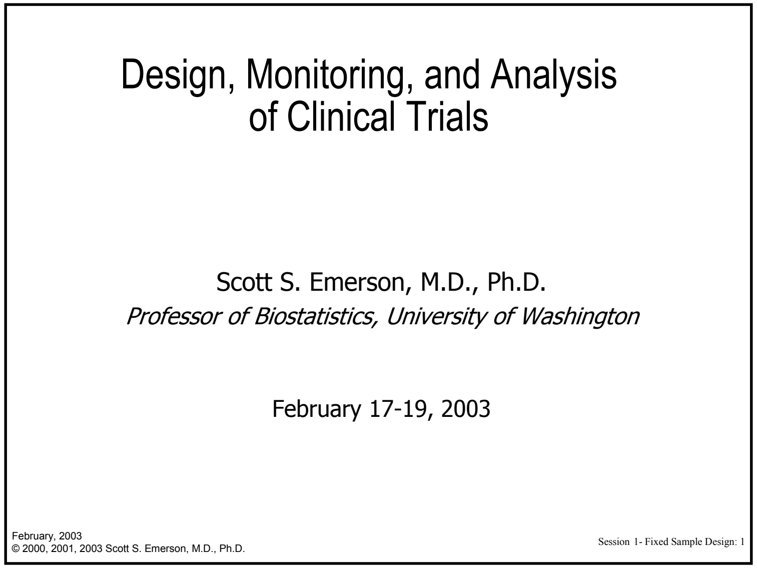
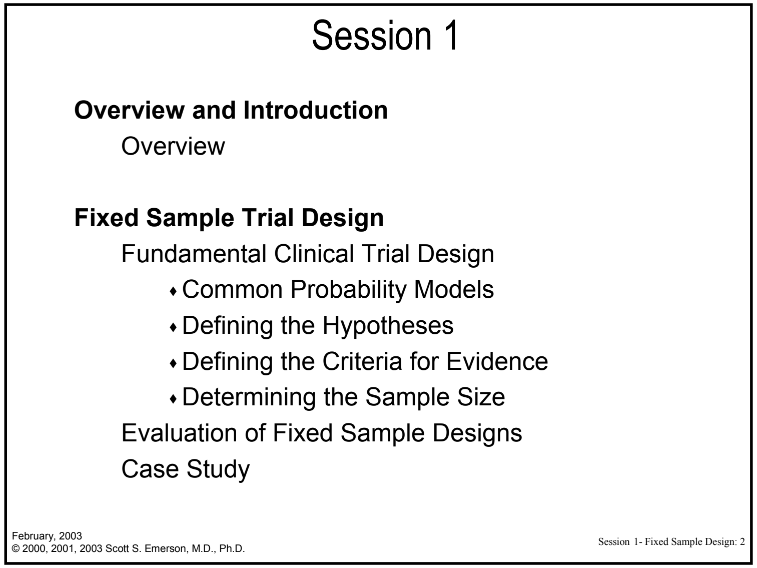


Design, Monitoring, and Analysis of Clinical Trials

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

February 17-19, 2003

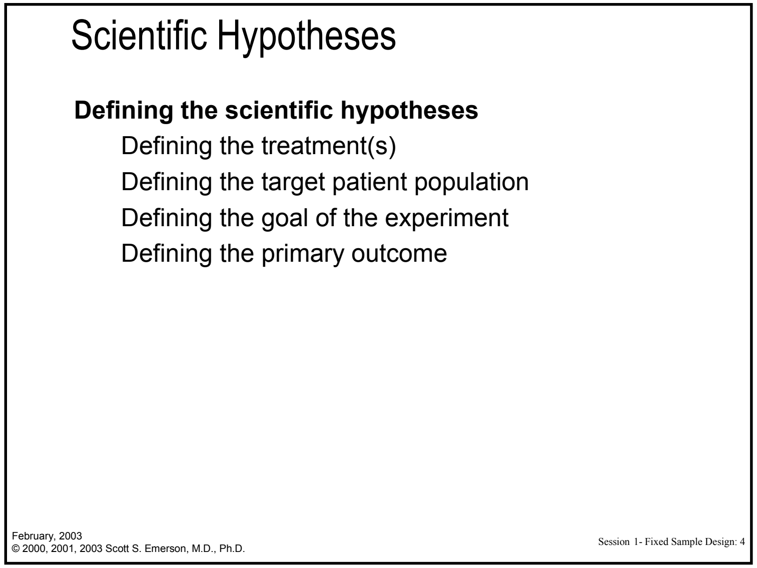


Session 1

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

Fundamental Clinical Trial Design



Scientific Hypotheses

Defining the scientific hypotheses
- Defining the treatment(s)
- Defining the target patient population
- Defining the goal of the experiment
- Defining the primary outcome

Scientific Hypotheses: Treatment

Defining the treatment(s)

The “treatment” that we test in a clinical trial must be completely defined at the time of randomization

- dose(s)
- administration(s)
- frequency and duration of treatment
- ancillary treatments and treatment reduction

Relevance to “intent-to-treat” analyses

Scientific Hypotheses: Patient Population

Defining the target patient population

Inclusion/exclusion criteria to identify a population for whom

- A new treatment is needed
- Experimental treatment is likely to work
  - and to work equally well in all subgroups
- All patients likely to eventually use the new treatment are represented
  - Safety issues
- Clinical experimentation with the new treatment is not unethical

Scientific Hypotheses: Experimental Goal

Defining the goal of experiment

Common scenarios for clinical trials

- Superiority
- Noninferiority
- Equivalence
- Nonsuperiority
- Inferiority

Scientific Hypotheses: Experimental Goal

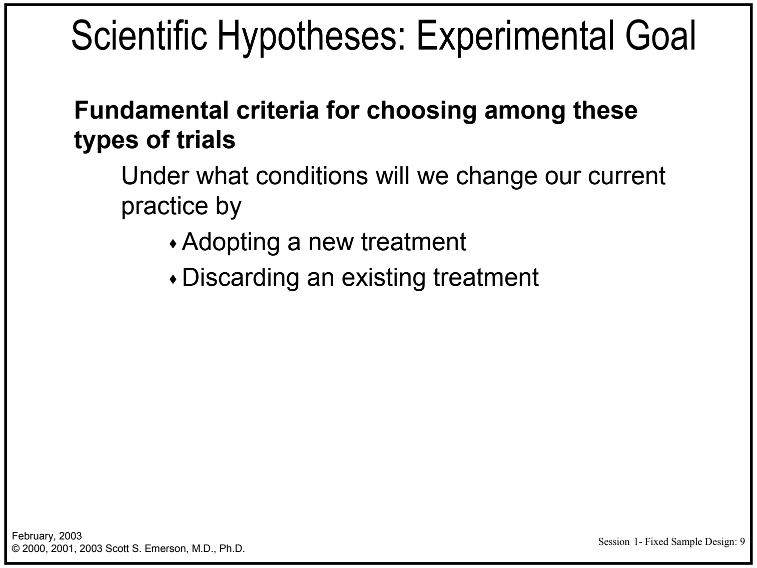
We will describe clinical trial designs that discriminate between several of these hypotheses

One sided hypothesis tests

Two sided hypothesis tests

Two sided equivalence tests (e.g., bioequivalence)

One-sided equivalence (noniferiority) tests

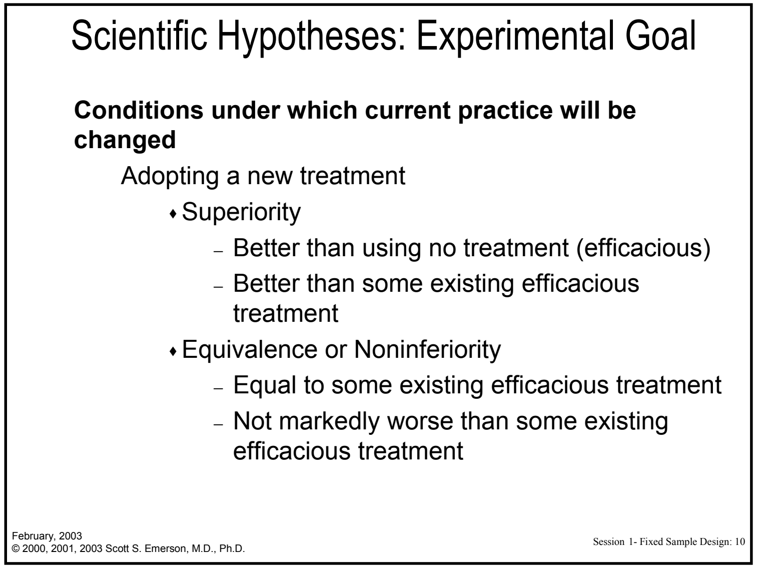


Scientific Hypotheses: Experimental Goal

Fundamental criteria for choosing among these types of trials

Under what conditions will we change our current practice by

- ♦ Adopting a new treatment
- ♦ Discarding an existing treatment

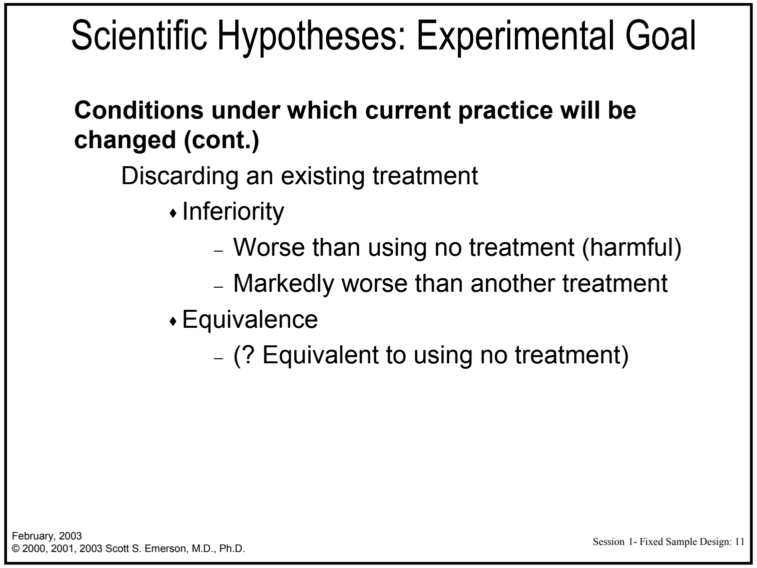


Scientific Hypotheses: Experimental Goal

Conditions under which current practice will be changed

Adopting a new treatment

- ♦ Superiority
  - – Better than using no treatment (efficacious)
  - – Better than some existing efficacious treatment
- ♦ Equivalence or Noninferiority
  - – Equal to some existing efficacious treatment
  - – Not markedly worse than some existing efficacious treatment

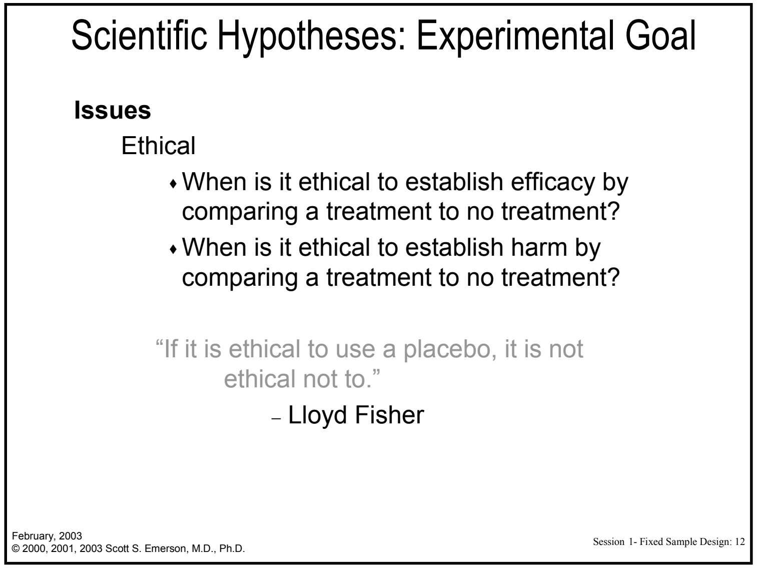


Scientific Hypotheses: Experimental Goal

Conditions under which current practice will be changed (cont.)

Discarding an existing treatment

- ♦ Inferiority
  - – Worse than using no treatment (harmful)
  - – Markedly worse than another treatment
- ♦ Equivalence
  - – (? Equivalent to using no treatment)



Scientific Hypotheses: Experimental Goal

Issues

Ethical

- ♦ When is it ethical to establish efficacy by comparing a treatment to no treatment?
- ♦ When is it ethical to establish harm by comparing a treatment to no treatment?

