Biost 524: Design of Medical Studies

Lecture 8:
Sample Size Formula;
Sequential Monitoring

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

May 17, 2010

1

Lecture Outline

- Sample Size Formula
- Sequential Monitoring
  - Motivation
  - Statistical Issues

2

Sample Size Formula

Common Settings

Where am I going?

- The most common RCT designs can all use the same sample size formulat

3

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?

4

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

5

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

6

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

7

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

8

Inference

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

9

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

10

Std Errors: Key to Precision

- Greater precision is achieved with smaller standard errors

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

(V related to average "statistical information")

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

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

11

Ex: One Sample Mean

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

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

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

12

Ex: Difference of Indep Means

$$ind\ Y_{ij} \sim \left(\mu_i, \sigma_i^2\right),\ i = 1,2;\ j = 1,\ldots,n_i$$
$$n = n_1 + n_2; \quad r = n_1 / n_2$$
$$\theta = \mu_1 - \mu_2 \qquad \hat{\theta} = \overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}$$

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

13

Ex: Difference of Paired Means

$$Y_{ij} \sim \left(\mu_i, \sigma_i^2\right),\ i = 1,2;\ j = 1,\ldots,n$$
$$corr\left(Y_{1j}, Y_{2j}\right) = \rho; \quad corr\left(Y_{ij}, Y_{mk}\right) = 0 \text{ if } j \neq k$$
$$\theta = \mu_1 - \mu_2 \qquad \hat{\theta} = \overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}$$

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

14

Ex: Mean of Clustered Data

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

$$V = \sigma^2\left(\frac{1 + (m-1)\rho}{m}\right) \qquad se\left(\hat{\theta}\right) = \sqrt{\frac{V}{n}}$$

15

Comment

- Even small correlations are important to consider if cluster sizes are large
  - Equal precision achieved with
    - 1000 clusters with m= 1, ρ = 0.00 (Tot N = 1000)
    - 650 clusters with m= 2, ρ = 0.30 (Tot N = 1300)
    - 550 clusters with m= 2, ρ = 0.10 (Tot N = 1100)
    - 190 clusters with m= 10, ρ = 0.10 (Tot N = 1900)
    - 109 clusters with m= 10, ρ = 0.01 (Tot N = 1090)
    - 20 clusters with m= 100, ρ = 0.01 (Tot N = 2000)

16

Ex: Independent Odds Ratios

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

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

$\theta = \log\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)} \right) \qquad \hat{\theta} = \log\left(\frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} \right)$

$\sigma_i^2 = \frac{1}{p_1(1-p_1)} = \frac{1}{p_1 q_1}$

$V = (r+1)\left[\sigma_1^2/r + \sigma_2^2\right] \qquad se(\hat{\theta}) = \sqrt{\frac{V}{n}} = \sqrt{\frac{1}{n_1 p_1 q_1} + \frac{1}{n_2 p_2 q_2}}$

17

Ex: Hazard Ratios

ind censored time to event (T_{ij}, δ_{ij}),

$i = 1,2;\ j = 1,\ldots,n_i; n = n_1 + n_2; \quad r = n_1 / n_2$

$\theta = \log(HR) \qquad \hat{\theta} = \hat{\beta}$ from PH regression

$V = \frac{(1+r)(1/r+1)}{\Pr[\delta_{ij} = 1]} \qquad se(\hat{\theta}) = \sqrt{\frac{V}{n}} = \sqrt{\frac{(1+r)(1/r+1)}{d}}$

18

Ex: Linear Regression

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

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

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

19

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

20

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)

21

Ex: Linear Regression

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

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

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

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

22

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)

23

Ex: Diff of Indep Proportions

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

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

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

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

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

24

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

25

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

26

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
    - $V = 2\sigma^2 (1 - \rho^2)$

27

Ex: ANCOVA (Baseline Adjustment)

$$ind\ Y_{fi} \mid X_i \sim \left(\beta_0 + \beta_1 \times X_i + \beta_1 \times Y_{0i}, \sigma^2_{Y|X,Y_0}\right),$$
$$i = 1, \ldots, n \qquad \rho = corr\left(Y_{0i}, Y_{fi}\right)$$
$$\theta = \beta_1 \qquad \hat{\theta} = \hat{\beta}_1 \text{ from LS regression}$$

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

28

Criteria for Precision

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

29

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

30

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

31

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: $$n = \frac{(\delta_{\alpha\beta})^2 V}{(\theta_1 - \theta_0)^2}$$

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

32

When Sample Size Constrained

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

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

  - Compute alternative detected with high power

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

33

Increasing Precision

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

34

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

35

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%

36

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

37

General Comments: Power

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

38

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

39

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

40

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

41

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

42

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

43

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

44

Sequential Monitoring

Motivation

Where am I going?

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

45

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

46

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

47

Question

- Under what conditions should we stop the study early?

48

Scientific Reasons

- Safety
- Efficacy
- Harm
- Approximate equivalence
- Futility

49

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

50

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

51

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

52

Criteria for Monitoring Plans

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

53

Sequential Monitoring

Statistical Issues

Where am I going?

