > 1 is 0.025
    - Mean 1.43: Prob that sample mean > 1 is 0.800
    - Mean 2.00: Prob that sample mean > 1 is 0.975
  - O'Brien-Fleming stopping rule: variance 26.02, max N 100
    - Mean 0.00: Prob that sample mean > 1 is 0.027
    - Mean 1.43: Prob that sample mean > 1 is 0.794
    - Mean 2.00: Prob that sample mean > 1 is 0.970

82

Power: Area Under Tail



83

Statistical Issues: Inference

- Measures of statistical inference should be based on the sampling density.
  - Frequentist inferential measures
    - Estimates which
      - minimize bias
      - minimize mean squared error
    - Confidence intervals
    - P values
    - Classical hypothesis testing

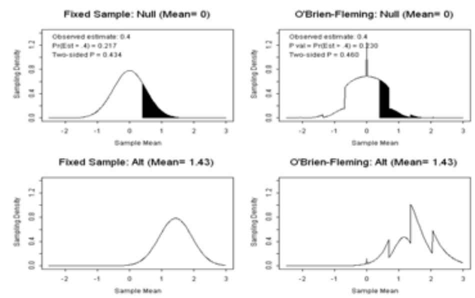
84

P value

- Null sampling density tail beyond observation
  - Fixed sample: Obs 0.4, Mean 0, variance 26.02, N 100
    - Prob that sample mean is greater than 0.4 is 0.217
    - Prob that sample mean is less than 0.4 is 0.783
    - Two-sided P value is 0.434
  - O'Brien-Fleming stopping rule: Obs 0.4, Mean 0, variance 26.02, max N 100
    - Prob that sample mean is greater than 0.4 is 0.230
    - Prob that sample mean is less than 0.4 is 0.770
    - Two-sided P value is 0.460

85

P value: Area Under Tail



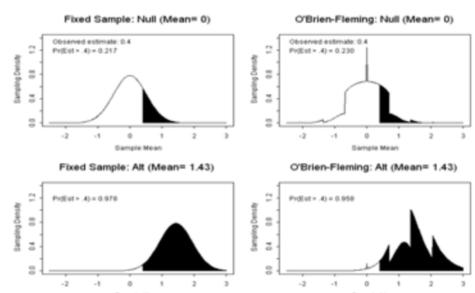
86

Confidence Interval

- Sampling density tail beyond observed value
  - Fixed sample: 95% CI for Obs 0.4, variance 26.02, N 100
    - Mean 0.00: Prob that sample mean > 0.4 is 0.217
    - Mean 1.43: Prob that sample mean > 0.4 is 0.978
    - 95% CI should include 0, but not 1.43
  - O'Brien-Fleming stopping rule: 95% CI for Obs 0.4, variance 26.02, max N 100
    - Mean 0.00: Prob that sample mean > 0.4 is 0.230
    - Mean 1.43: Prob that sample mean > 0.4 is 0.958
    - 95% CI should include 0 and 1.43

87

Confidence Interval



88

Point Estimates

- Effect of sampling distribution on estimates
  - For observed sample mean of 0.4, some point estimates are computed based on summary measures of the sampling distribution.
  - We can examine how the stopping rule affects the summary measures for sampling distribution
    - If they differ, then the corresponding point estimates should differ

89

Sampling Distn Functionals

- Effect of sampling distribution on estimates
  - Sampling distribution summary measures for variance 26.02, max N 100

True treatment effect: Mean = 0.000

Sampling Dist Summary Measure	Fixed Sample	O'Brien-Fleming
Mean	0.000	0.000
Median	0.000	0.000
Mode	0.000	0.000
Maximal for	0.000	0.000

90

Sampling Distn Functionals

- Effect of sampling distribution on estimates
  - Sampling distribution summary measures for variance 26.02, max N 100

True treatment effect: Mean = 0.400

Sampling Dist Summary Measure	Fixed Sample	O'Brien-Fleming
Mean	0.400	0.380
Median	0.400	0.374
Mode	0.400	0.000
Maximal for	0.400	0.400