> “If it is ethical to use a placebo, it is not ethical not to.”
>
> – Lloyd Fisher

Scientific Hypotheses: Experimental Goal

Issues (cont.)

Scientific

- How to define scientific hypotheses when trying to establish
  - efficacy by comparing a new treatment to no treatment
  - efficacy by comparing a new treatment to an existing efficacious treatment
  - superiority of one treatment over another
- How to choose the comparison group when trying to establish efficacy by comparing a new treatment to an existing efficacious treatment

Scientific Hypotheses: Primary Endpoint

Scientific basis

A clinical trial is planned to detect the effect of a treatment on some outcome

- Safety
- Efficacy
- Effectiveness

Statement of the outcome is a fundamental part of the scientific hypothesis

Scientific Hypotheses: Primary Endpoint

Ethical basis

Generally, subjects participating in a clinical trial are hoping that they will benefit in some way from the trial

Clinical endpoints are therefore of more interest than purely biological endpoints

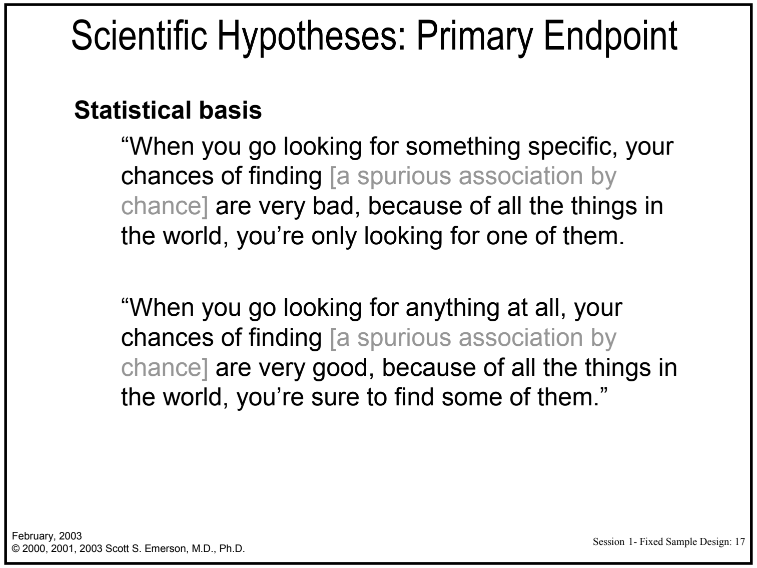
Scientific Hypotheses: Primary Endpoint

Statistical basis

"When you go looking for something specific, your chances of finding it are very bad, because of all the things in the world, you're only looking for one of them.

"When you go looking for anything at all, your chances of finding it are very good, because of all the things in the world, you're sure to find some of them."

- Darryl Zero in "The Zero Effect"

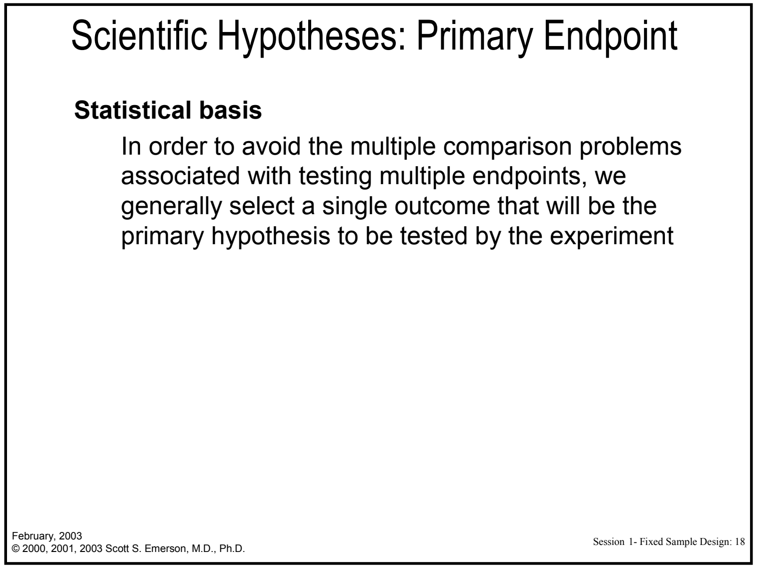


Scientific Hypotheses: Primary Endpoint

Statistical basis

“When you go looking for something specific, your chances of finding [a spurious association by chance] are very bad, because of all the things in the world, you’re only looking for one of them.

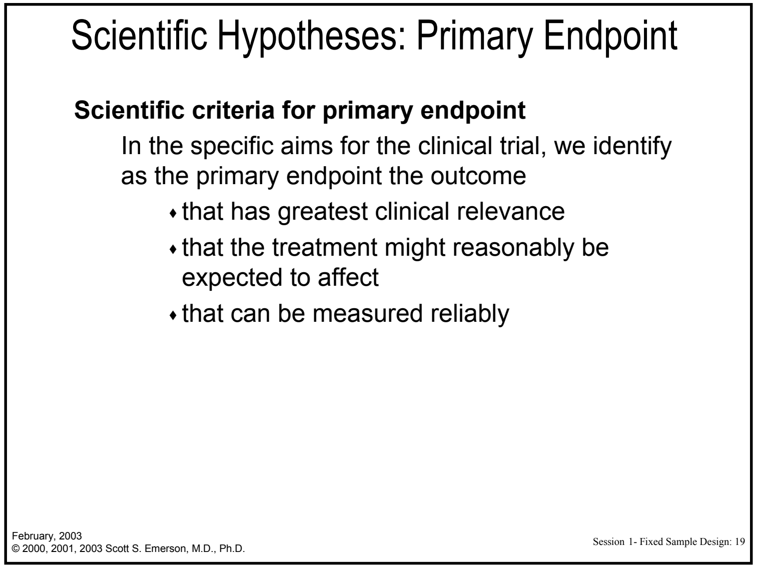
“When you go looking for anything at all, your chances of finding [a spurious association by chance] are very good, because of all the things in the world, you’re sure to find some of them.”



Scientific Hypotheses: Primary Endpoint

Statistical basis

In order to avoid the multiple comparison problems associated with testing multiple endpoints, we generally select a single outcome that will be the primary hypothesis to be tested by the experiment

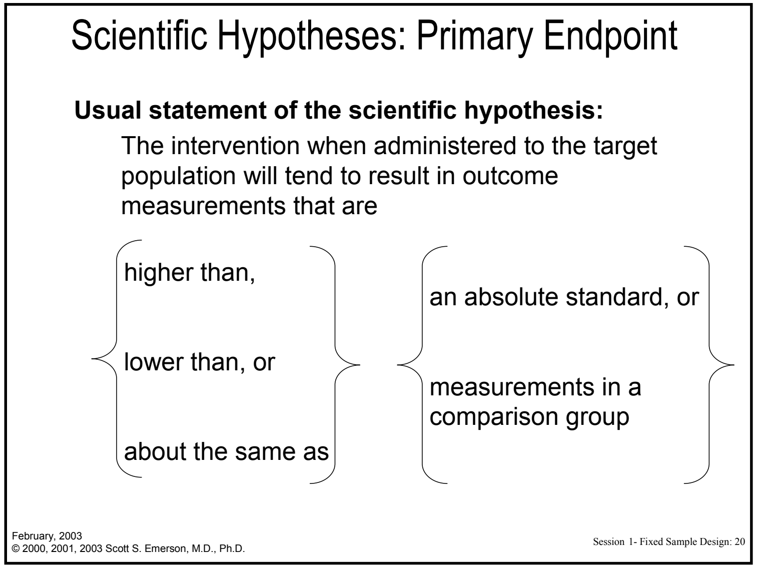


Scientific Hypotheses: Primary Endpoint

Scientific criteria for primary endpoint

In the specific aims for the clinical trial, we identify as the primary endpoint the outcome

- that has greatest clinical relevance
- that the treatment might reasonably be expected to affect
- that can be measured reliably



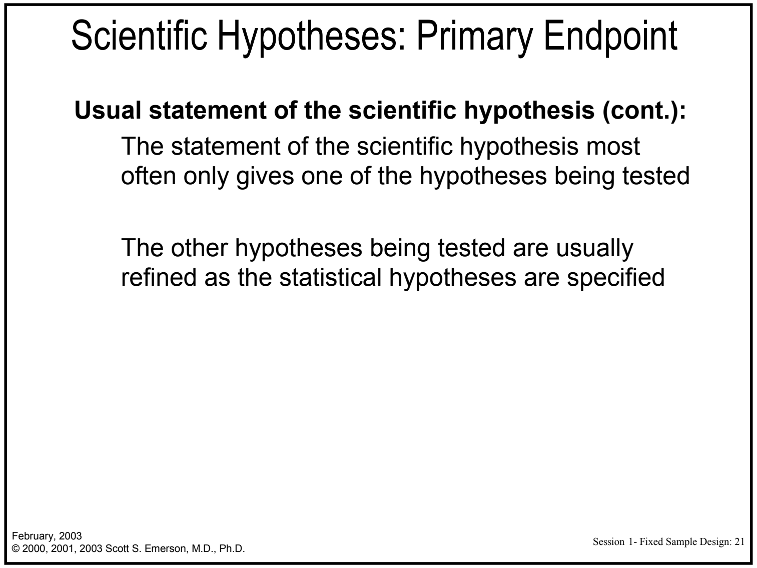
Scientific Hypotheses: Primary Endpoint

Usual statement of the scientific hypothesis:

The intervention when administered to the target population will tend to result in outcome measurements that are

{ higher than,
lower than, or
about the same as }

{ an absolute standard, or
measurements in a comparison group }



Scientific Hypotheses: Primary Endpoint

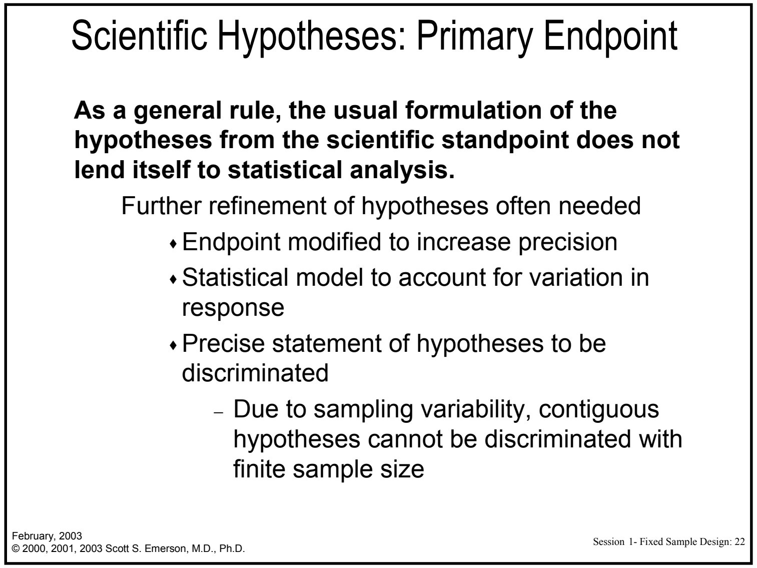
Usual statement of the scientific hypothesis (cont.):

The statement of the scientific hypothesis most often only gives one of the hypotheses being tested

The other hypotheses being tested are usually refined as the statistical hypotheses are specified

February, 2003

Session 1- Fixed Sample Design: 21



Scientific Hypotheses: Primary Endpoint

As a general rule, the usual formulation of the hypotheses from the scientific standpoint does not lend itself to statistical analysis.

