Design, Monitoring, and Analysis of Clinical Trials

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

February 17-19, 2003

Course Outline
Day 1

Session 1

- Introduction and Overview
- Fixed Sample Trial Design
- Evaluation of Fixed Sample Designs
- Case study: Fixed sample design
  - Two sample comparison of proportions

Course Outline
Day 1

Session 2

- Group Sequential Stopping Rules
- Families of Designs
- Evaluation of Group Sequential Designs
- Case Study: Group sequential design
  - Two sample comparison of proportions
- Practicum: Basic design using GUI
  - Probability models & hypotheses
  - Power and sample size determination
  - Evaluation of candidate designs

Course Outline
Day 2

Session 3

- Issues in Implementing Stopping Rules
- Recomputation of Sample Size
- Constraining Boundaries at Prior Analyses
- Monitoring Secondary Endpoints
- Case Study: Monitoring a clinical trial
  - Boundary scales: Unified family versus error spending functions
  - Re-estimation of sample size

Course Outline
Day 2

Session 4

Analyses Adjusted for Stopping Rules

Choice of Inferential Methods

Documentation of Design, Monitoring and Analysis

Practicum: Basic monitoring using GUI

- Constrained boundaries
  - Sample mean and error spending scales
- Sample size recomputation
- Adjusted inference

Workshop Outline
Day 3

Session 1

Practicum: Group sequential design

- Further examples
- Advanced GUI features
- Using command line functions
  - Plots, reports, simulations
  - Less common evaluation criteria

Workshop Outline
Day 3

Session 2

Practicum: Special topics

- Nonparametric applications
  - Nonproportional hazards
- Poorly specified stopping rules
- Bayesian stopping rules

Session 1

Overview and Introduction

Overview

Fixed Sample Trial Design

Fundamental Clinical Trial Design

- Common Probability Models
- Defining the Hypotheses
- Defining the Criteria for Evidence
- Determining the Sample Size

Evaluation of Fixed Sample Designs

Case Study

Overview

Overview

Science and statistics

What is science?

- Clinical trial setting

Why statistics?

Sequential clinical trials

Ethical concerns

Statistical issues

Overview

Clinical trials

Experimentation in human volunteers

Investigation of a new treatment or 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?

Overview

Often competing goals must be considered

Scientific (basic science):

- focus on questions about mechanisms

Ethical:

- focus on minimizing harm to human volunteers

Clinical:

- focus on improving overall health of patients

Statistical:

- focus on questions that can be answered precisely

Overview

As an experiment, a clinical trial must meet scientific standards

It must address a meaningful question

- discriminate between viable hypotheses (Science)

Its results must be credible to scientific community

- Valid materials, methods (Science, Statistics)
- Valid measurement of experimental outcome (Science, Clinical, Statistics)
- Valid quantification of uncertainty in experimental procedure (Statistics)

Scientific Experiments

Scientific Experimentation

Goals

A well designed experiment **<u>discriminates between</u>** hypotheses <u>(The Scientist Game)</u>

- The hypotheses should be the most important, viable hypotheses
- All other things being equal, it should be equally informative for all possible outcomes
  - Binary search (using prior probability of being true)
  - But may need to consider simplicity of experiments, time, cost

Scientific Experimentation

At the end of the experiment, we want to present results that are convincing to the scientific community

The limitations of the experiment must be kept in mind

- Statistics means never having to say you are certain.

  -ASA T-shirt

This also holds more generally for science

- Distinguish results from conclusions

Phases of Clinical Trials

Classification of stages of investigation

- Gradual accumulation of experience in humans
  - Phase I: Initial safety / dose finding
  - Phase II: Preliminary efficacy / further safety
  - Phase III: Establishment of efficacy
  - Phase IV:
    - Therapeutics: Post-marketing surveillance
    - Prevention: Effectiveness
- Differing focus across phases leads to different choices for design of studies

Role of Statistical Inference

Role of Statistical Inference

A scientific study is conducted to answer some question

- Prediction of values
  - Single best estimate
  - Interval estimates
- Clustering of measurements across variables
- Relationships among variables
  - Distribution of measurements within groups
  - Comparison of distributions across groups
  - Interactions

Role of Statistical Inference

Why Statistics?