91

Sampling Distn Functionals

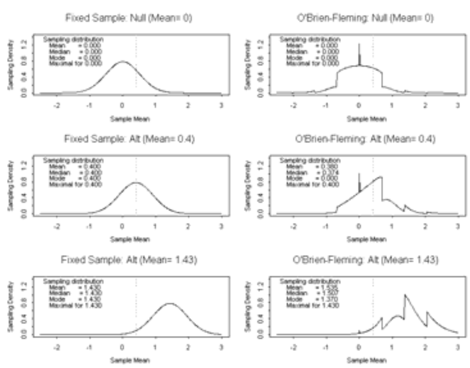
- Effect of sampling distribution on estimates
  - Sampling distribution summary measures for variance 26.02, max N 100

True treatment effect: Mean = 1.430

Sampling Dist Summary Measure	Fixed Sample	O'Brien-Fleming
Mean	1.430	1.535
Median	1.430	1.507
Mode	1.430	1.370
Maximal for	1.430	1.430

92

Alternative Sampling Densities



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)

94

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

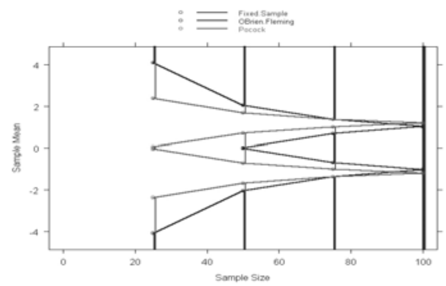
95

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.

96

OBF vs Pocock Stopping Rules



97

Sampling Distn Functionals

- Effect of sampling distribution on estimates
  - Sampling distribution summary measures for variance 26.02, max N 100

True treatment effect: Mean = 0.000

Sampling Dist Summary Measure	Fixed Sample	O'Brien-Fleming	Pocock
Mean	0.000	0.000	0.000
Median	0.000	0.000	0.000
Mode	0.000	0.000	0.000
Maximal for	0.000	0.000	0.000

98

Sampling Distn Functionals

- Effect of sampling distribution on estimates
  - Sampling distribution summary measures for variance 26.02, max N 100

True treatment effect: Mean = 0.400

Sampling Dist Summary Measure	Fixed Sample	O'Brien-Fleming	Pocock
Mean	0.400	0.380	0.372
Median	0.400	0.374	0.333
Mode	0.400	0.000	0.040
Maximal for	0.400	0.400	0.400

99

Sampling Distn Functionals

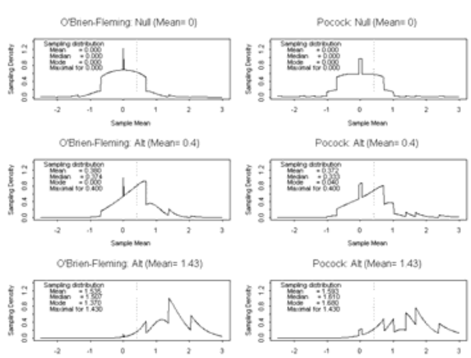
- Effect of sampling distribution on estimates
  - Sampling distribution summary measures for variance 26.02, max N 100

True treatment effect: Mean = 1.430

Sampling Dist Summary Measure	Fixed Sample	O'Brien-Fleming	Pocock
Mean	1.430	1.535	1.593
Median	1.430	1.507	1.610
Mode	1.430	1.370	1.680
Maximal for	1.430	1.430	1.430

100

Alternative Sampling Densities



Nonbinding Futility Boundaries

102

Reasons for Early Stopping

- 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

103

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

104

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

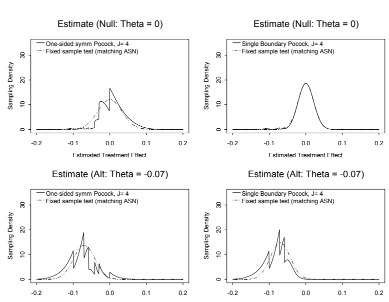
105

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

106

Impact on Sampling Densities



107

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

108

Boundary Scales

109

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

110

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

111

Boundary Scales: Notation

- One sample inference about means
  - Generalizable to most other commonly used models