Further refinement of hypotheses often needed

- ♦ Endpoint modified to increase precision
- ♦ Statistical model to account for variation in response
- ♦ Precise statement of hypotheses to be discriminated
  - – Due to sampling variability, contiguous hypotheses cannot be discriminated with finite sample size

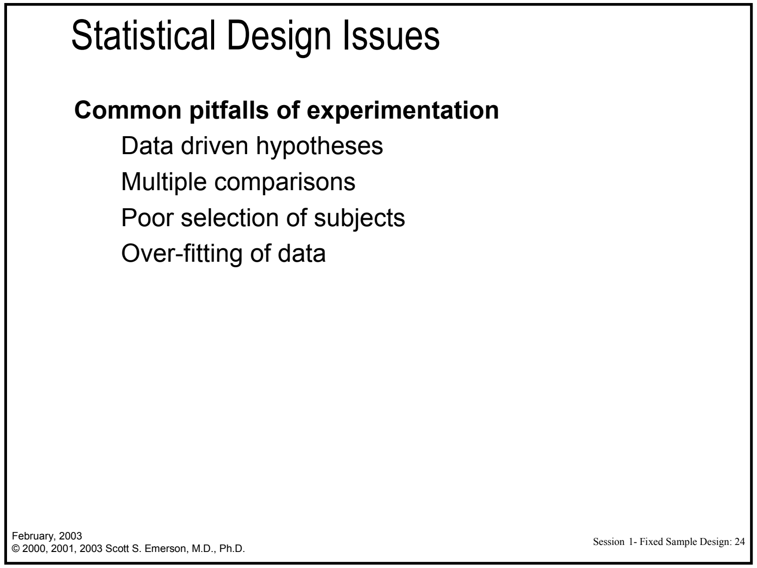
February, 2003

Session 1- Fixed Sample Design: 22

Statistical Design of Clinical Trials

February, 2003

Session 1- Fixed Sample Design: 23



Statistical Design Issues

Common pitfalls of experimentation

Data driven hypotheses

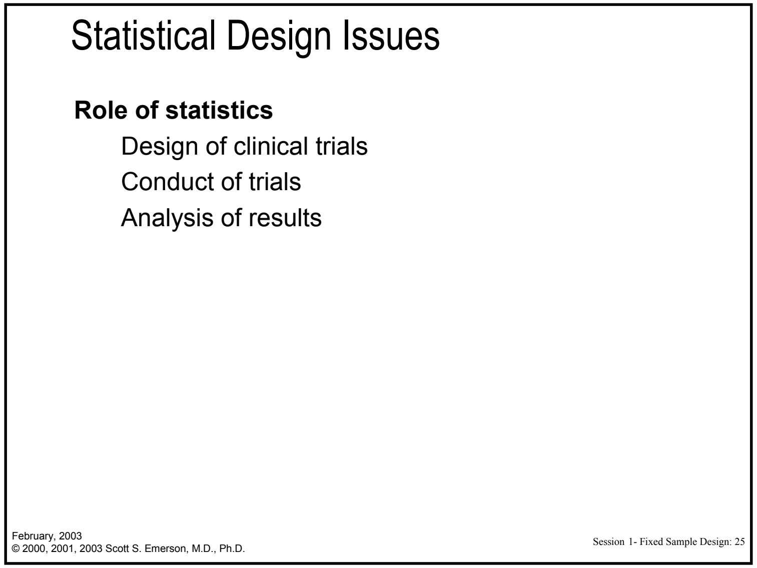
Multiple comparisons

Poor selection of subjects

Over-fitting of data

February, 2003

Session 1- Fixed Sample Design: 24



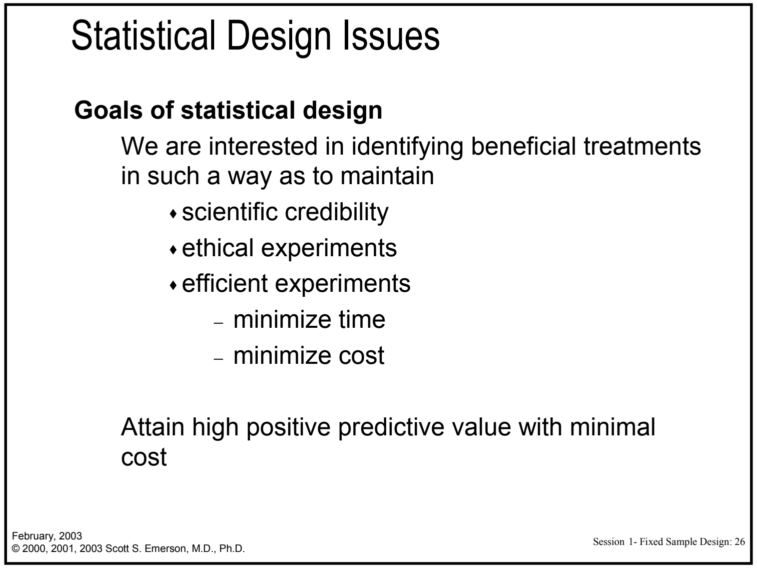
Statistical Design Issues

Role of statistics

Design of clinical trials

Conduct of trials

Analysis of results



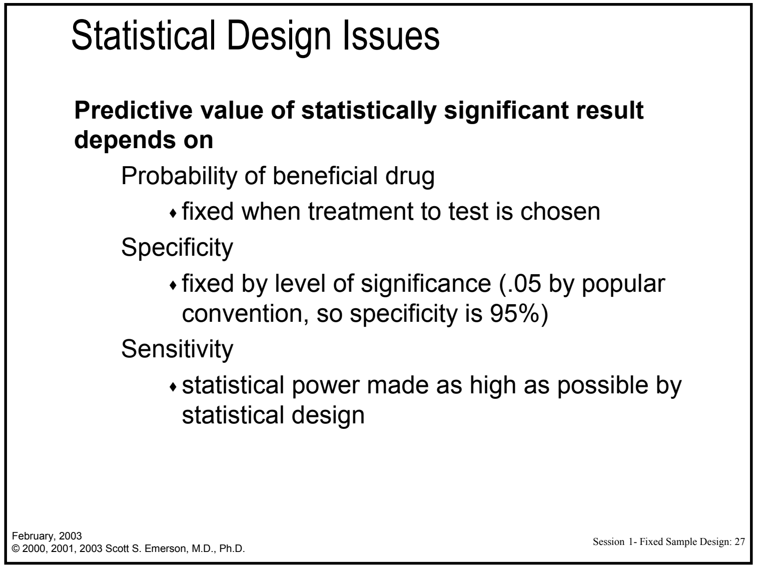
Statistical Design Issues

Goals of statistical design

We are interested in identifying beneficial treatments in such a way as to maintain

- scientific credibility
- ethical experiments
- efficient experiments
  - minimize time
  - minimize cost

Attain high positive predictive value with minimal cost



Statistical Design Issues

Predictive value of statistically significant result depends on

Probability of beneficial drug

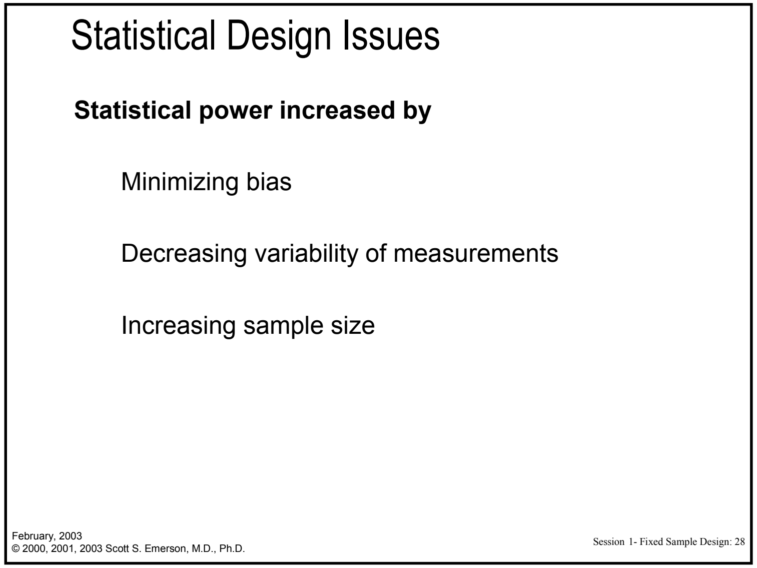
- fixed when treatment to test is chosen

Specificity

- fixed by level of significance (.05 by popular convention, so specificity is 95%)

Sensitivity

- statistical power made as high as possible by statistical design



Statistical Design Issues

Statistical power increased by

Minimizing bias

Decreasing variability of measurements

Increasing sample size

Statistical Design Issues

Bias is minimized by

Removing confounding by other risk factors

Addressing issues of effect modification

Removing ascertainment bias, etc.

Statistical Design Issues

Variability of measurements decreased by

Homogeneity of patient population

Precise definition of treatment(s)

Appropriate choice of endpoints

High precision in measurements

Appropriate sampling strategy

Statistical Design

Statistical Design Issues

Defining the probability model

- ♦ Defining the comparison group
- ♦ Refining the primary endpoint
- ♦ Defining the method of analysis

Defining the statistical hypotheses

Statistical Design

Statistical Design Issues (cont.)

Defining the statistical criteria for evidence

Determining the sample size

Evaluating the operating characteristics

Planning for monitoring

Plans for analysis and reporting results

Defining the Probability Model

Number of Arms

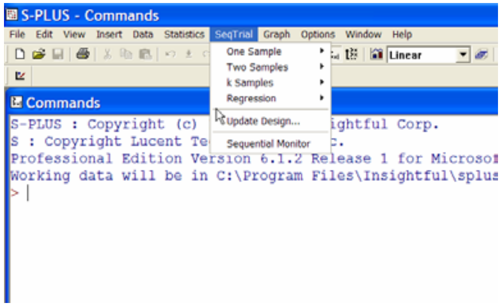
Defining the comparison groups

- One arm trial: Historical or matched controls
  - comparison to absolute reference
- Two arm trial: Placebo or Active controls
  - comparison between arms
- Multiple arm trial: Several treatments or doses
  - global tests vs pairwise comparisons
- Regression trial: Continuous dose response
  - test slope parameters

Also need to consider randomization ratio across treatment arms (or distribution of dose levels)

S+SeqTrial: Choosing the Number of Arms

Select the number of arms from the pull down SeqTrial menu



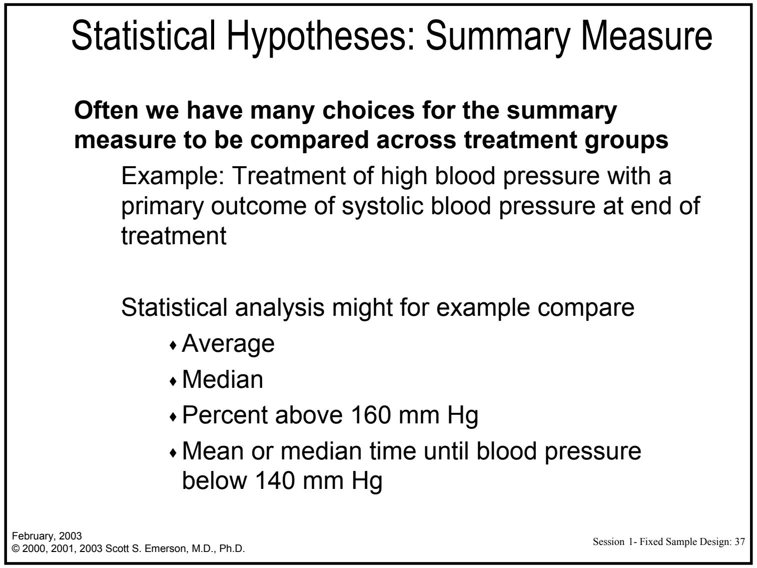
```
S-PLUS : Copyright (c)                 ightful Corp.
S : Copyright Lucent Te                C.
Professional Edition Version 6.1.2 Release 1 for Microso
Working data will be in C:\Program Files\Insightful\splus
> |
```

Probability Model: Summary Measure

Summarize tendencies of response using a summary measure

- Determine the tendency for a new treatment to have a beneficial effect on a clinical outcome

- Consider the distribution of outcomes for individuals receiving intervention
  - Usually choose a summary measure of the distribution
    - e.g.. mean, median, proportion cured, etc.
  - Hypotheses then expressed for values of summary measure



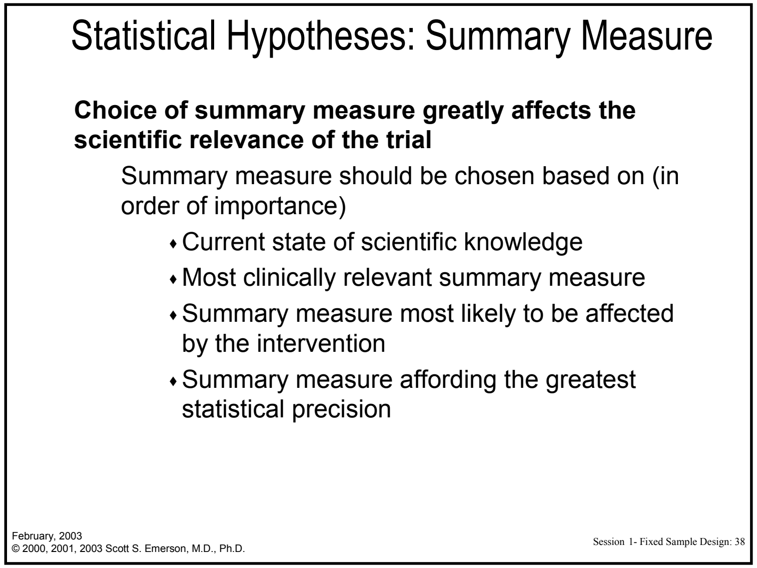
Statistical Hypotheses: Summary Measure

Often we have many choices for the summary measure to be compared across treatment groups