- Observations Subject to Error
  - In the real world, few patterns are deterministic
    - Hidden (unmeasured) variables
    - Inherent randomness

Goal is to use a sample to identify treatments that are truly beneficial

Problem is similar to that in diagnostic testing in patients

Role of Statistical Inference

Typically, a sample of data is obtained in order to try to answer the scientific question

- Sampling schemes
  - Observational studies
    - Cross-sectional
    - Cohort
    - Case-control
  - Interventions
- Time of observation
  - Single point in time
  - Longitudinal

Role of Statistical Inference

Descriptive statistics are computed for the sample

- Detection of errors
- Materials and methods
- Validity of assumptions for analysis
- Estimates of association, etc.
- Hypothesis generation

Role of Statistical Inference

Attempts are then made to use the sample to make inference about the entire population from which the sample was drawn

- Need to quantify the uncertainty in the estimates computed from the sample

- To what extent does the random variation inherent in sampling affect our ability to draw conclusions?

Role of Statistical Inference

In statistical inference, we are interested in finding optimal estimates of future observations or population parameters

- Single best estimate

- (We must define what we mean by “best”)

Role of Statistical Inference

In statistical inference, we are interested in putting bounds on the certainty with which we draw conclusions

Interval estimates for population parameters

Decisions about plausible values for population parameters

Hierarchy of Statistical Goals

Hierarchy of experimental goals

Determinism:

- What works?

Probability model:

- What works most often?

Bayesian statistics:

- What probably works most often?

Frequentist statistics:

- If it weren't likely to work most often, what is the probability that it would have worked now?

Hierarchy of Statistical Goals

Tradeoffs between Bayesian and frequentist approaches

Bayesian: A vague (subjective) answer to the right question

- (How could the Bayesian know my propensity to cheat?)

Frequentist: A precise (objective) answer to the wrong question

- (The frequentist would give the same answer even if it were impossible that I were a cheater)

Hierarchy of Statistical Goals

Tradeoffs between Bayesian and frequentist approaches (cont.)

In fact, there is no real reason to regard tradeoffs as necessary.

Both approaches contribute complementary information about the strength of statistical evidence.

It is valid to consider both measures.

Hierarchy of Statistical Goals

In light of the fact that <u>all</u> trial designs have both a Bayesian and a frequentist interpretation, it is incorrect to regard that either approach is statistically more efficient than the other

Any effort to sell Bayesian methods on the basis of their requiring smaller sample sizes is merely changing the standards of statistical evidence required for the trial

Similar changes to frequentist standards of evidence will also result in smaller sample sizes

Hierarchy of Statistical Goals

Tradeoffs between Bayesian and frequentist approaches (cont.)

Bayesian inference:

- How likely are the hypotheses to be true based on the observed data (and a presumed prior distribution)?

Frequentist inference:

- Are the data that we observed typical of the hypotheses?

Statistical Criteria for Evidence

At the end of the study use frequentist and/or Bayesian data analysis to provide

Decision for or against hypotheses

- Binary decision
- Quantification of strength of evidence

Estimate of the treatment effect

- Single best estimate
- Range of reasonable estimates

Ethical Issues

Ethical Issues

Conducted in human volunteers, the clinical trial must be ethical for participants on the trial

Individual ethics

- Minimize harm and maximize benefit for participants in clinical trial
- Avoid giving trial participants a harmful treatment
- Do not unnecessarily give trial participants a less effective treatment

Ethical Issues

The clinical trial must ethically address the needs of the greater population of potential recipients of the treatment

Group ethics

- Approve new beneficial treatments as rapidly as possible
- Avoid approving ineffective or (even worse) harmful treatments
- Do not unnecessarily delay the new treatment discovery process

Ethical Issues

Mechanisms for ensuring ethical treatment of study subjects

Before starting the study:

- Institutional review board (IRB)

During conduct of the study:

- Data safety monitoring board (DSMB)

After studies completed

- Regulatory agencies (e.g., FDA)

Ethical Issues

Institutional review board (Human subjects committee)

Membership

- Scientists, clinicians, ethicists, statisticians

Reviews

- Protocols
- Informed consent

IRB approval necessary before study can start

Ethical Issues

Data safety monitoring committee

Independent advisory committee which meets periodically to review

- Conduct of the study
- Interim analysis of study data
  - Safety and efficacy data
- Secular trends in clinical setting
  - Changes in diagnosis of disease
  - Changes in treatment of disease
  - Changes in treatment of adverse events

Ethical Issues

Data safety monitoring committee (cont.)

