Sample Size Determinations for Clinical Trials

Janet Raboud

Advanced Statistical Methods for Clinical Trials

Reasons for Sample Size Calculations

- An undersized trial will miss the effect of interest
- An oversized trial is a waste of resources
- Both are unethical
  - Not fair to enrol patients in a trial which will not yield meaningful results.
  - Also not fair to expose more patients than necessary to trial procedures, placebo treatment, and possibly reduce their chances of enrolling in other trials.

Framing the Question

- The most important step in planning a randomized trial is framing the research question.
- What is the most important outcome?
  - Eg. Efficacy, toxicity?
- How will it be measured?
- What are the comparator arms?

"I want to measure adherence"

- Pill counting, MEMS caps, patient report?
- Percentage of pills taken divided by pills prescribed? (ie. a continuous outcome)
- Proportion of patients with perfect adherence. (binary outcome)
- Proportion of patients at least 80% adherent.
- Average over entire study or a particular time period?

"I want to measure change in body fat"

- Right arm fat, left arm fat, average arm fat, leg fat, limb fat, total fat, …
- Grams of body fat at week 24.
- Absolute change in body fat to week 24.
- Percentage change in body fat by week 24.
- Proportion of patients with a 5g increase in fat.
- Proportion of patients with a 5% increase in fat.
- Different time points.

Primary Outcome

- The first step to sample size calculation is specification of the primary outcome.
  - In a grant, the variable specified as the primary outcome should be the same one used for the sample size calculations.
  - The primary analysis should match the analysis specified in the sample size calculation section.
- If some secondary outcomes are of particular interest, power calculations should be performed and presented for these variables too, to show what differences can be detected with the proposed sample size.

Primary outcome - details

- Specify whether the primary outcome is continuous, categorical, binary, time to event.
- Specify the time point during the study at which the comparison of the primary endpoint is of most interest.
- If interested in a change in the primary outcome from baseline, think about whether it makes more sense to look at absolute change or percentage change in the variable.

The Sample Size "Dance"

- Physician: "How many patients do I need?"
- Statistician: "How many do you have?"
- Sample size calculation is an iterative process involving a discussion of
  - The primary outcome.
  - The effect size of interest.
  - Guesstimate of standard deviation/other parameters
  - The number of patients available, eligible, and expected to be willing to participate.
  - Other ongoing, competing trials.
  - Time frame necessary to complete the trial so that the results are still of interest to the scientific community
  - Other possible sites that could join a multi-site trial.

Range of Calculations

- A sample size calculation is not usually a single calculation but a **set** of calculations, which can be presented in a table or graph.
- It is important to be aware of how the required sample size varies according to variations in input parameters, **especially those whose values you are least certain of**.

Limited Pool of Patients

- If the sample size is limited by either
  - Financial constraints
  - Available pool of patients or
  - Time frame for recruiting

  Then calculations should focus on what difference can be detected with the sample size available, and/or the power you have to detect a particular effect of interest.

Sample Size

- The calculated sample size is the evaluable number of patients required for the **analysis**.
- The last step in the calculation is to determine the number of patients you need to randomize in order to achieve the required number of evaluable patients.

Dropouts

- How will the primary endpoint be measured for participants who do not complete the trial?
- Can the endpoint be inferred, imputed or assigned a "worst possible score"?
- If not, are patients who do not complete the trial expected to be "missing at random"?
- Sample size calculations are often adjusted for proportion of patients who are not expected to complete the trial.
- N* = N/percentage expected to complete the trial

Adjustment for Incomplete Follow-up

- Depending on the outcome, the sample size calculations may already adjust for incomplete follow-up
  - Survival analysis – by specification of median duration of follow-up
  - Longitudinal outcome – by specification of number of follow-up visits
  - Binary outcome – if missing is assumed to be "failure".

Parameters for Sample Size Calculation

- Significance level (α): The probability of a type I error, that the null hypothesis will be rejected and it will be concluded that there is an effect when in fact there is none. This is usually set to .05.
- Power ($1-\beta$): where β is the probability of a type II error, the chance of missing a clinically significant difference. Power is usually set to 80%.
- Clinically important difference: The smallest difference in means/proportions/etc that the clinicians feel is worth detecting. This difference can be smaller or larger than *observed* differences from other trials.

Parameters, ctd

- One-sided vs two-sided test.
  - 2 sided tests are more conservative
  - Only use one-sided tests when one is only interested in outcomes in one tail of the distribution
  - Use a 2 sided test unless you have a good reason not to.
- Standard deviation
  - If an estimate of the standard deviation is not available, a reasonable guess is the range/4.
  - Information on standard deviation of changes can be trickier to find.