Example: Treatment of high blood pressure with a primary outcome of systolic blood pressure at end of treatment

Statistical analysis might for example compare

- Average
- Median
- Percent above 160 mm Hg
- Mean or median time until blood pressure below 140 mm Hg

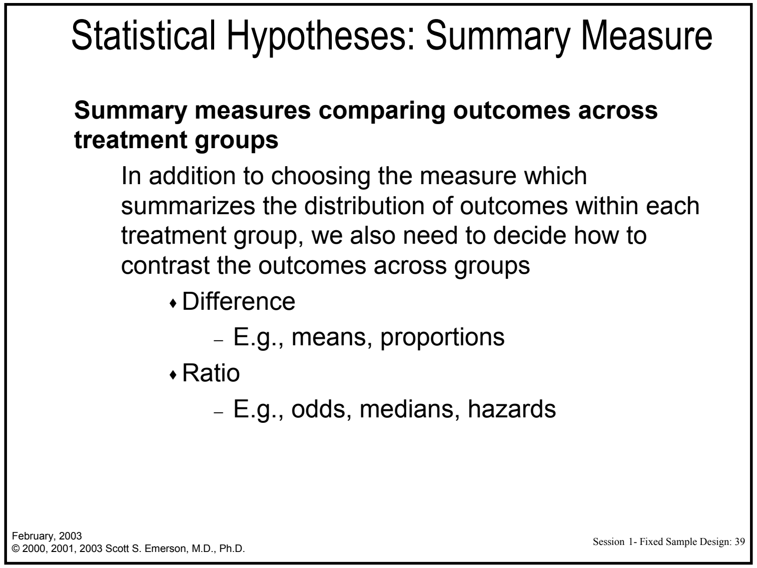


Statistical Hypotheses: Summary Measure

Choice of summary measure greatly affects the scientific relevance of the trial

Summary measure should be chosen based on (in order of importance)

- Current state of scientific knowledge
- Most clinically relevant summary measure
- Summary measure most likely to be affected by the intervention
- Summary measure affording the greatest statistical precision

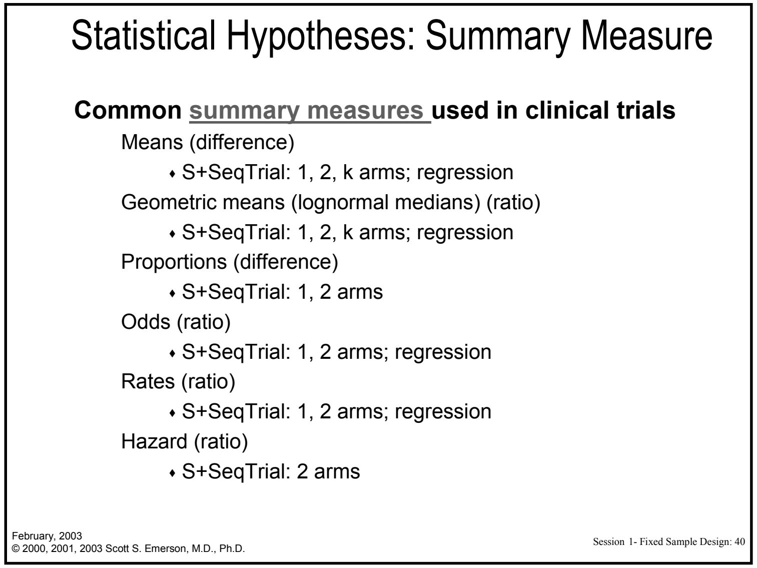


Statistical Hypotheses: Summary Measure

Summary measures comparing outcomes across treatment groups

In addition to choosing the measure which summarizes the distribution of outcomes within each treatment group, we also need to decide how to contrast the outcomes across groups

- Difference
  - E.g., means, proportions
- Ratio
  - E.g., odds, medians, hazards



Statistical Hypotheses: Summary Measure

Common summary measures used in clinical trials

Means (difference)

- S+SeqTrial: 1, 2, k arms; regression

Geometric means (lognormal medians) (ratio)

- S+SeqTrial: 1, 2, k arms; regression

Proportions (difference)

- S+SeqTrial: 1, 2 arms

Odds (ratio)

- S+SeqTrial: 1, 2 arms; regression

Rates (ratio)

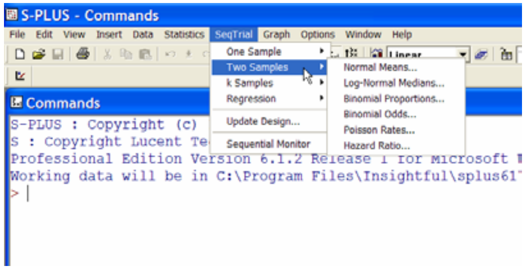
- S+SeqTrial: 1, 2 arms; regression

Hazard (ratio)

- S+SeqTrial: 2 arms

S+SeqTrial: Choosing the Summary Measure

Select the probability model from the pull down SeqTrial menu



```
S-PLUS : Copyright (c)
S : Copyright Lucent Te
Professional Edition Version 6.1.2 Release 1 for Microsoft 
Working data will be in C:\Program Files\Insightful\splus61
> |
```

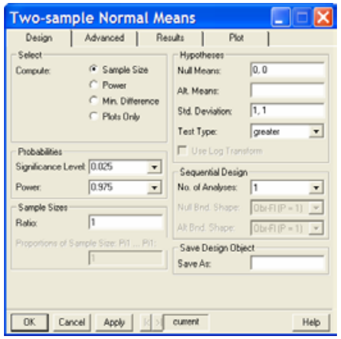
S+SeqTrial: Launching the Dialog

Upon selection of the probability model, a dialog is launched to allow entry of

- Computational task
- Hypotheses (form is specific to probability model)
- Error probabilities
- Number of interim analyses and boundary shapes
- Sample size, randomization ratio, timing of analyses
- Name of saved object
- Tabs for
  - Advanced group sequential design
  - Report options
  - Plotting options

S+SeqTrial: Example of a Dialog

Dialog for two sample comparison of (normal) means



Compute: ● Sample Size ○ Power ○ Min. Difference ○ Plots Only

Null Means: 0, 0

Alt. Means:

Std. Deviation: 1, 1

Test Type: greater

□ Use Log Transform

Significance Level: 0.025

Power: 0.975

Ratio: 1

Proportions of Sample Size: Pi1 ... Pi1: 1

No. of Analyses: 1

Null Bnd. Shape: Obr-Fl (P = 1)

Alt Bnd. Shape: Obr-Fl (P = 1)

Save As:

Defining Statistical Hypotheses

Statistical Hypotheses

Problem: The distribution (or summary measure) for the outcome is not directly observable

Use a sample to estimate the distribution (or summary measure) of outcomes

Such an estimate is thus subject to sampling error

- Quantify precision of estimates
- Make decisions about whether population summary measure is in some range of values
  - Hypotheses

Statistical Hypotheses: Definition

We would like to have high precision as we discriminate between hypotheses

Scientifically, we are interested in performing experiments to decide which of two (or more) hypotheses might be true

Discrimination with high precision suggests that if we are highly confident that one hypothesis is true, we are equally confident that all other, mutually exclusive hypotheses are false

Statistical Hypotheses: Definition

Ideally we might like to discriminate between two contiguous hypotheses with high precision

Example of contiguous hypotheses

- The treatment tends to decrease blood pressure, or
- The treatment tends to increase blood pressure or leave it unchanged

Statistical Hypotheses: Definition

In the presence of sampling error, we must be careful in how we define hypotheses

In the presence of sampling error, an infinite sample size is required to be able to address both of those contiguous hypotheses with high precision simultaneously

Possible solutions

- Treat hypotheses asymmetrically
- Define experiment in terms of noncontiguous hypotheses

Classical Hypothesis Testing

Classical Hypothesis Testing

Classical hypothesis testing usually defines null and alternative hypotheses for some summary measure of probability model

Null Hypothesis: Represents the status quo

- Classically, the decision in the absence of evidence to the contrary

Alternative Hypothesis: What we hope is true

- Classically, the decision in the presence of evidence against the null
- Usually defined as a contiguous hypothesis

Classical Hypothesis Testing

Classical One-sided Test (Greater Alternative)

Let

- θ (a parameter) summarize treatment effect
- T (a statistic) tend to be large for larger θ
- c_U be a critical value chosen by some criterion

Null Hypothesis: $H_0: \theta = \theta_0$
Alternative Hypothesis: $H_1: \theta > \theta_0$

Reject H_0 $\Longleftrightarrow$ $T \geq c_U$
Do not reject H_0 $\Longleftrightarrow$ $T < c_U$

Classical Hypothesis Testing

Classical One-sided Test (Lesser Alternative)

Let

- θ (a parameter) summarize treatment effect
- T (a statistic) tend to be large for larger θ
- c_L be a critical value chosen by some criterion

Null Hypothesis: $H_0: \theta = \theta_0$
Alternative Hypothesis: $H_1: \theta < \theta_0$

Reject H_0 $\Longleftrightarrow$ $T \leq c_L$
Do not reject H_0 $\Longleftrightarrow$ $T > c_L$

Classical Hypothesis Testing

Classical Two-sided Test

Let

- θ (a parameter) summarize treatment effect
- T (a statistic) tend to be large for larger θ
- c_L, c_U critical values chosen by some criterion

Null Hypothesis: H_0: $\theta = \theta_0$

Alternative Hypothesis: H_1: $\theta \neq \theta_0$

Reject H_0 $\Longleftrightarrow$ $T \leq c_L$ or $T \geq c_U$

Do not reject H_0 $\Longleftrightarrow$ $c_L < T < c_U$

Decision Theoretic Approach

Decision Theoretic Approach

Decision Theoretic Issues: Asymmetry in treatment of null and alternative when using classical frequentist hypothesis testing

How to interpret a failure to reject the null?

- Would like to distinguish between
  - inadequate precision (sample size)
  - evidence against the alternative

Solution: Noncontiguous hypotheses

- Statistical design such that all but one hypothesis rejected with high confidence

Decision Theoretic Approach

The statistical hypotheses can be defined by either the hypotheses rejected or the hypotheses accepted

Classical hypothesis testing is described in terms of the hypotheses rejected

- In finite sample sizes, these hypotheses must not be contiguous if we maintain a fixed level of statistical evidence (see later)

If we describe the hypotheses being accepted, those hypotheses will overlap

Decision Theoretic Approach: Hypotheses

Statistical Hypotheses defined for summary measure of probability model

Null Hypothesis: Usually still the status quo

Alternative Hypothesis: Usually still what we hope is true

- Stated in terms of the minimal difference that it is important to detect
  - (May account for attenuation due to dropout, delayed effect, etc.)