At periodic meetings, interim study results are reviewed and recommendations made to the sponsor

- Terminate the study early
- Modify the protocol
- Issue alerts to the investigators
- Modify study monitoring procedures
- Continue as planned

Ethical Issues

Data safety monitoring committee (cont.)

Membership: Usually 3 or 4 members independent of study sponsor and investigators

- Scientists, clinicians
  - Experts in disease
  - Experts in treatment
  - Experts in anticipated adverse events
- Statisticians
- Ethicists
- Patient advocates

Ethical Issues

Data safety monitoring committee (cont.)

Review of interim data

- DSMB is unblinded to treatment assignment
  - Interim analyses results kept confidential
- Recommendations for early termination are often guided by formal stopping rules
  - Recommendations are advisory to sponsor

Ethical Issues

Regulatory agencies

Grant approval to study investigational new drugs

Review progress of studies from phase I to phase III

Review all data from studies of new treatment before granting approval

Ethical Issues

Regulatory agencies (cont.)

Usually require 2 - 3 independent phase III studies

- Concurrent control group to assess efficacy and rates of common adverse experiences

Usually require experience treating some minimal number of patients in order to put upper bounds on rates of serious adverse experiences that went unobserved

- Rule of 3: If no events were observed in N patients, the upper 95% confidence bound is asymptotically 3 / N (4.6 / N for 99% bound)

Statistical Issues

Statistical Issues

Bottom Line

The wide variety of situations addressed by clinical trials demand a broad variety of study designs

In every case, however, it is of paramount importance that the clinical trial design be fully evaluated to ensure

- scientific credibility
- ethical experiments
- efficient experiments

Statistical Issues

Really Bottom Line

"You better think (think)
think about what you're
trying to do…"
- Aretha Franklin

Statistical Issues

Role of statistical software:

A variety of statistical operating characteristics should be considered in order to ensure that the clinical trial design appropriately addresses the scientific, clinical, and statistical issues.

Ethical and efficiency concerns often lead to sequential monitoring, which does not greatly affect which operating characteristics are to be examined, but does affect the computation of those operating characteristics.

Statistical Issues

Many measures used to quantify statistical evidence for treatment effect are based on the sampling density for a test statistic

- Design operating characteristics
  - Type I error, power
    - Sample size computation
- Statistical inference
  - P values
  - Confidence intervals
  - Some optimality properties of estimators:
    - bias
    - mean squared error

Statistical Issues

In fixed sample testing (no interim analyses), frequentist inference is most often obtained using test statistics that are normally distributed.

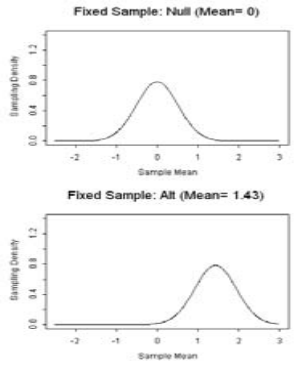
Hence, the sampling density must be numerically integrated to find some operating characteristics.

Due to properties of the normal distribution, it is feasible to table a standardized form.

The frequentist estimates, confidence intervals, and P values are then derived from the normal sampling distribution.

Example

Fixed sample (no interim analyses) sampling density



Statistical Issues

In monitoring a study, ethical considerations may demand that a study be stopped early.

The conditions under which a study might be stopped early constitutes a stopping rule

- At each analysis, the values that would cause a study to stop early are specified

The stopping boundaries might vary across analyses due to the imprecision of estimates

- At earlier analyses, estimates are based on smaller sample sizes and are thus less precise

Statistical Issues

The choice of stopping boundaries is typically governed by a wide variety of often competing goals.

The process for choosing a stopping rule is the substance of this course.

For the present, however, we consider only the basic framework for a stopping rule.

Statistical Issues

The stopping rule must account for ethical issues.