Types of Sample size formulae

- Comparison of means between 2 groups
- Comparison of proportions
- Comparison of time to event
- Comparison of Poisson rates
- Longitudinal outcomes
  - Rate of change
  - Average value
  - Binary outcomes
- Crossover trials
- Equivalence trials

Comparison of Two Means

The numbers of patients, n_1 and n_2, required in Groups 1 and 2 to detect a difference Δ in means with significance level α and power 1-β assuming variances σ_1^2 and σ_2^2 in populations 1 and 2 are:

$$n_1 = (\sigma_1^2 + \sigma_2^2/k)(z_{1-\alpha/2} + z_\beta)^2 / \Delta^2$$

$$n_2 = (k\sigma_1^2 + \sigma_2^2)(z_{1-\alpha/2} + z_\beta)^2 / \Delta^2$$

where $k = n_2/n_1$ = ratio of 2 sample sizes.

Power: computer program for sample size and power calculations. Download for free from http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize

Example of Sample Size to Compare Means

- Randomized trial of two weight loss programs
- Primary outcome = change in weight at 6 months
- Effect size of interest = 5 pounds
- Expected range of change in weight: -40 lbs to 5 lbs
  - Can infer that std dev of change in weight is approximately 45/4 = 11
- Assuming Δ=5, σ_1=σ_2, σ=11, α=.05, 80% power, the sample size per group is 77.

Sample size per group with varying assumptions

	Effect Size of Interest (Δ)		
Std Dev (σ)	2.5 lbs	5 lbs	7.5 lbs
8	162	41	19
9	204	52	24
10	252	64	29
11	305	**77**	35
12	363	91	41
13	425	107	48
14	493	124	56
15	566	142	64

Comparison of Proportions between Two Groups

The number of participants *per group* required to detect a difference p_1-p_2 in the proportions with significance level α and power 1-β is

$$n = \frac{[p_1(1-p_1) + p_2(1-p_2)](z_{1-\alpha/2} + z_{1-\beta})^2}{(p_1 - p_2)^2}$$

where p_1 is the expected proportion in Group 1

and p_2 is the expected proportion in Group 2.

Example of Sample Size Calculation to Compare Proportions

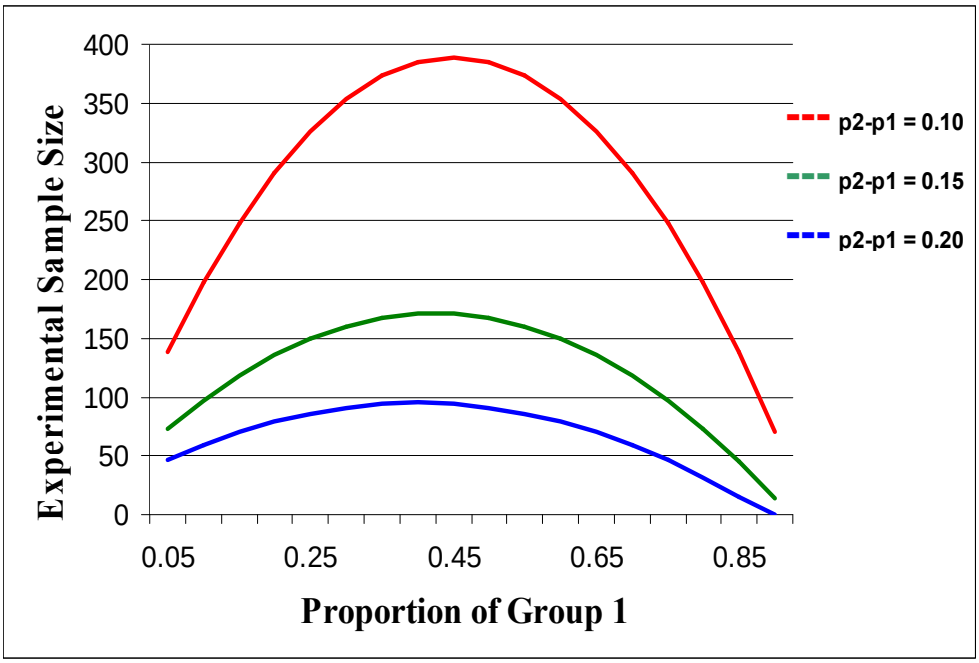
- Randomized trial of two treatments for HIV patients
- Primary outcome = proportion of patients with viral load (VL) less than the limit of detection at 48 weeks
- Expect that 60% of patients in the standard of care arm will have suppressed VL
- Interested in an effect size of 20%
- Assuming p_1=.6, p_2=.8, α=.05, 80% power, the sample size **per group** is 81.