In the decision theoretic framework, we want to be certain that we will reject either the null or the alternative at the end of the study

Decision Theoretic Approach: Tests

We can describe clinical trial designs that discriminate between two or more hypotheses

One sided hypothesis tests

- Tests for superiority
- Tests for inferiority
- Shifted tests for noninferiority (one-sided equivalence)

Two sided hypothesis tests

- Tests for superiority/inferiority
- Two sided equivalence tests (e.g., bioequivalence)

Defining the Hypotheses

Decision Theoretic Framework: One-sided tests

(Assume large θ corresponds to benefit)

Test of a greater alternative ($\theta_+ > \theta_0$)

- Null: H_0: $\theta \leq \theta_0$ (equivalent or inferior)
- Alternative: H_1: $\theta \geq \theta_+$ (sufficiently superior)

Decisions:

- Reject $H_0 \Longleftrightarrow T \geq c_U$ (superior)
- Reject $H_1 \Longleftrightarrow T \leq c_U$ (not sufficiently superior)

Defining the Hypotheses

Decision Theoretic Framework: One-sided tests

(Assume large θ corresponds to benefit)

Test of a lesser alternative ($\theta_- < \theta_0$)

- Null: H_0: $\theta \geq \theta_0$ (equivalent or superior)
- Alternative: H_1: $\theta \leq \theta_-$ (dangerously inferior)

Decisions:

- Reject $H_0 \Longleftrightarrow T \leq c_L$ (inferior)
- Reject $H_1 \Longleftrightarrow T \geq c_L$ (not dangerously inferior)

Defining the Hypotheses

Decision Theoretic Framework: Testing one-sided equivalence (noninferiority)

(Assume large θ corresponds to benefit)

A shifted one-sided test ($\theta_{0-} < \theta_{0+}$)

- Unacceptably inferior: $H_{0-}: \theta \leq \theta_{0-}$
- Reference for futility: $H_{0+}: \theta \geq \theta_{0+}$

Decisions:

- Reject $H_{0+} \Longleftrightarrow T \leq c_L$ (not worth continuing)
- Reject $H_{0-} \Longleftrightarrow T \geq c_L$ (not unacceptably inferior)

Defining the Hypotheses

Decision Theoretic Framework: Two-sided tests

(Assume large θ corresponds to benefit)

Test of a two-sided alternative ($\theta_+ > \theta_0 > \theta_-$)

- Upper Alt: $H_+: \theta \geq \theta_+$ (sufficiently superior)
- Null: $H_0: \theta = \theta_0$ (equivalent)
- Lower Alt: $H_-: \theta \leq \theta_-$ (dangerously inferior)

Decisions:

- Reject H_0, $H_- \Longleftrightarrow T \geq c_U$ (superior)
- Reject H_+, $H_- \Longleftrightarrow c_L \leq T \leq c_U$ (approx equiv)
- Reject H_+, $H_0 \Longleftrightarrow T \leq c_L$ (inferior)

S+SeqTrial: Specifying Hypotheses

Hypotheses groupbox varies with probability model

General principles

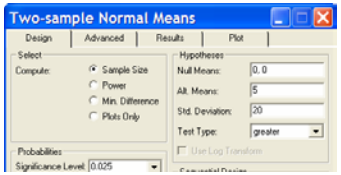
- Specify summary measures for each arm
  - Treatment, comparison
    - (Sometimes may log transform)
  - Can specify just the value for treatment arm
    - Default for the comparison group is the null value for treatment group
- Specify variance
- Specify type of test
  - Greater, less, two-sided, (one-sided) equivalence

S+SeqTrial: Specifying Hypotheses

Example: Two arm comparison of means

One-sided test of a greater hypothesis

- Null: Mean of 0 in each group
- Alt: Mean of 5 in treatment, 0 in comparison
- Standard deviation: 20 in each group



Two-sample Normal Means

Compute: ● Sample Size ○ Power ○ Min. Difference ○ Plots Only

Null Means: 0, 0

Alt. Means: 5

Std. Deviation: 20

Test Type: greater

□ Use Log Transform

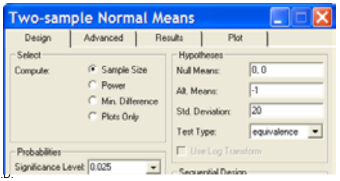
Significance Level: 0.025

S+SeqTrial: Specifying Hypotheses

Example: Two arm comparison of means

One-sided equivalence test

- ♦ Null: Mean of 0 in each group
  - – Defines exact equivalence
  - – Interpretation of rejection of noninferiority is obtained from examining design
- ♦ Alt: Mean of -1 in treatment, 0 in comparison
  - – Limit of acceptable inferiority



Two-sample Normal Means

Select
Compute: ● Sample Size ○ Power ○ Min. Difference ○ Plots Only

Hypotheses
Null Means: 0, 0
Alt. Means: -1
Std. Deviation: 20
Test Type: equivalence
□ Use Log Transform

Probabilities
Significance Level: 0.025

Sequential Design

Quantifying Statistical Evidence

(Defining "Rejection of Hypotheses")

Definition of "Reject"

In order for the decision theoretic approach to make scientific sense, we need to consider the statistical criteria used to "reject" hypotheses

We need to consider the criteria used to measure

- ♦ the precision of estimates, and
- ♦ the precision of decisions about the hypotheses

Hierarchy of Statistical Questions

It is necessary to refine the (usually vaguely stated) scientific hypotheses into a form that lends itself to statistical analysis

It is useful to consider the hierarchy of refinements to the scientific question that are necessary to obtain a statistical hypothesis

- ♦ Deterministic: Does it work?
- ♦ Probability model: What proportion of the time does it work?
- ♦ Bayesian: What is the probability that it works most the time?
- ♦ Frequentist: If it didn't work most the time, would we see data like this?

Statistical Evidence

Classifications of methods to quantify statistical evidence

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?

Quantifying Statistical Evidence

In either case, we make probability statements to quantify the strength of evidence

Bayesian inference:

- Credible interval covers 100(1- α)% of posterior distribution of θ
- Posterior probability of hypotheses

Frequentist inference:

- Confidence interval defines hypotheses for which observed data is in central 100(1- α)% of sampling distribution
- P value is probability under the null of observing as extreme results as observed data

Ensuring Sufficient Statistical Evidence

In frequentist inference, it is common to choose sample size to ensure adequate precision

Width of confidence interval

Choosing the power: Issues

- Sample size requirements
- Interpretation of a negative study
  - Is the definition of “reject” the same for all hypotheses?

Standards for Statistical Evidence

None yet agreed upon, but…

The concept of having confidence in your conclusions exists in both Bayesian and frequentist inference

- Both Bayesian credible intervals and frequentist confidence intervals satisfy frequentist coverage criteria

In either type of inference, it seems reasonable to use a consistent definition of “rejection” of hypotheses

- Use of credible interval or confidence interval
- Frequentist tests: equal type I and type II errors

Quantifying Statistical Evidence

Standard choices for "level of confidence" in conclusions

Bayesian inference:

- 95% credible intervals

Frequentist inference:

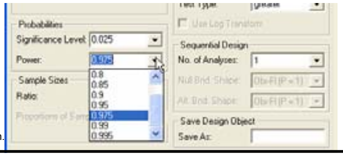
- 95% confidence intervals
- Type I error
  - one-sided level .025; two-sided level .05
- Power to detect the alternative
  - .975 is consistent with 95% CI

S+SeqTrial: Specifying Error Probabilities

S+SeqTrial generally finds frequentist designs which can then be evaluated for Bayesian properties

Probabilities groupbox

- Significance level (size, type I error)
  - One-sided or two-sided according to test type
  - .025 by default, but arbitrary choices possible
- Power
  - Upper or lower according to test type
  - .975 by default, but arbitrary choices possible



Probabilities

Significance Level: 0.025

Power: 0.975

Sample Sizes

Ratio:

Sequential Design

No. of Analyses: 1

Null Bnd. Shape: Obr-Fl (P = 1)

Alt. Bnd. Shape: Obr-Fl (P = 1)

Save Design Object

Save As:

Determining the Sample Size

Determining the Sample Size

Criteria for sample size

Sample size required to provide discrimination between hypotheses

- Frequentist: provide desired power to detect specified alternative
- Bayesian: provide sufficient precision for posterior distribution of treatment effect

Determining the Sample Size

Criteria for sample size (cont.)

Sometimes constrained by practical limitations

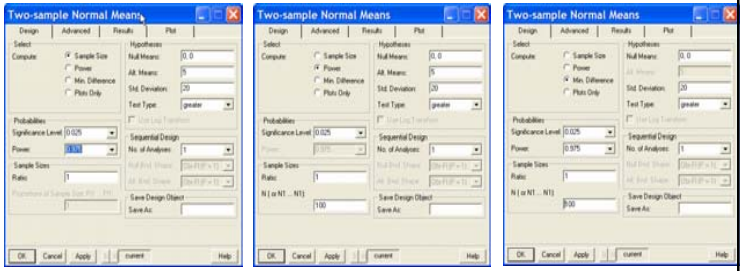
- Compute alternatives that can be discriminated from the null
  - Frequentist: alternative for which study has desired power
  - Bayesian: alternative for which posterior probability is sufficiently high when null posterior probability is low
- Determine power (frequentist) or posterior probability (Bayesian) criterion which corresponds to a particular alternative

February, 2003

S+SeqTrial: Specifying Task

S+SeqTrial can compute sample size, alternative or power

Corresponding entry field will be grayed out



Two-sample Normal Means

- Compute: ● Sample Size ○ Power ○ Min. Difference ○ Plots Only; Null Means: 0, 0; Alt. Means: 5; Std. Deviation: 20; Test Type: greater; □ Use Log Transform; Significance Level: 0.025; Power: 0.975; No. of Analyses: 1; Null Bnd. Shape: Obr-Fl (P = 1); Alt. Bnd. Shape: Obr-Fl (P = 1); Ratio: 1; Proportions of Sample Size: Pi1 ... Pi1: 1; Save As:

Two-sample Normal Means

- Compute: ○ Sample Size ● Power ○ Min. Difference ○ Plots Only; Null Means: 0, 0; Alt. Means: 5; Std. Deviation: 20; Test Type: greater; □ Use Log Transform; Significance Level: 0.025; Power: 0.975; No. of Analyses: 1; Null Bnd. Shape: Obr-Fl (P = 1); Alt. Bnd. Shape: Obr-Fl (P = 1); Ratio: 1; N (or N1 ... N1): 100; Save As:

Two-sample Normal Means

- Compute: ○ Sample Size ○ Power ● Min. Difference ○ Plots Only; Null Means: 0, 0; Alt. Means: 5; Std. Deviation: 20; Test Type: greater; □ Use Log Transform; Significance Level: 0.025; Power: 0.975; No. of Analyses: 1; Null Bnd. Shape: Obr-Fl (P = 1); Alt. Bnd. Shape: Obr-Fl (P = 1); Ratio: 1; N (or N1 ... N1): 100; Save As:

February, 2003

Determining the Sample Size

Criteria for sample size (cont.)

Special case: Censored survival studies

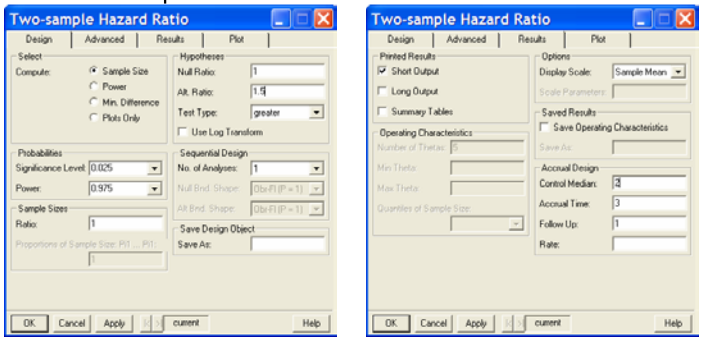
- Statistical information proportional to number of events
- Sample size computations for desired number of events
- Additional probability models used to account for accrual and follow-up patterns

February, 2003

S+SeqTrial: Accrual for Survival Studies

S+SeqTrial calculates accrual time or sample size based on uniform accrual and exponential time to event

Enter parameters on Results tab



Two-sample Hazard Ratio

- Compute: ● Sample Size ○ Power ○ Min. Difference ○ Plots Only; Null Ratio: 1; Alt. Ratio: 1.5; Test Type: greater; □ Use Log Transform; Significance Level: 0.025; Power: 0.975; No. of Analyses: 1; Null Bnd. Shape: Obr-Fl (P = 1); Alt. Bnd. Shape: Obr-Fl (P = 1); Ratio: 1; Proportions of Sample Size: Pi1 ... Pi1: 1; Save As:

Two-sample Hazard Ratio

- Printed Results: ☑ Short Output □ Long Output □ Summary Tables; Display Scale: Sample Mean; Scale Parameters:; Saved Results: □ Save Operating Characteristics; Save As:; Operating Characteristics: Number of Thetas: 5; Min Theta:; Max Theta:; Quantiles of Sample Size:; Accrual Design: Control Median: 2; Accrual Time: 3; Follow Up: 1; Rate:

February, 2003

Evaluation of Fixed Sample Designs

Evaluation of Fixed Sample Designs

Evaluating the operating characteristics

Clinical trial design is most often iterative

- Specify an initial design
- Evaluate operating characteristics
- Modify the design
- Iterate

Evaluation of Fixed Sample Designs

Operating characteristics for fixed sample studies

Level of Significance (often pre-specified)
Sample size requirements
Power Curve
Decision Boundary
Frequentist inference on the Boundary
Bayesian posterior probabilities

Evaluation of Fixed Sample Designs

Sample size requirements

Feasibility of accrual

Credibility of trial

- Validity of assumptions for statistical analysis
- Preconceived notions of sample size requirements

Evaluation of Fixed Sample Designs

Power curve

Probability of rejecting null for arbitrary alternatives

- Power under null: level of significance
- Power for specified alternative
  - Lower and upper power curves

Alternative rejected by design

- Alternative for which study has high power

Evaluation of Fixed Sample Designs

Decision boundary

Value of test statistic leading to rejection of null

- Variety of boundary scales possible

Often has meaning for applied researchers (especially on scale of estimated treatment effect)

- Estimated treatment effects may be viewed as unacceptable for ethical reasons based on prior notions
- Estimated treatment effect may be of little interest due to lack of clinical importance or futility of marketing

Evaluation of Fixed Sample Designs

Frequentist inference on the boundary

Consider confidence intervals when observation corresponds to decision boundary

Ensure desirable precision for negative studies

- Confidence interval identifies hypotheses not rejected by analysis
- Have all scientifically meaningful hypotheses been rejected?