Early stopping might be based on

- Individual ethics
  - the observed statistic suggests efficacy
  - the observed statistic suggests harm
- Group ethics
  - the observed statistic suggests equivalence

Exact choice will vary according to scientific / clinical setting

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

(Example stopping rule taken from a two-sided symmetric design (Pampallona & Tsiatis, 1994) with a maximum of four analyses and O'Brien-Fleming (1979) boundary relationships)

Example

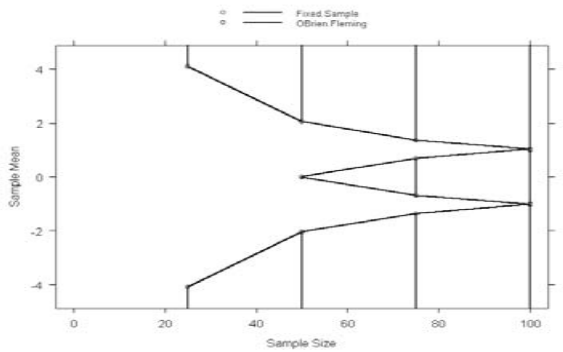
"O'Brien-Fleming" stopping rule

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

Example

"O'Brien-Fleming" stopping rule



Statistical Issues

In sequential testing (1 or more interim analyses), more specialized software is necessary.

The sampling density at each stage depends on continuation from previous stage

Recursive numerical integration of convolutions

The sampling density is not so simple: skewed, multimodal, with jump discontinuities

The treatment effect is no longer a shift parameter

Example

"O'Brien-Fleming" stopping rule

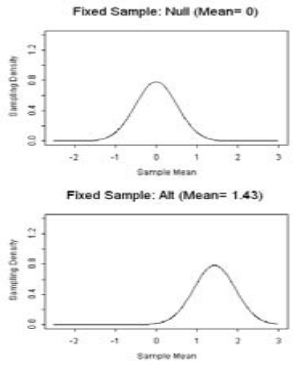
Possibility for early stopping introduces jump discontinuities at values corresponding to stopping boundaries

- Size of jump will depend upon true value of the treatment effect (mean)

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

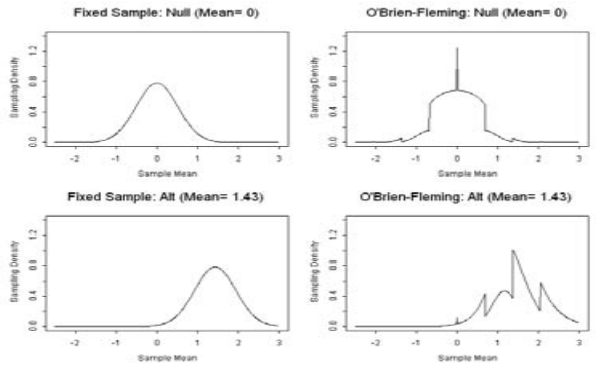
Example

Fixed sample (no interim analyses) sampling density



Example

Sampling density under stopping rule



Statistical Issues

Because the estimate of the treatment effect is no longer normally distributed in the presence of a stopping rule, the frequentist inference typically reported by statistical software is no longer valid

The standardization to a Z statistic does not produce a standard normal

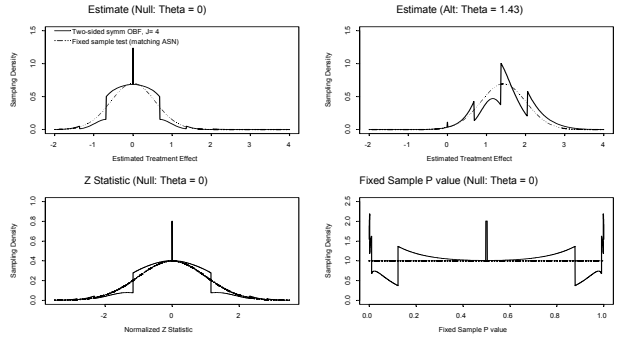
- The number 1.96 is now irrelevant

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

Sampling Densities for Z, Fixed P

Sampling densities for Z statistic, fixed sample P value in the presence of a stopping rule



Statistical Issues

Because a stopping rule changes the sampling distribution, the use of a stopping rule should change the computation of those design operating characteristics based on the 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

Example

Type I error: Null sampling density tails beyond crit value

Fixed sample test: Mean 0, variance 26.02, N 100

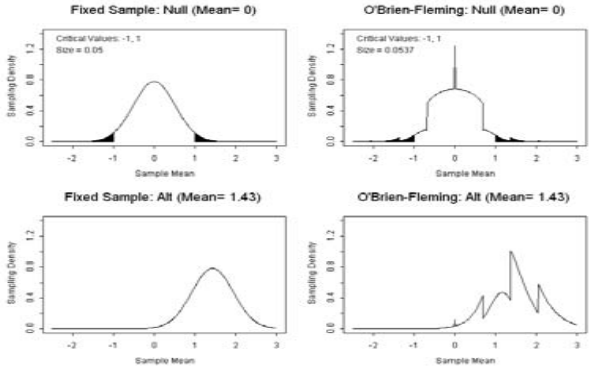
- Prob that sample mean is greater than 1 is 0.025
- Prob that sample mean is less than -1 is 0.025
- Two-sided type I error (size) is 0.05