Probability model:	$X_1, \ldots, X_N \quad iid \quad (\mu, \sigma^2)$
Null hypothesis:	$H_0 : \mu = \mu_0$
Analyses after	$N_1, \ldots, N_J = N$
Data at jth analysis:	$x_1, \ldots, x_{N_j}$

Distributional assumptions:

in absence of a stopping rule $\overline{X}_j \,\dot{\sim}\, N\left(\mu, \frac{\sigma^2}{N_j}\right)$

112

Partial Sum Scale

$$S_j = \sum_{i=1}^{N_j} X_i$$

- Uses:
  - Cumulative number of events
    - Boundary for 1 sample test of proportion
  - Convenient when computing density

113

MLE Scale

$$\overline{X}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} X_i = \frac{S_j}{N_j}$$

- Uses:
  - Natural (crude) estimate of treatment effect

114

Normalized (Z) Statistic Scale

$$z_j = \sqrt{N_j}\,\frac{\left[\overline{x}_j - \mu_0\right]}{\sigma}$$

- Uses:
  - Commonly computed in analysis routines

115

Fixed Sample P Value Scale

$$\begin{aligned} p_j &= 1 - \Phi(z_j) \\ &= 1 - \int_{-\infty}^{z_j} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du \end{aligned}$$

- Uses:
  - Commonly computed in analysis routine
  - Robust to use with other distributions for estimates of treatment effect

116

Error Spending Scale

$$E_{dj} = \frac{1}{\alpha_u}\left[\sum_{i=1}^{j-1} \Pr\left(S_i \ge d_i, \bigcap_{k=1}^{i-1}\left(S_k \in (a_k, b_k) \cup (c_k, d_k)\right) ; \mu_d \right) + \Pr\left(S_j \ge s_j ; \mu_d\right)\right]$$

- Uses:
  - Implementation of stopping rules with flexible determination of number and timing of analyses

117

Bayesian Posterior Scale

Prior distribution $\mu \sim N(\zeta, \tau^2)$

$$\begin{aligned} B_j(\mu_*) &= \Pr\left(\mu \ge \mu_* \mid (X_1, \dots, X_{N_j})\right) \\ &= 1 - \Phi\left(\frac{\mu_*\left[N_j\tau^2 + \sigma^2\right] - N_j\tau^2\overline{x}_j - \sigma^2\zeta}{\sigma\tau\sqrt{N_j\tau^2 + \sigma^2}} \right) \end{aligned}$$

- Uses:
  - Bayesian inference (unaffected by stopping)
  - Posterior probability of hypotheses

118

Conditional Power Scale

Threshold at final analysis $t_{\overline{X}_J}$

Hypothesized value of mean μ_*

$$\begin{aligned} C_j\left(t_{\overline{X}_J}, \mu_*\right) &= \Pr\left(\overline{X}_J \ge t_{\overline{X}_J} \mid \overline{X}_j ; \mu = \mu_*\right) \\ &= 1 - \Phi\left(\frac{N_J\left[t_{\overline{X}_J} - \mu_*\right] - N_j\left[\overline{x}_j - \mu_*\right]}{\sigma\sqrt{N_J - N_j}} \right) \end{aligned}$$

- Uses:
  - Conditional power
    - Probability of significant result at final analysis conditional on data so far (and hypothesis)
  - Futility of continuing under specific hypothesis

119

Conditional Power Scale (MLE)

Threshold at final analysis $t_{\overline{X}_J}$

Hypothesized value of mean $\mu_* = \overline{x}_j$

$$\begin{aligned} C_j\left(t_{\overline{X}_J}, \mu_*\right) &= \Pr\left(\overline{X}_J \ge t_{\overline{X}_J} \mid \overline{X}_j ; \mu = \overline{x}_j\right) \\ &= 1 - \Phi\left(\frac{N_J\left[t_{\overline{X}_J} - \overline{x}_j\right]}{\sigma\sqrt{N_J - N_j}} \right) \end{aligned}$$