Evaluation of Fixed Sample Designs

Bayesian posterior probabilities

Examine the degree to which the frequentist inference leads to sensible decisions under a range of prior distributions for the treatment effect

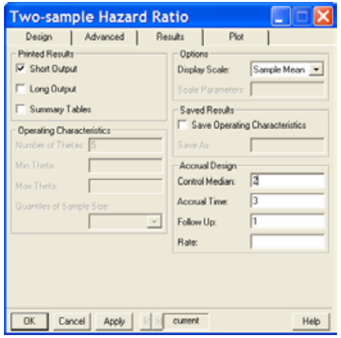
- Contour plots of Bayesian inference using conjugate normal priors

Bayesian estimates of treatment effect

S+SeqTrial: Evaluation of Designs

S+SeqTrial generates reports with critical values, power tables, frequentist inference on the boundary, Bayesian posterior probabilities

Enter desired output on Results tab



Printed Results
- ☑ Short Output
- □ Long Output
- □ Summary Tables

Operating Characteristics
- Number of Thetas: 5
- Min Theta:
- Max Theta:
- Quantiles of Sample Size:

Options
- Display Scale: Sample Mean
- Scale Parameters:

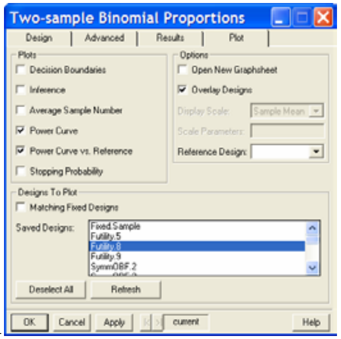
Saved Results
- □ Save Operating Characteristics
- Save As:

Accrual Design
- Control Median: 2
- Accrual Time: 3
- Follow Up: 1
- Rate:

S+SeqTrial: Evaluation of Designs

S+SeqTrial generates plots of power curves, both absolute and relative to some reference design

Enter desired output on Plots tab



Plots
- □ Decision Boundaries
- □ Inference
- □ Average Sample Number
- ☑ Power Curve
- ☑ Power Curve vs. Reference
- □ Stopping Probability

Options
- □ Open New Graphsheet
- ☑ Overlay Designs
- Display Scale: Sample Mean
- Scale Parameters:
- Reference Design:

Designs To Plot
- □ Matching Fixed Designs
- Saved Designs: Fixed.Sample, Futility.5, Futility.8, Futility.9, SymmOBF.2

Evaluation of Fixed Sample Designs

Sensitivity to assumptions about variability

Comparison of means, geometric means

- Need to estimate variability of observations

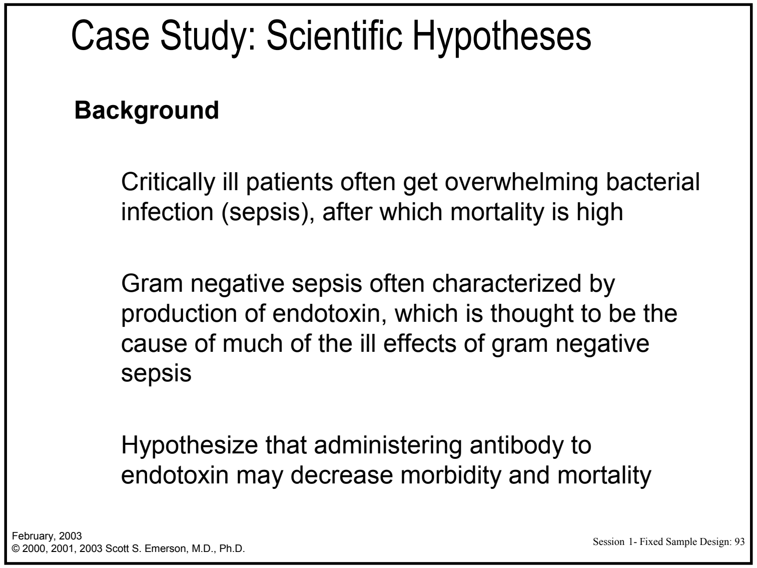
Comparison of proportions, odds, rates

- Need to estimate event rate

Comparison of hazards

- Need to estimate number of subjects and time required to observe required number of events

Case Study



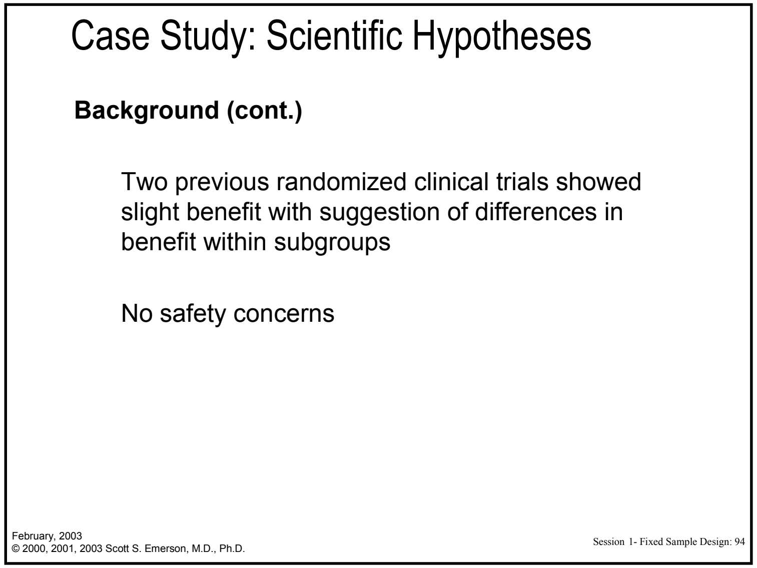
Case Study: Scientific Hypotheses

Background

Critically ill patients often get overwhelming bacterial infection (sepsis), after which mortality is high

Gram negative sepsis often characterized by production of endotoxin, which is thought to be the cause of much of the ill effects of gram negative sepsis

Hypothesize that administering antibody to endotoxin may decrease morbidity and mortality

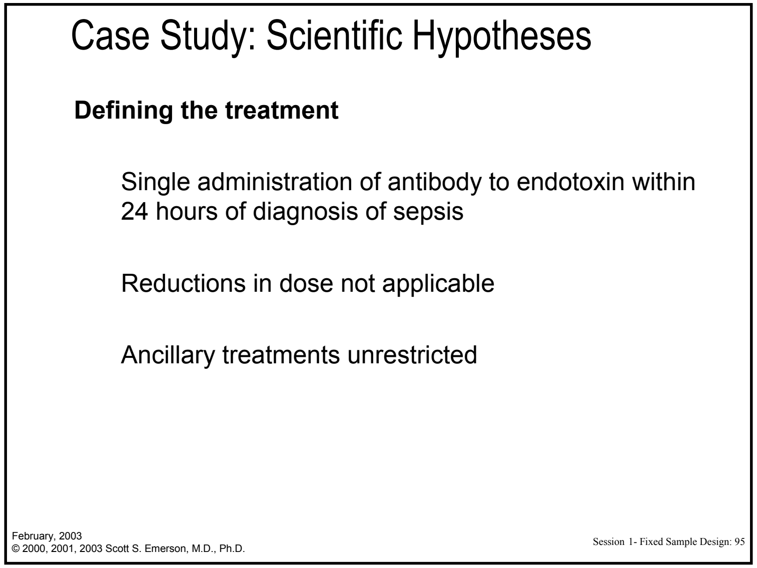


Case Study: Scientific Hypotheses

Background (cont.)

Two previous randomized clinical trials showed slight benefit with suggestion of differences in benefit within subgroups

No safety concerns



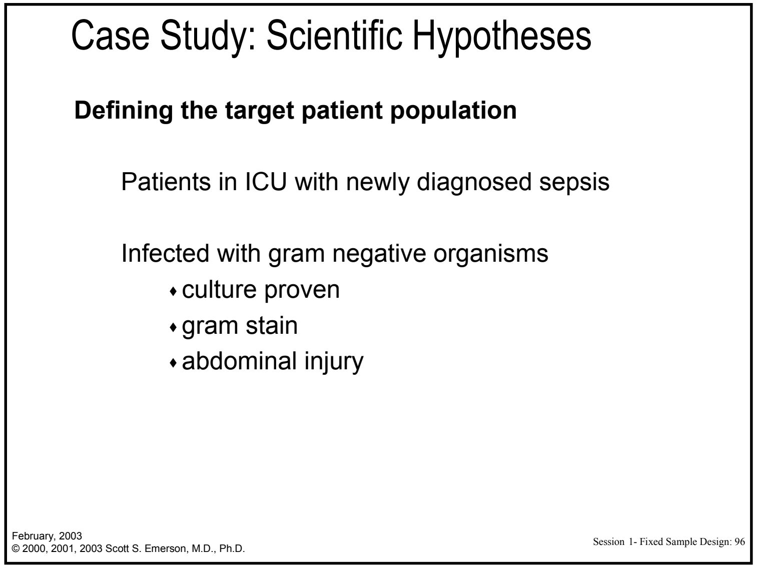
Case Study: Scientific Hypotheses

Defining the treatment

Single administration of antibody to endotoxin within 24 hours of diagnosis of sepsis

Reductions in dose not applicable

Ancillary treatments unrestricted



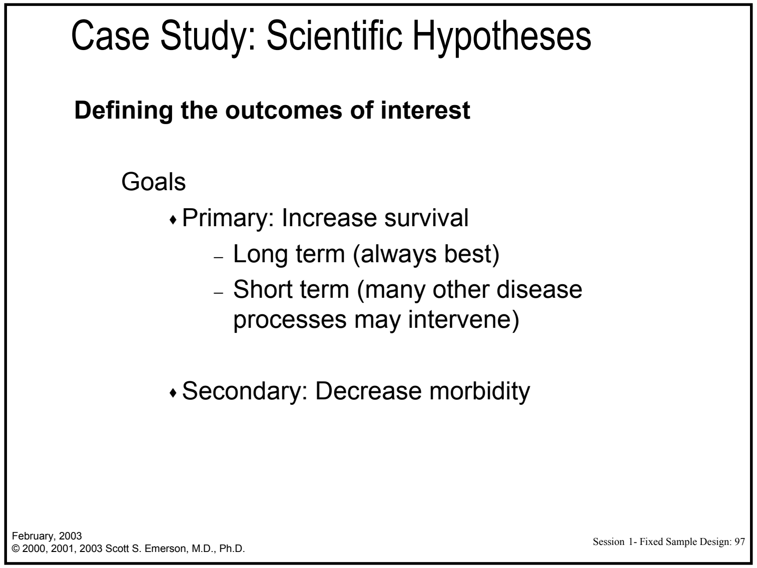
Case Study: Scientific Hypotheses

Defining the target patient population

Patients in ICU with newly diagnosed sepsis

Infected with gram negative organisms

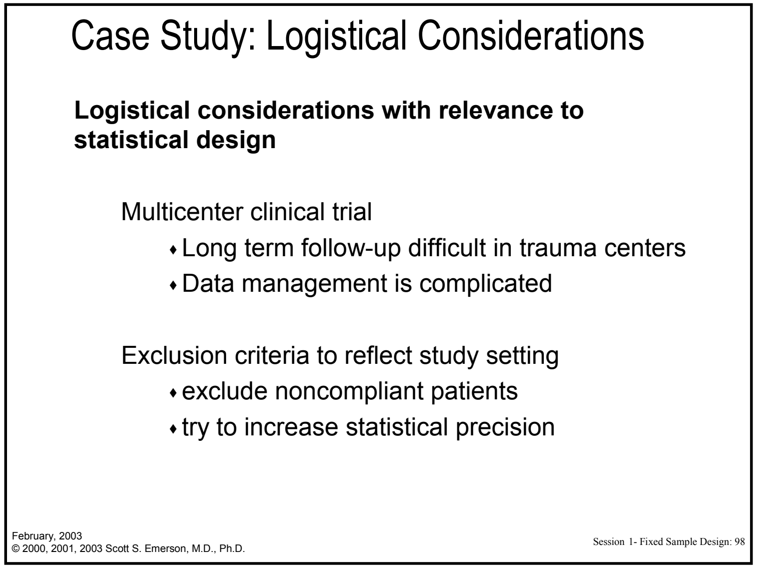
- culture proven
- gram stain
- abdominal injury