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

Example

Type I error: Null sampling density tails beyond crit value



Statistical Issues

We can of course maintain the type I error when using a stopping rule by altering the critical value used to declare statistical significance

This only involves finding the correct quantiles of the true sampling density to use at the final analysis

Example

"O'Brien-Fleming" stopping rule

At each interim analysis, stop early if sample mean is indicated range

At the final analysis, the stopping must occur

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
100	< -1.023	(-1.023,1.023)	> 1.023

Example

"Pocock" stopping rule

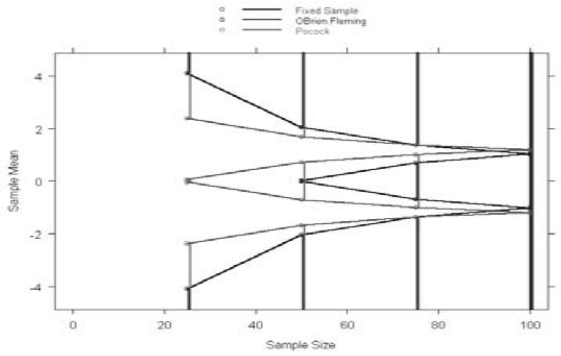
At each interim analysis, stop early if sample mean is indicated range

At the final analysis, the stopping must occur

N	Harm	Equiv	Efficacy
25	< -2.37	(-0.048,0.048)	> 2.37
50	< -1.68	(-0.715,0.715)	> 1.68
75	< -1.37	(-1.011,1.011)	> 1.37
100	< -1.187	(-1.187,1.187)	> 1.187

Example

"Pocock" vs "O'Brien-Fleming" stopping rules



Example

Power: Alternative sampling density tail beyond crit value

O'Brien-Fleming stopping rule: variance 26.02, max N 100

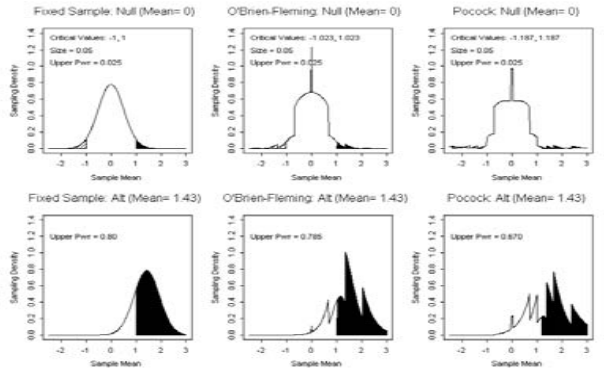
- Mean 0.00: Prob that sample mean > 1.023 is 0.025
- Mean 1.43: Prob that sample mean > 1.023 is 0.785
- Mean 2.00: Prob that sample mean > 1.023 is 0.970

Pocock stopping rule: variance 26.02, max N 100

- Mean 0.00: Prob that sample mean > 1.187 is 0.025
- Mean 1.43: Prob that sample mean > 1.187 is 0.670
- Mean 2.00: Prob that sample mean > 1.187 is 0.922

Example

Power: Alternative sampling density tail beyond crit value



Statistical Issues

The use of a stopping rule allows greater efficiency on average

Sample size requirements are a random variable

- Efficiency characterized by some summary of the sample size distribution
  - Average sample N (ASN)
  - Median, 75%ile of sample size distribution
  - Stopping probabilities at each analysis

Sample size distribution depends on true treatment effect

- (This was the goal of using a stopping rule)

Example

Sample size distribution for designs considered here

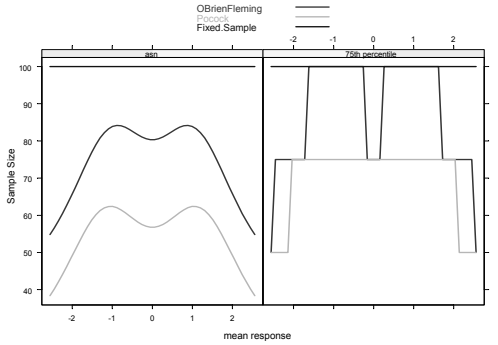
Fixed sample design requires 100 subjects no matter how effective (or harmful) the treatment is