- Uses:
  - Conditional power
  - Futility of continuing under specific hypothesis

120

Predictive Power Scale

Threshold at final analysis $t_{\bar{X}_J}$

Prior distribution $\mu \sim N\left(\zeta, \tau^2\right)$

$$H_j\left(t_{\bar{X}_J}\right) = \int \Pr\left(\bar{X}_j \ge t_{\bar{X}_J} \mid \bar{X}_j, \mu\right) \lambda\left(\mu \mid \bar{X}_j\right) d\mu$$

$$= 1 - \Phi\left(\frac{N_J\left[N_j\tau^2 + \sigma^2\right]\left[t_{\bar{X}_J} - \bar{x}_j\right] + \sigma^2\left[N_J - N_j\right]\left[\bar{x}_j - \zeta\right]}{\sigma\sqrt{\left[N_J - N_j\right]\left[N_J\tau^2 + \sigma^2\right]\left[N_j\tau^2 + \sigma^2\right]}} \right)$$

- Uses:
  - Futility of continuing study

121

Predictive Power (Flat Prior)

Threshold at final analysis $t_{\bar{X}_J}$

Prior distribution $\mu \sim N\left(\zeta, \tau^2 \to \infty\right)$

$$H_j\left(t_{\bar{X}_J}\right) = \int \Pr\left(\bar{X}_j \ge t_{\bar{X}_J} \mid \bar{X}_j, \mu\right) \lambda\left(\mu \mid \bar{X}_j\right) d\mu$$

$$= 1 - \Phi\left(\frac{N_J\left[t_{\bar{X}_J} - \bar{x}_j\right]}{\sigma\sqrt{\frac{N_J}{N_j}\left[N_J - N_j\right]}} \right)$$

- Uses:
  - Futility of continuing study

122

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

123

Dangers of Statistical Scales

124

My View

- Statistically

- Scientifically

125

My View

- Statistically
  - It doesn't really matter
- Scientifically
  - You see what a difference it makes

126

Setting

- 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

127

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%)

128

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

129

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

130

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?

131

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?

132

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

133

Protocol Stopping Rule

	N Accrue	Futility Boundary			Efficacy Boundary		
		Z			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		

134

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

135

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

136

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

137

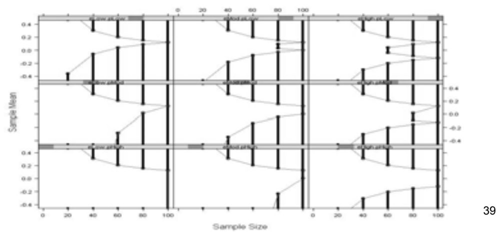
Current Relevance

- Many statisticians (unwisely?) focus on error spending scales, Z statistics, fixed sample P values when describing designs
- Some statisticians have (unwisely?) suggested the use of stochastic curtailment for
  - Defining prespecified sampling plans
  - Adaptive modification of sampling plans

138

Unified Family: MLE Scale

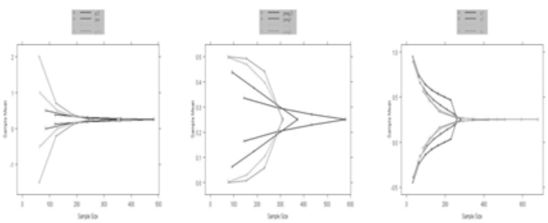
- Down columns: Early stopping vs no early stopping
- Across rows: One-sided vs two-sided decisions



39

Unified Family: MLE Scale

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



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 θ < -0.07
    - Point estimate with low bias, MSE; 95% CI