Case Study: Scientific Hypotheses

Defining the outcomes of interest

Goals
- Primary: Increase survival
  - Long term (always best)
  - Short term (many other disease processes may intervene)
- Secondary: Decrease morbidity

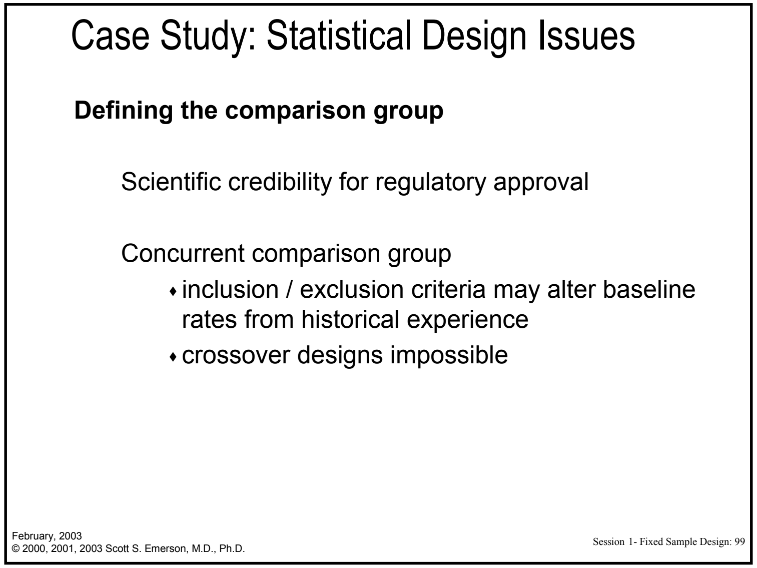


Case Study: Logistical Considerations

Logistical considerations with relevance to statistical design

Multicenter clinical trial
- Long term follow-up difficult in trauma centers
- Data management is complicated

Exclusion criteria to reflect study setting
- exclude noncompliant patients
- try to increase statistical precision

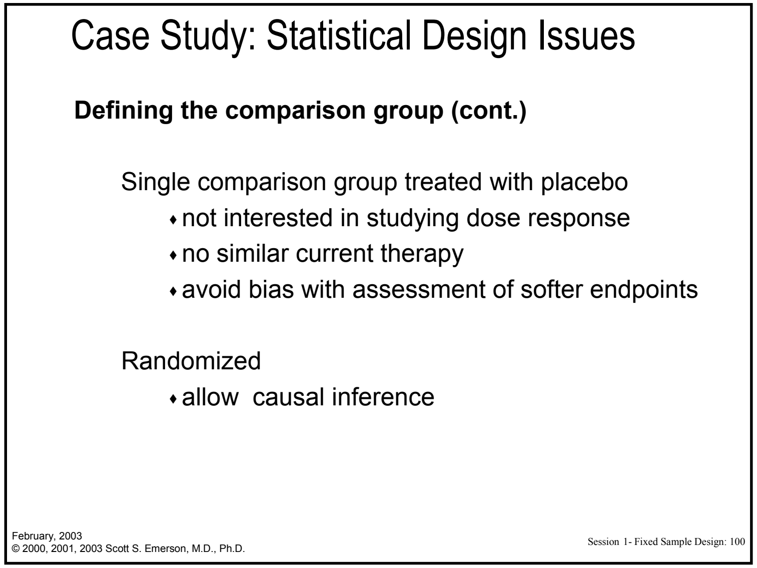


Case Study: Statistical Design Issues

Defining the comparison group

Scientific credibility for regulatory approval

Concurrent comparison group
- inclusion / exclusion criteria may alter baseline rates from historical experience
- crossover designs impossible

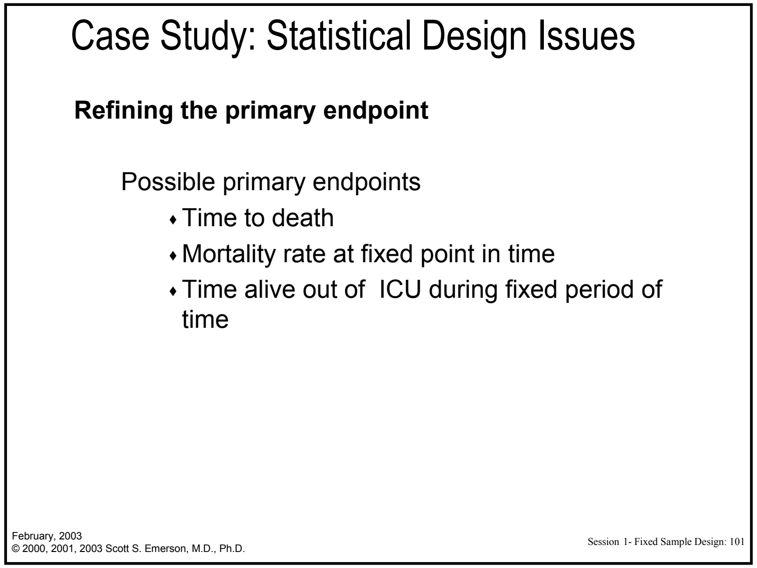


Case Study: Statistical Design Issues

Defining the comparison group (cont.)

Single comparison group treated with placebo
- not interested in studying dose response
- no similar current therapy
- avoid bias with assessment of softer endpoints

Randomized
- allow causal inference

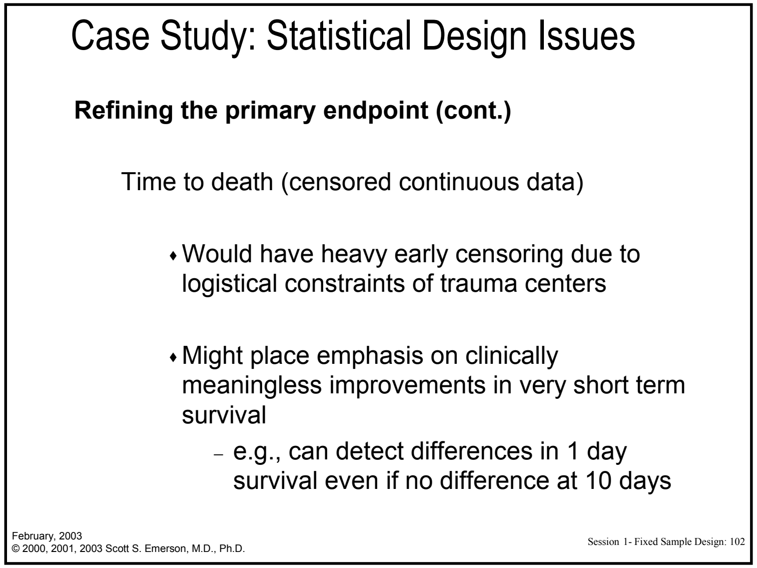


Case Study: Statistical Design Issues

Refining the primary endpoint

Possible primary endpoints

- Time to death
- Mortality rate at fixed point in time
- Time alive out of ICU during fixed period of time

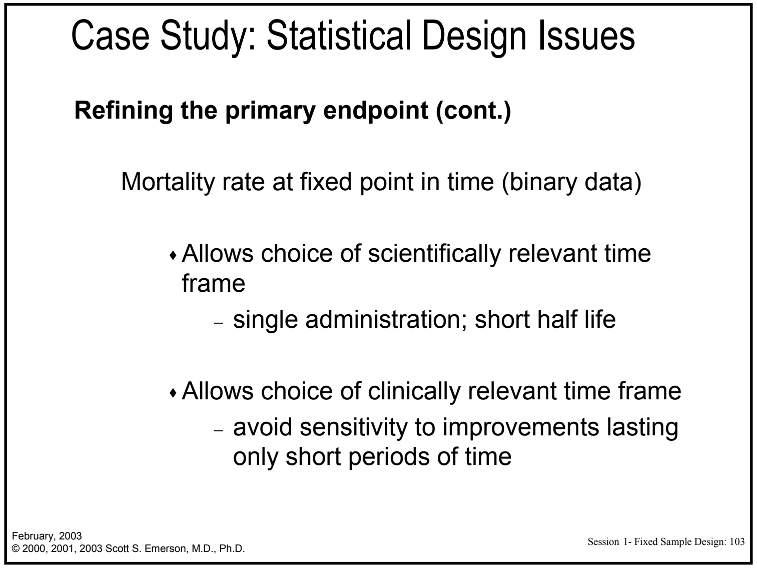


Case Study: Statistical Design Issues

Refining the primary endpoint (cont.)

Time to death (censored continuous data)

- Would have heavy early censoring due to logistical constraints of trauma centers
- Might place emphasis on clinically meaningless improvements in very short term survival
  - e.g., can detect differences in 1 day survival even if no difference at 10 days

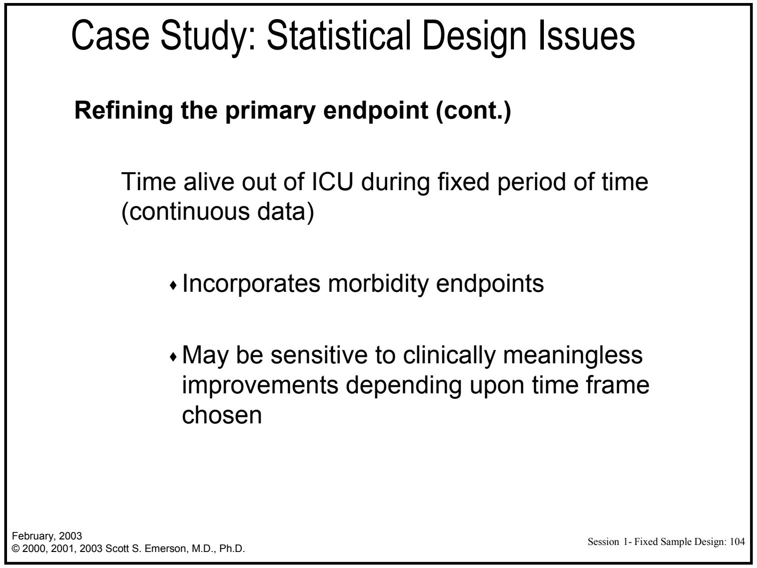


Case Study: Statistical Design Issues

Refining the primary endpoint (cont.)

Mortality rate at fixed point in time (binary data)

- Allows choice of scientifically relevant time frame
  - single administration; short half life
- Allows choice of clinically relevant time frame
  - avoid sensitivity to improvements lasting only short periods of time



Case Study: Statistical Design Issues

Refining the primary endpoint (cont.)

Time alive out of ICU during fixed period of time (continuous data)

- Incorporates morbidity endpoints
- May be sensitive to clinically meaningless improvements depending upon time frame chosen

Case Study: Statistical Design Issues

Refining the primary endpoint (cont.)

Primary endpoint selected (binary data)

- Sponsor: 14 day mortality
- FDA: ? 28 day mortality

Case Study: Statistical Design Issues

Refining the primary endpoint (cont.)