O'Brien-Fleming stopping rule requires fewer subjects on average (worst case: about 84)

Pocock stopping rule requires even fewer subjects on average over a wide range of alternatives (worst case: about 62)

Example

Sample size distribution as a function of treatment effect



Example

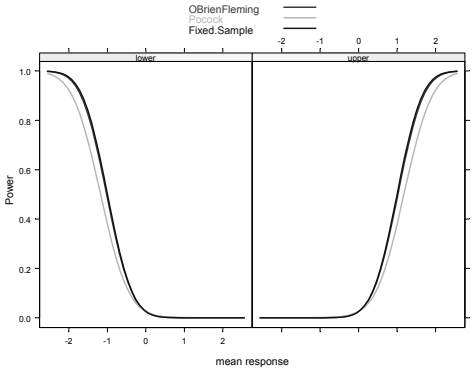
Failure to adjust the maximal sample size does affect the power of the clinical trial design

The introduction of the stopping rule will decrease the power of the design relative to a fixed sample design with the same maximal sample size

In the examples considered so far, we maintained the maximal sample size at 100 subjects

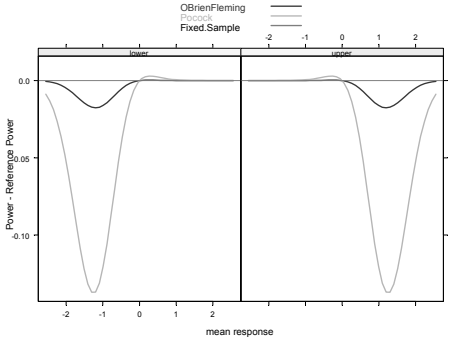
Example

Power as a function of treatment effect



Example

Power as a function of treatment effect relative to fixed sample design



Statistical Issues

We can maintain both the type I error and power when using a stopping rule by altering the critical value used to declare statistical significance and maximal sample size

This involves a search for the sample size that will provide the power.

Example

"O'Brien-Fleming" stopping rule with desired power

At each interim analysis, stop early if sample mean is indicated range

At the final analysis, the stopping must occur

N	Harm	Equiv	Efficacy
26	< -4.01	----	> 4.09
52	< -2.01	(-0.006,0.006)	> 2.01
78	< -1.34	(-0.670,0.670)	> 1.34
104	< -1.003	(-1.003,1.003)	> 1.023

Example

"Pocock" stopping rule with desired power

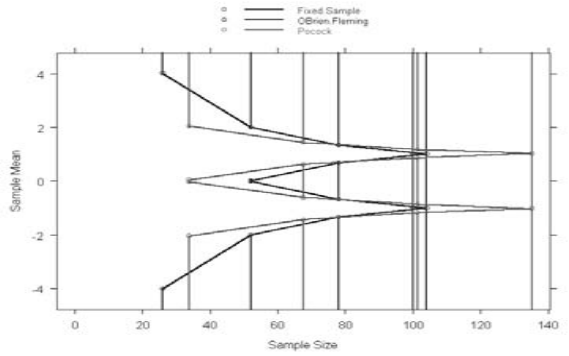
At each interim analysis, stop early if sample mean is indicated range

At the final analysis, the stopping must occur

N	Harm	Equiv	Efficacy
34	< -2.04	(-0.042,0.042)	> 2.04
68	< -1.44	(-0.615,0.615)	> 1.44
101	< -1.18	(-0.869,0.869)	> 1.18
135	< -1.021	(-1.021,1.021)	> 1.021

