Sample Size Calculations in Randomized Clinical Trials

Martin L. Lesser, PhD

Feinstein Institute for Medical Research
Biostatistics Unit
CME Disclosure Statement

• The North Shore LIJ Health System adheres to the ACCME's new Standards for Commercial Support. Any individuals in a position to control the content of a CME activity, including faculty, planners and managers, are required to disclose all financial relationships with commercial interests. All identified potential conflicts of interest are thoroughly vetted by the North Shore-LIJ for fair balance and scientific objectivity and to ensure appropriateness of patient care recommendations.

• Course Director, Kevin Tracey, has disclosed a commercial interest in Setpoint, Inc. as the cofounder, for stock and consulting support. He has resolved his conflicts by identifying a faculty member to conduct content review of this program who has no conflicts.

• The speaker, Martin L. Lesser, PhD, has no conflicts.
Why is Sample Size Important in Clinical Trials?

• Must know whether a “negative” result is likely to be a “false negative” (Type II error)
• Sample size must be sufficient in order to yield reliable results
• It is not ethical to place subjects at risk if the study has a high likelihood of being inconclusive
Common Settings for Sample Size Calculations

- **Dichotomous (binary) outcomes (rates and proportions)**
  - Pearson chi-square test, odds ratios, McNemar’s test

- **Continuous outcomes**
  - t-test, paired t-test, ANOVA, repeated measures ANOVA

- **Survival (time to event) outcomes**
  - log rank test, Gehan-Wilcoxon test, Cox regression
Basics of Hypothesis Testing
(in a 2-arm clinical trial)

• Subjects randomized to Arm A or Arm B
• Investigator wants to show that B is better than A (or, more generally, A and B are different)
• Null hypothesis denoted by $H_0: A=B$
• Alternative hypothesis denoted by $H_A: B>A$ (or, $H_A: B < A$ or, $H_A: B \neq A$)
• In general, $H_A$ is what the investigator wants to prove
### Decision Making in Hypothesis Testing

<table>
<thead>
<tr>
<th>True Situation</th>
<th>No Difference ((H_0))</th>
<th>Difference Exists ((H_A))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difference Exists</td>
<td>Type I error ((\alpha \text{ error}))</td>
<td>No error</td>
</tr>
<tr>
<td>(Reject (H_0))</td>
<td></td>
<td>No error</td>
</tr>
<tr>
<td>No Difference</td>
<td>No error</td>
<td>Type II error ((\beta \text{ error}))</td>
</tr>
<tr>
<td>(Do not reject (H_0))</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
More on Hypothesis Testing

• “Trial by Jury” analogy: “not guilty” is not the same as “innocent”
• Usually, $\alpha=0.05$ (false positive error rate, Type I error)
• Usually, $\beta=0.10$ or $\beta=0.20$ (false negative error rate, Type II error)
• Power = $1 - \beta$ (usually 0.90 or 0.80)
Power of a Statistical Test

• Power = 1 − β
  = Probability (reject $H_0$, given $H_A$ is true)
  = Probability (conclude that the two treatments are different, when, in fact they are different)

• Power depends on:
  - $\alpha$ level
  - 1-tailed or 2-tailed test (2-tailed test preferred)
  - $\beta$ level
  - Members of $H_A$ for which one wants to achieve specified power (there are many possibilities for $H_A$)
How Does This All Work in the Real World?

• Investigator and statistician set up the problem and specify $\alpha$, 1- or 2-tailed test, and, usually, $\beta$.
• Investigator specifies the true parameters (e.g., A and B) for which he/she wants to achieve the specified power $1 - \beta$.
• Statistician computes the required sample size (n per arm) to achieve the specified power.

- Typically, the investigator states that the sample size estimate is too high!!
Example: Binary Outcomes

• RCT to compare two response rates
• Arm A is current standard therapy, used on large numbers of patients for past 10 years
• Arm B is new treatment
• Want to compare response rates (chi-square test)
• Assume $p_A = 30\%$. Want to see an increase in response rate to at least 50\% in order for it to be considered a clinical improvement
Example: Binary Outcomes (cont’d)

• $H_0: p_A = p_B$ vs. $H_A: p_A \neq p_B$

• $\alpha = 0.05$, $1 - \beta = 0.90$, 2-tailed test

• $n = 125$ subjects / arm

• For 80% power, $n=93$/arm

• For 30% vs. 60%, with 80% power, $n=42$/arm

• For 30% vs. 35%, with 80% power, $n=1377$/arm
A Note About Rates and Proportions

• Different combinations of rates yield similar sample size requirements
  - 10% vs. 35%, 80% power: n=54 subjects/arm
  - 25% vs. 55%, 80% power: n=54
  - 45% vs. 75%, 80% power: n=54

• Sample size/power is a function of the two proportions, not just the difference between the proportions

• Rates near 50% or that average near 50% generally require larger sample sizes
Example: Continuous Outcome

- RCT to compare two diets for weight loss
- Arm A is current standard therapy, used on large numbers of patients for past 10 years
- Arm B is new treatment
- Want to compare mean weight loss (t-test)
- Assume mean weight loss $\mu_A = 20.5$ pounds.
  Want to see an increase in weight loss to 30.0 pounds in order for it to be considered a clinical improvement
Example: Continuous Outcome (cont’d)