Summary measures within groups for binary data

- Proportion with event
- Odds of event

Measures of treatment effect (comparison across groups)

- Difference in proportions
- Odds ratio

Case Study: Statistical Design Issues

Defining the method of analysis

Test for differences in binomial proportions

- Ease of interpretation
- 1:1 correspondence with tests of odds ratio (for known baseline event rates)

No adjustment for covariates

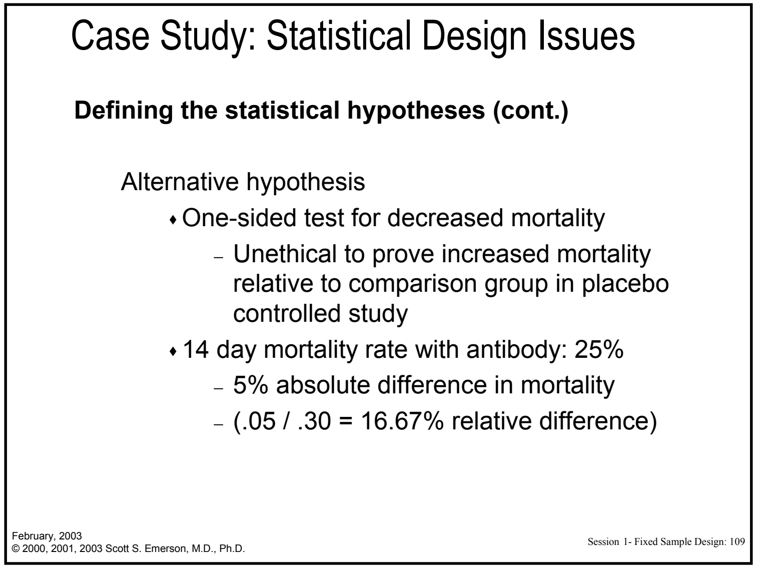
Statistical information dictated by mean-variance relationship of Bernoulli random variable: p(1-p)

Case Study: Statistical Design Issues

Defining the statistical hypotheses

Null hypothesis

- No difference in mortality between groups
- Estimated baseline rate
  - 14 day mortality: 30%
  - (needed for estimates of variability)



Case Study: Statistical Design Issues

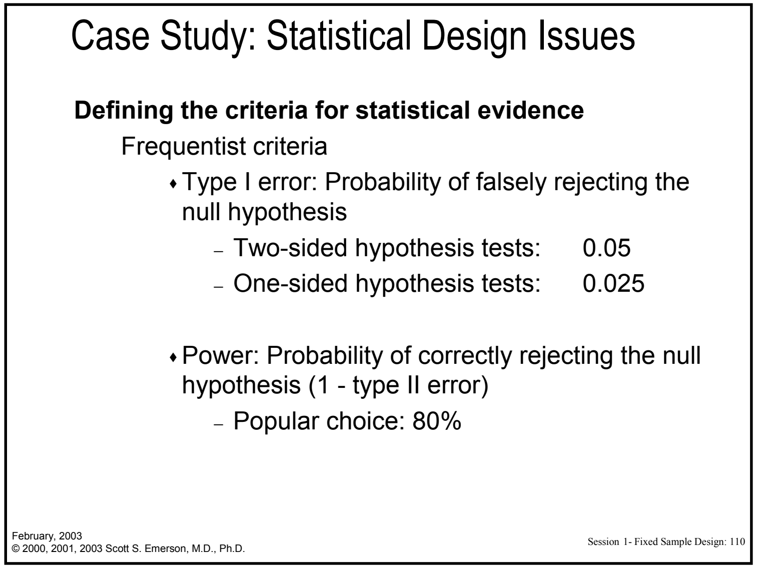
Defining the statistical hypotheses (cont.)

Alternative hypothesis

- One-sided test for decreased mortality
  - Unethical to prove increased mortality relative to comparison group in placebo controlled study
- 14 day mortality rate with antibody: 25%
  - 5% absolute difference in mortality
  - (.05 / .30 = 16.67% relative difference)

February, 2003
© 2000, 2001, 2003 Scott S. Emerson, M.D., Ph.D.

Session 1- Fixed Sample Design: 109



Case Study: Statistical Design Issues

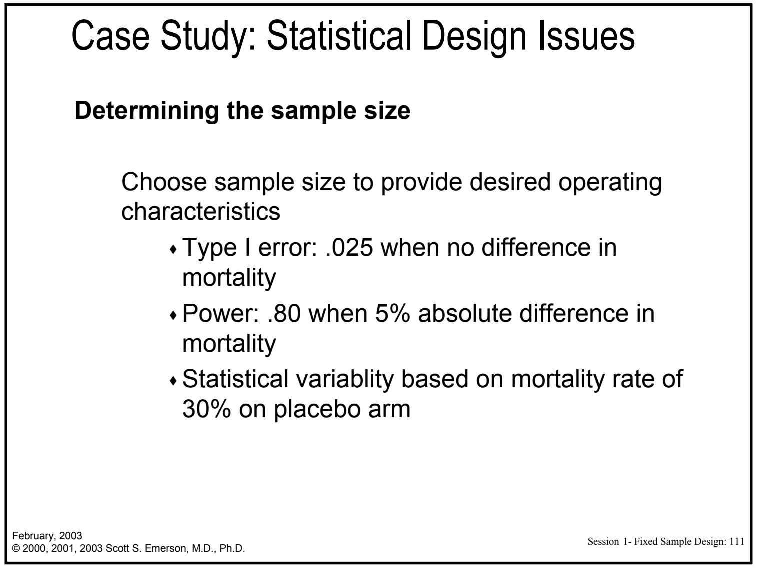
Defining the criteria for statistical evidence

Frequentist criteria

- Type I error: Probability of falsely rejecting the null hypothesis
  - Two-sided hypothesis tests: 0.05
  - One-sided hypothesis tests: 0.025
- Power: Probability of correctly rejecting the null hypothesis (1 - type II error)
  - Popular choice: 80%

February, 2003
© 2000, 2001, 2003 Scott S. Emerson, M.D., Ph.D.

Session 1- Fixed Sample Design: 110



Case Study: Statistical Design Issues

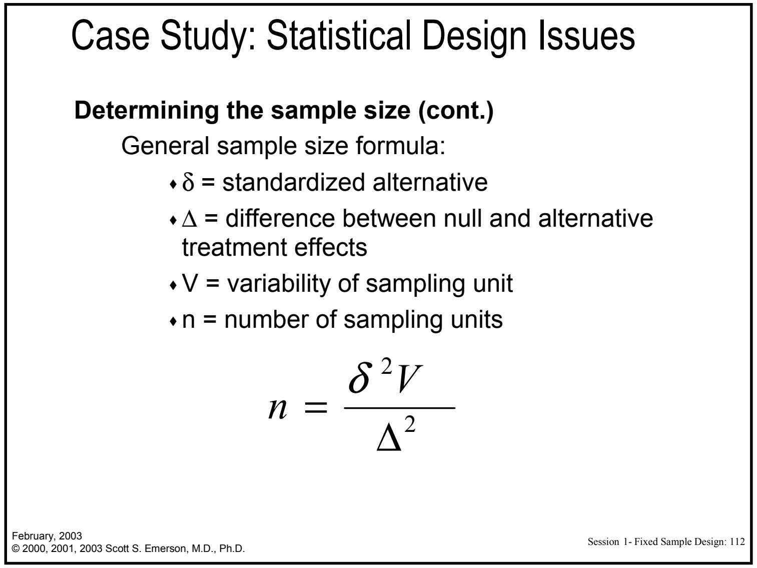
Determining the sample size

Choose sample size to provide desired operating characteristics

- Type I error: .025 when no difference in mortality
- Power: .80 when 5% absolute difference in mortality
- Statistical variablity based on mortality rate of 30% on placebo arm

February, 2003
© 2000, 2001, 2003 Scott S. Emerson, M.D., Ph.D.

Session 1- Fixed Sample Design: 111



Case Study: Statistical Design Issues

Determining the sample size (cont.)

General sample size formula:

- δ = standardized alternative
- Δ = difference between null and alternative treatment effects
- V = variability of sampling unit
- n = number of sampling units

$$n = \frac{\delta^2 V}{\Delta^2}$$

February, 2003
© 2000, 2001, 2003 Scott S. Emerson, M.D., Ph.D.

Session 1- Fixed Sample Design: 112

Case Study: Statistical Design Issues

Determining the sample size (cont.)

Fixed sample test (no interim analyses):

- $\delta = (z_{1-\alpha} + z_{\beta})$ for size α, power β

Two sample test of binomial proportions

- $\Delta = (p_{T1} - p_{C1}) - (p_{T0} - p_{C0})$
- $V = p_{T1}(1 - p_{T1}) + p_{C1}(1 - p_{C1})$ under H_1
- n = sample size per arm

Case Study: Statistical Design Issues

Determining the sample size (cont.)

Sample size planned trial:

$$n = \frac{(z_{1-\alpha} + z_{\beta})^2 (p_T q_T + p_C q_C)}{(p_T - p_C)^2}$$

$$n = \frac{(1.96 + .841)^2 (.25 \times .75 + .3 \times .7)}{.05^2} = 1247.97$$

S+SeqTrial

Designing the fixed sample study

Sample size and critical value

```
PROBABILITY MODEL and HYPOTHESES:
 Two arm study of binary response variable
 Theta is difference in probabilities (Treatment - Comparison)
 One-sided hypothesis test of a lesser alternative:

           Null hypothesis : Theta >=  0      (size  = 0.025)
    Alternative hypothesis : Theta <= -0.05   (power =  0.8)
   (Fixed sample test)

STOPPING BOUNDARIES: Sample Mean scale
                             a      d
    Time 1 (N= 2495.95) -0.035 -0.035
```

Case Study: Statistical Design Issues

Evaluating the operating characteristics

Critical values

- Observed value which rejects the null
- Point estimate of treatment effect
  - Will that effect be considered important?
  - (Clinical and marketing relevance)

Case Study: Statistical Design Issues

Evaluating the operating characteristics (cont.)

Confidence interval at the critical value

- Observed value which fails to reject null
- Set of hypothesized treatment effects which might reasonably generate data like that observed
  - Have we excluded all scientifically meaningful alternatives with a negative study?
  - (Basic science relevance)

Case Study: Statistical Design Issues

Evaluating the operating characteristics (cont.)

Power curve

- Probability of rejecting null hypothesis across various alternatives

Bayesian

- Posterior probability of hypotheses at the critical values

Case Study: Statistical Design Issues

Evaluating the operating characteristics (cont.)

Sensitivity to assumptions used in design

- variance estimates
- different sample sizes

S+SeqTrial

Evaluating the fixed sample study

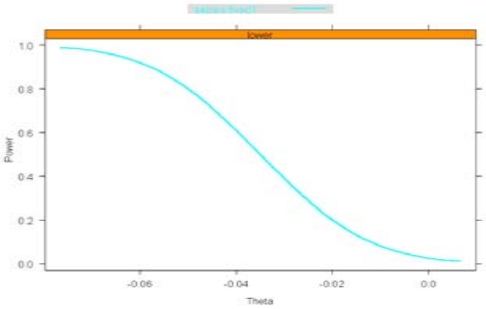
Frequentist inference at the boundaries

```
Inferences at the Boundaries
                    *** a Boundary *** *** d Boundary ***
Time 1     Boundary             -0.035             -0.035
                MLE             -0.035             -0.035
            P-value              0.025              0.025
       95% Conf Int        (-0.07, 0)         (-0.07, 0)
```

S+SeqTrial

Evaluating the fixed sample study

Power curve



S+SeqTrial

Re-designing the fixed sample study

Sample size 1700 subjects (850 / arm)

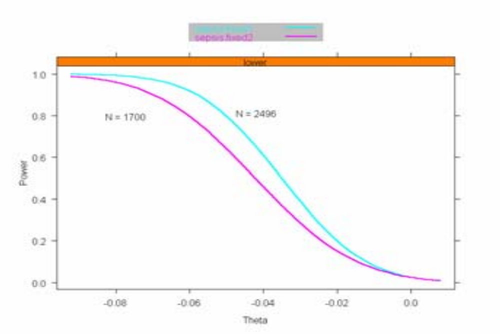
```
PROBABILITY MODEL and HYPOTHESES:
 Two arm study of binary response variable
 Theta is difference in probabilities (Treatment - Comparison)
 One-sided hypothesis test of a lesser alternative:

           Null : Theta >=  0           (size  =  0.025 )
    Alternative : Theta <= -0.06019     (power =  0.8   )
 (Fixed sample test)
STOPPING BOUNDARIES: Sample Mean scale
                           a       d
    Time 1 (N= 1700) -0.0421 -0.0421
```

S+SeqTrial

Comparing the effect of sample size

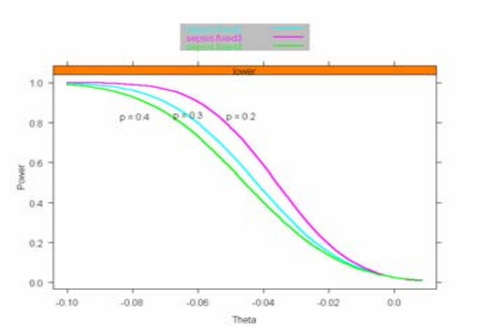
Power curve



S+SeqTrial

Comparing the effect of baseline mortality

Power curve


sepsis.fixed3
lower
Power
p = 0.4
p = 0.3
p = 0.2
Theta