Example

"Pocock", "O'Brien-Fleming" with desired power



Example

Power: Alternative sampling density tail beyond crit value

O'Brien-Fleming stopping rule: variance 26.02, max N 104

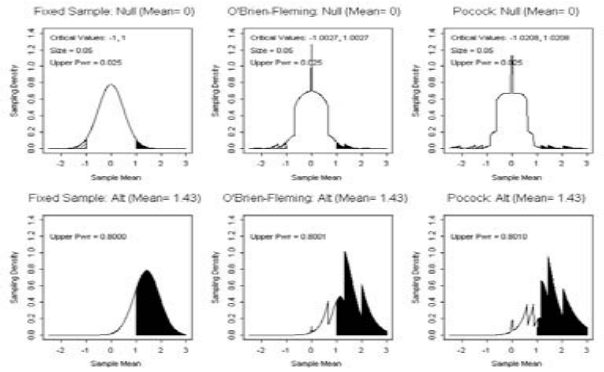
- Mean 0.00: Prob that sample mean > 1.003 is 0.025
- Mean 1.43: Prob that sample mean > 1.003 is 0.8001
- Mean 2.00: Prob that sample mean > 1.003 is 0.975

Pocock stopping rule: variance 26.02, max N 135

- Mean 0.00: Prob that sample mean > 1.021 is 0.025
- Mean 1.43: Prob that sample mean > 1.021 is 0.801
- Mean 2.00: Prob that sample mean > 1.021 is 0.975

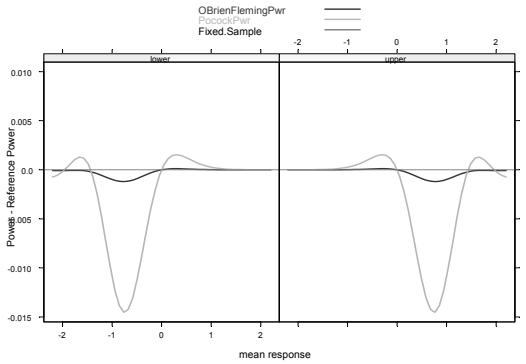
Example

Power: Alternative sampling density tail beyond crit value



Example

Power curves relative to fixed sample design



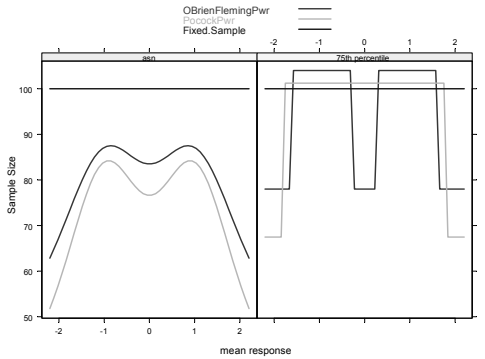
Example

The increased <u>maximal</u> sample size need not mean a less efficient design when using a stopping rule

- Fixed sample design requires 100 subjects no matter how effective (or harmful) the treatment is
- O'Brien-Fleming stopping rule requires fewer subjects on average (worst case: about 88) and the increase in the maximal sample size is only 4%
- Pocock stopping rule requires even fewer subjects on average over a wide range of alternatives, but requires a 35% increase in the maximal sample size
  - However, there is always less than a 25% chance that a trial would continue to the last analysis

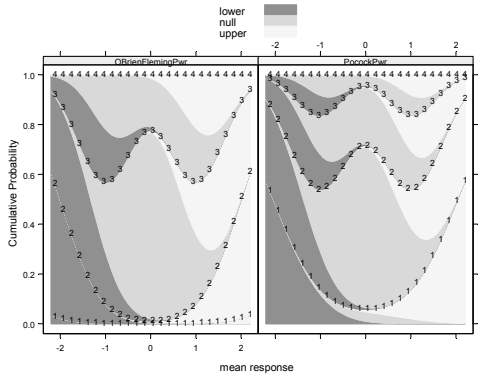
Example

Sample size distribution as a function of treatment effect



Example

Stopping probabilities as a function of treatment effect



Statistical Issues

In this course

Focus on study designs appropriate for phase II and phase III clinical trials

Focus on statistical design issues especially as they relate to the design, monitoring, and analysis of the clinical trials

Emphasize the choice of statistical designs to address scientific questions

S+SeqTrial

Selection of clinical trial design is iterative, involving scientists, statisticians, management, and regulators

Encourage use of measures with scientific meaning

Facilitate search through extensive space of designs

Facilitate comparison of designs with respect to variety of operating characteristics

Seamless progression from design to monitoring to analysis

S+SeqTrial

Interface with more routine analysis methods

Sequential aspects only part of clinical trial needs

Design

- might also want to consider effects of drop-in, drop-out, compliance, missing data, etc.

Analysis

- Descriptive statistics, graphics
- Statistical analysis
- Models adjusting for covariates