- Important: Must be able to specify the standard deviation ($\sigma$). Suppose $\sigma = 10.1$ pounds.
- Then, standardized difference ("effect size") is 
  \[ d = \frac{30.0 - 20.5}{10.1} = 0.94 \]
- $H_0: \mu_A = \mu_B$ vs. $H_A: \mu_A \neq \mu_B$
- $\alpha = 0.05, 1 - \beta = 0.90$ (90% power), 2-tailed test
- $n = 25$ subjects / arm (Note: $n=21$, 1-tailed test)
- For 80% power, $n=19$ / arm (Note: $n=15$, 1-tailed test)
Plot of Power vs. Sample Size for Effect Size = 0.94

Two group t-test of equal means (equal n's)

Sample Size per Group
5 10 15 20 25

Power
50
60
70
80
90
100

Sample Size per Group

Æ = 0.050 (2) ᴬ = 1.000

© 2006 Martin L Lesser, PhD
Survival Analysis: A Bit More Complicated

• Applies to any endpoint that is “time until an event” and allows for censored data
• Censoring complicates the problem
• Power depends heavily on the number of censored observations
• Typically, sample size calculation depends on:
  • Median survival in each arm
  • Fixed or variable follow-up on each subject?
  • Length of accrual period
  • Length of follow-up period
The Challenge: Specifying the Parameters of Clinical Importance

• What is the smallest difference in the rates (or means) that is of clinical interest?
• Often a hard question to answer
• Answer may need to balance efficacy vs. safety vs. cost vs. other considerations
• Different clinicians may differ on specifications
• Need good data from prior trials on which to base the specification
• Justification of a randomized trials usually requires extensive prior data
Sample Size “Negotiations”

• What happens when the investigator can’t provide the specifications?
  • Use data from analogous situations
  • “Generic” calculations

• What happens when the initial calculation yields an unrealistically large sample size?
  • Change the assumptions
  • Change the desired power
  • Negotiate
Patient Flow in Clinical Trials

Available

Considered

Eligible

Consented

Enrolled

Compliant

Adequately Followed
Related Topics

• Sample size for paired data
  – For means, need to specify within subject variation, i.e., $\sigma_{\text{diff}}$; can be done if $\sigma_1$, $\sigma_2$, and $\rho_{12}$
  – For proportions (McNemar’s test), need to specify proportion of discordant observations

• More than 2 treatment arms
  – Adjustments for multiple comparisons

• Repeated measures and crossover designs
  – Need to specify within subject correlations
Related Topics (cont’d)

• Equivalence and non-inferiority trials
  – Need to specify a “difference” that is to be considered equivalent
  – Sample sizes can be very large for proving equivalence
  – Null and alternative hypotheses are reversed
    • $H_0: B > A$ vs. $H_A: B = A$
      more specifically,
    • $H_0: B - A \geq \Delta$ vs. $H_A: B - A < \Delta$
Related Topics (cont’d)

• Sample size based on estimation and confidence intervals
• “Retrospective” power calculations (after a study has been completed)
• Use of simulations for sample size calculations
• Effect of multiple endpoints
• Non-parametric methods
• Effect of “intention to treat” on sample size calculations
**Simon 2-Stage Optimal Design**

- $H_0: p \leq p_0$ vs. $H_A: p \geq p_1$
- Where response rate $\leq p_0$ is uninteresting and response rate $\geq p_1$ is the desired target
- Simon’s “Optimal Design”: Observe $n_1$ subjects in stage 1. If response rate $r_1 \leq a_1/n_1$, then stop the trial and reject the drug.
- If $r_1 > a_1/n_1$, then study an additional $n_2$ subjects in stage 2, for a total of $n=n_1+n_2$. If the “total” response rate $r \leq a/n$, then reject the drug. If $r > a/n$, then consider the drug for further testing and Phase III trials.

Simon 2-Stage Optimal Design (cont’d)

- For given $\alpha$, $\beta$, $p_0$, and $p_1$, this design minimizes $EN(p_0)$, the expected number of subjects studied under $H_0$.
- Example: Let $\alpha=0.05$ $\beta=0.20$ $p_0=0.30$ $p_1=0.45$
  
  Stage 1: Enter 27 subjects; stop trial and reject drug if $r_1 \leq \frac{9}{27}$.
  
  If $r_1 > \frac{9}{27}$, then go on to Stage 2.
  
  Stage 2: Enter 54 additional subjects (total=81).
  
  If $r \leq \frac{30}{81}$, then reject the drug.
  
  If $r > \frac{30}{81}$, then trial is favorable toward drug.

Note: $E(N(p_0)) = 41.7$. Prob(early termination) = 0.73
Simon 2-Stage Minimax Design

• Similar to the 2-stage optimal design
• Minimizes the maximum total sample size (n) among all optimal designs
• Minimax design is attractive when subject accrual is low
• Previous example worked with minimax:
  \[ r_1 \leq \frac{16}{46}, r \leq \frac{25}{65}, \text{EN}(p_0)=49.6, \text{PET}(p_0)=0.81 \]
  (Optimal design had n=81.)