Sample size **per group** with varying assumptions

	Effect Size of Interest (p_2-p_1)		
p_1	**15%**	**20%**	**25%**
40	173	97	61
45	173	96	60
50	169	93	58
55	162	88	54
60	152	**81**	49
65	138	72	43
70	120	62	35
75	100	49	30*

* Assuming p_2 = .99.

Assumes: $\alpha = 0.05$, 80% power
Varying p_1 and p_2



Power Calculations if Sample Size is Fixed

- If the maximum sample size for the study is 100 patients (50 in each group), we would want to calculate
  - The **power** to detect the difference of interest given the assumed proportions of patients in the "standard of care group" with the outcome
  - The **detectable difference** for a range of proportions of patients with the outcome in the standard of care group for the fixed sample size.

Power to detect difference of interest

$$\text{Power} = \Phi\left(Z < \frac{|p_1 - p_2|}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} + z_{\alpha/2} \frac{\sqrt{2p(1-p)}}{\sqrt{p_1(1-p_1) + p_2(1-p_2)}} \right)$$

- Assume p_1=.6, p_2=.8, α=.05, N per group = 50
- Then power to detect the difference is 59%

Detectable difference with N=50 per group, significance level of .05 and 80% power

p_1	Detectable p_2
.4	.15 or .68
.5	.23 or .77
.6	.32 or .85
.7	.43 or .92
.8	.54 or .98

Count Data

- Examples of Outcomes with Poisson Distributions
  - Rates of nosocomial infections per 1000 patient-days in hospital
  - Rates of Strep A per 100,000 population
- Characteristics of variables with Poisson dists:
  - Rare events
  - Cases reported in two distinct time periods are independent
  - The probability of observing an event is directly proportional to the length of the time period
- Probability of k events in time period T is

$$\Pr(k) = e^{-\lambda T}(\lambda T)^k / k!$$

Comparison of Two Poisson Rates

- Consider a situation where we want to determine the sample size necessary to compare two Poisson rates
- λ_1 and λ_2 are the rates for the two populations
- $\rho = \lambda_1 / \lambda_2$ is the ratio of the two rates
- d is the relative size of the two sampling frames
- The number of events needed to be observed in population 1 to detect a rate of ρ with significance level α and power $1-\beta$ is

$$n_1 = \frac{(\rho/d + 1)(f z_{1-\alpha} + z_\beta)^2}{(\rho - 1)^2} \qquad \text{where} \qquad f = \sqrt{\frac{1 + d\rho}{d + \rho}}$$

- The size of population 1 needed is: $r_1 = n_1 / \lambda_1$
- The size of population 2 needed is $r_2 = d\, r_1$