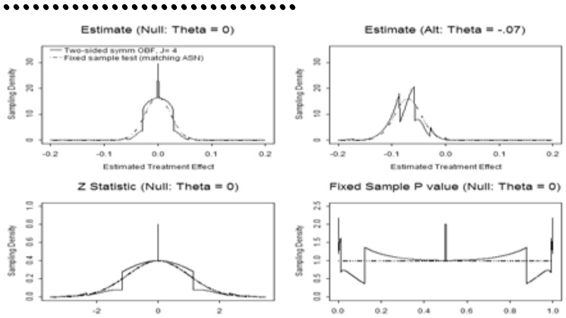
141

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

142

Z scale, Fixed Sample P value

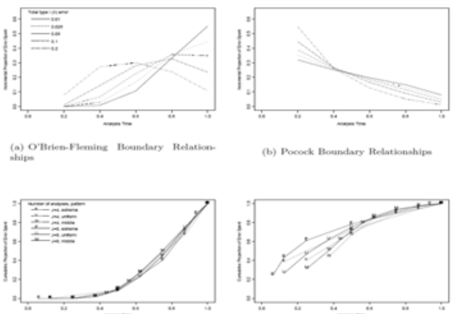


Error Spending Functions

- 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

144

OBF, Pocock Error Spending



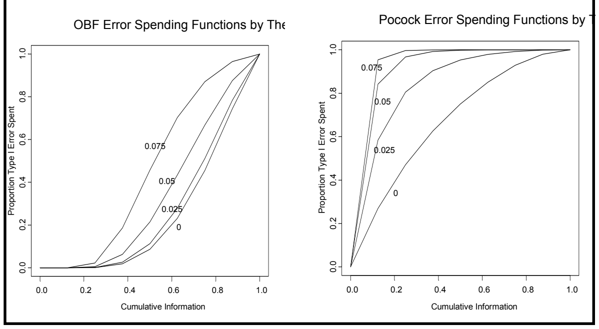
145

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?

146

Error Spent by Alternative



Stochastic Curtailment

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

148

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

149

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"

150

Conditional/Predictive Power

	Symmetric O'Brien-Fleming					O'Brien-Fleming Efficacy, P=0.8 Futility				
		Conditional Power		Predictive Power			Conditional Power		Predictive Power	
N	MLE	Design	Estimate	Sponsor	Noninf	MLE	Design	Estimate	Sponsor	Noninf
	Efficacy (rejects 0.00)					*Efficacy (rejects 0.00)*				
425	**-0.171**	**0.500**	**0.000**	**0.002**	**0.000**	**-0.170**	**0.500**	**0.000**	**0.002**	**0.000**
850	**-0.085**	**0.500**	**0.002**	**0.015**	**0.023**	**-0.085**	**0.500**	**0.002**	**0.015**	**0.023**
1275	**-0.057**	**0.500**	**0.091**	**0.077**	**0.124**	**-0.057**	**0.500**	**0.093**	**0.077**	**0.126**
	Futility (rejects -0.0855)					*Futility (rejects -0.0866)*				
425	**0.085**	**0.500**	**0.000**	**0.077**	**0.000**	**0.047**	**0.719**	**0.000**	**0.222**	**0.008**
850	**0.000**	**0.500**	**0.002**	**0.143**	**0.023**	**-0.010**	**0.648**	**0.015**	**0.247**	**0.063**
1275	**-0.028**	**0.500**	**0.091**	**0.241**	**0.124**	**-0.031**	**0.592**	**0.142**	**0.312**	**0.177**

151

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

152

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

153

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)

154

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%

155

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)

156

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%

157

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

158

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

159

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)

160

Comparable Power for θ = -0.07

- Boundaries based on MLE

Group Sequential test

	Efficacy	Futility
N= 425	< - 0.170	> 0.047
N= 850	< - 0.085	> - 0.010
N= 1275	< - 0.057	> - 0.031
N= 1700	< - 0.042	> - 0.042

Fixed Sample Test

N= 1598	< - 0.043	> - 0.043

161

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

Fixed Sample Test

N= 1598	0.888	< - 0.043	> - 0.043	0.112

162

Cond Prob of Discordance

Group Sequential test

		Efficacy	Futility	
N= 425	0.002	< - 0.170	> 0.047	0.348
N= 850	0.003	< - 0.085	> - 0.010	0.263
N= 1275	0.009	< - 0.057	> - 0.031	0.172
N= 1700	0.094	< - 0.042	> - 0.042	0.182
Total	0.024			0.197