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

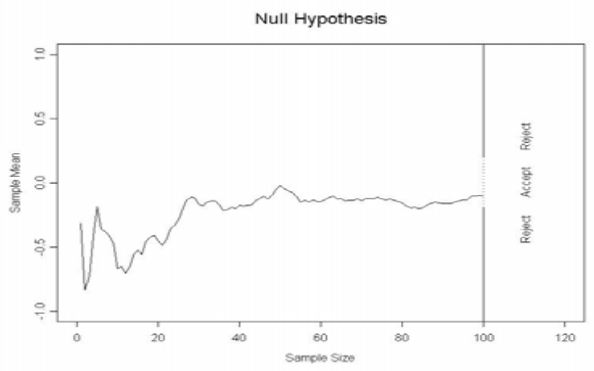
54

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

55

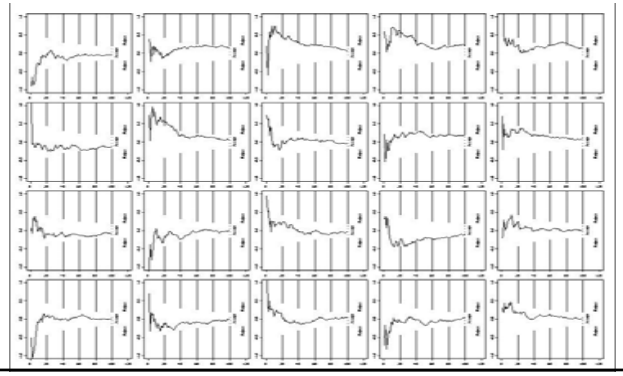
Sample Path for a Statistic



56

Fixed Sample Methods Wrong

- Simulated trials under null stop too often



57

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

58

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

59

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

60

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

61

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	

62

"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

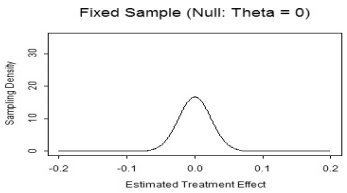
63

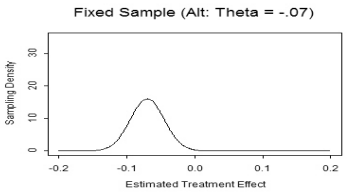
Major Issue

- 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)

64

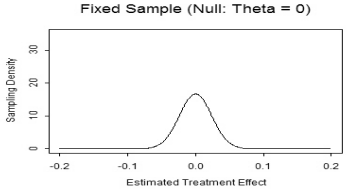
Sampling Distribution of MLE

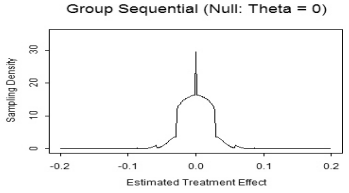


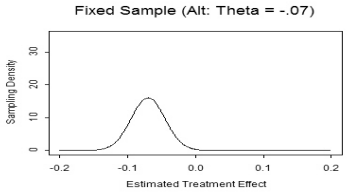


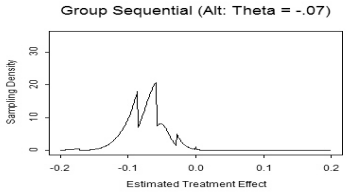
65

Sampling Distribution of MLE

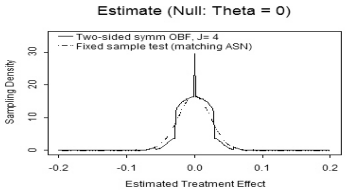


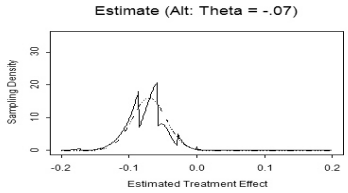


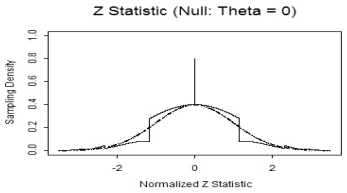


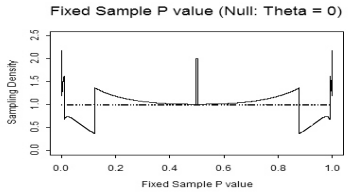


Sampling Distributions









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.)

68

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

69

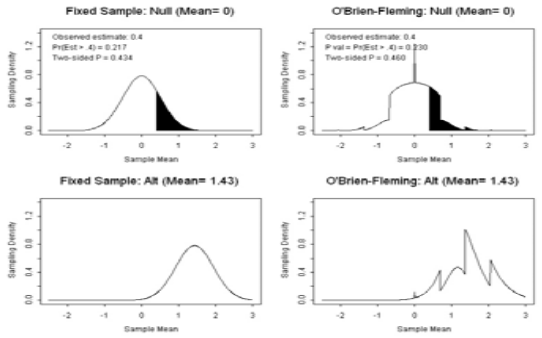
Familiarity and Contempt

- 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

70

Example: P Value

- Null sampling density tail



71

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

72

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)

73

Interval Estimates

- Quantile unbiased estimates
  - Assume you observed the 2.5^{th} or 97.5^{th} 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

74

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

75