Fixed Sample Test

N= 1598		< - 0.043	> - 0.043	

163

Uncond Prob of Discordance

Group Sequential test

		Efficacy	Futility	
N= 425	0.009	< - 0.170	> 0.047	0.001
N= 850	0.001	< - 0.085	> - 0.010	0.006
N= 1275	0.004	< - 0.057	> - 0.031	0.007
N= 1700	0.017	< - 0.042	> - 0.042	0.009
Total	0.022			0.022

Fixed Sample Test

N= 1598		< - 0.043	> - 0.043	

164

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

Fixed Sample Test

N= 1598	0.888	< - 0.043	> - 0.043	0.112

165

Cond/Uncond Comparison

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

166

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)

167

Other Comparisons

- In the previous example, the fixed sample design had the same power as the GST
  - If we instead compare a fixed sample test having same worst case ASN, the GST would have greater power
  - If we compare a fixed sample test having same maximal N, the GST has less power

168

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

169

Evaluation of Designs

170

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

171

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

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)

173

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

174

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

175

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

176

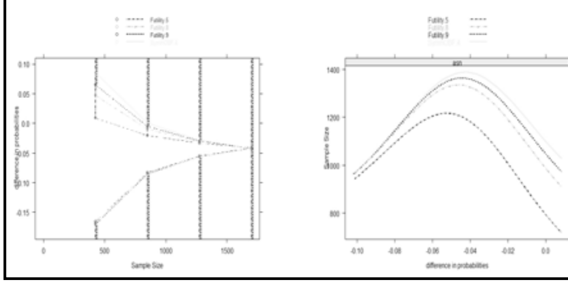
Sample Size

- 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)

177

ASN Curve

- Expected sample size as function of true effect



Evaluation: Power Curve

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

179

Evaluation: Boundaries

- Decision boundary at each analysis: Value of test statistic leading to early stopping
  - On the scale of estimated treatment effect (DMC, Statisticians)
    - Inform DMC of precision
    - 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)

180

Evaluation: Inference

- Inference on the boundary at each analysis
  - Frequentist (Scientists, Statisticians, Regulatory)
    - Adjusted point estimates
    - Adjusted confidence intervals
    - Adjusted P values
  - Bayesian (Scientists, Statisticians, Regulatory)
    - Posterior mean of parameter distribution
    - Credible intervals
    - Posterior probability of hypotheses
    - Sensitivity to prior distributions

181

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

182

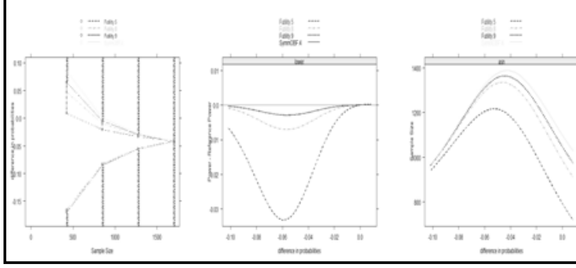
Evaluation: Futility

- 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 (Often asked for, but of questionable relevance)
    - Assume specific hypotheses
    - Assume current best estimate
  - Predictive power
    - Assume Bayesian prior distribution

183

Efficiency / Unconditional Power

- Tradeoffs between early stopping and loss of power



Evaluation: Marketable Results

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

185

Adaptive Sampling Plans

186

Sequential Sampling Strategies

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

187

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

188

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

189

Estimate Alternative

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

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

190

Estimate Sample Size

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

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

191

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

192

Incremental Statistics

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

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

193

Conditional Distribution

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

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

$$P_j^* \mid N_j^* \overset{H_0}{\sim} U(0,1).$$

194

Unconditional Distribution

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

195

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)

196

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

197

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

198

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

199

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

200

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

201

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

202

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

203

Documentation of Design, Monitoring, and Analysis Plans

204

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)

205

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

206

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)

207

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

208