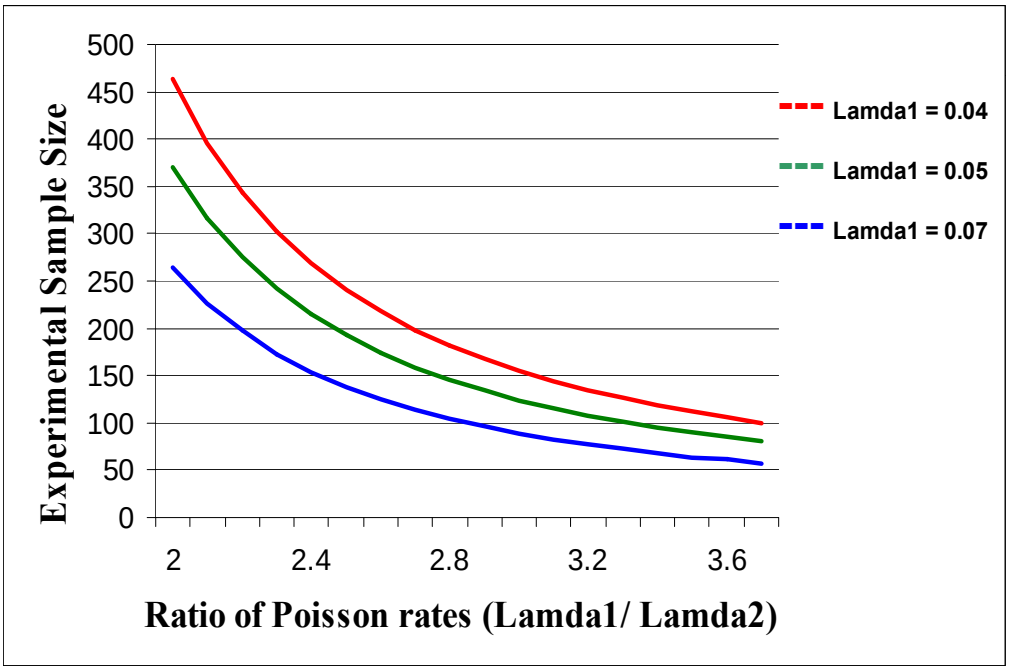
Example of Comparison of Poisson Rates

- Assume we are comparing rates of infection in two hospitals.
- We measure rates of infection as the number of infections per patient day.
- Hospital A is a 200-bed hospital and hospital B is a 300 bed hospital.
- d is the relative size of the two sampling frames = 300/200 = 1.5
- We want to detect a difference of $\rho = \lambda_1 / \lambda_2 = 2$ in the ratio of the rates of infection of the two hospitals.
- f = sqrt[(1+1.5*2)/(1.5+2)] = 1.069
- The number of events needed to be observed in population 1 to detect a rate of ρ with significance level α and power 1-β is

$$n_1 = \frac{(\rho/d+1)(fz_{1-\alpha/2}+z_\beta)^2}{(\rho-1)^2} = \frac{(2/3+1)(1.069*1.96+.84)^2}{(2-1)^2} = 14.36$$

- The size of population 1 needed is: $r_1 = n_1/\lambda_1 = 14.36/.001 = 14{,}360$
- The size of population 2 needed is $r_2 = d\, r_1 = 1.5 * 14{,}360 = 21{,}540$

Assumes: $\alpha = 0.05$, $\beta = 0.2$, $d = 1$
Varying λ_1 and ρ



Time to Event Studies

- The time to an event may be specified when we are not only interested in *whether* an event occurs but in *when* it occurs.
- Events could be: mortality, time to remission, time to viral load suppression, time to stop smoking.
- In some cases, the occurrence of the event is of interest (eg. 30 day mortality, smoking cessation, ever suppressing viral load while on treatment regimen) and then the sample size should be determined with a comparison of proportions.

Time to Event Studies

- Assume patients are recruited over an accrual period A and followed for an additional follow-up time F
- The median survival times for patients on treatments 1 and 2 are m_1 and m_2.
- $R = m_2/m_1$= relative hazard

Sample size for Time to event outcomes

- To compute sample size, you need to specify:
  - The estimated proportions of subjects in each group who are "endpoint-free" at a fixed time
  - or – the estimated hazard ratio (R) and the median survival time for the control group (m_1)
  - or – the estimated median survival times for each group (m_1 and m_2)

Sample size using proportions of patients with events by time *t*

- This calculation assumes that patients are followed for a **fixed length of time** (t) and that the hazard ratio (R) is constant over time. If p_i is the proportion of subjects who are endpoint free at time t for group i, then

$$R = \frac{\ln p_1}{\ln p_2} = \frac{\ln(\exp(-\lambda_1))}{\ln(\exp(-\lambda_2))} = \frac{\lambda_1}{\lambda_2}$$

- And the sample size per group is

$$n = \frac{(z_{1-\alpha/2} + z_\beta)^2 \left[\frac{1}{1-p_1} + \frac{1}{1-p_2} \right]}{(\log R)^2}$$

Schoenfeld & Richter
Biometrics 1982

Example

- We want to determine if a new drug for the treatment of lung cancer lengthens survival time. All patients in the trial will be followed for 2 years. If the 2-year survival rate under standard therapy is p_1=0.25, the number of patients **per group** needed to detect R=1.5 with α=.05 and 80% power is

$$n = \frac{(1.96+0.84)^2\left[\frac{1}{1-.25}+\frac{1}{1-.397}\right]}{(\ln 1.5)^2} = 143$$

since $$R = 1.5 = \frac{\ln(.25)}{\ln(p_2)} \Rightarrow p_2 = .397$$

Time to event calculations

- Note that the sample size depends on the number of expected events, and the actual follow-up time does not appear in the equation.
- The follow-up time is only important in relation to the median survival time or the expected proportion event free at time *t*, so that the expected number of events during the study can be calculated.
- When patients are not followed for a fixed length of time, the time to accrue patients also needs to be considered.
- Sample size may also need to be adjusted for drop-outs.
- The program "Power" adjusts for follow-up time and accrual time.

Ordinal Data

- Sometimes the outcome is an ordered categorical variable, such as pain, quality of life, etc.
- This outcome would be compared between treatment groups with a Wilcoxon rank sum test.
- In order to calculate the required sample size, you need to estimate the proportions of each treatment group expected to be in each category.
- If the outcome is continuous, then it needs to be discretized into categories.

Longitudinal Outcomes

- Comparison of a **rate of change** in a continuous outcome between two groups
  - $Y_{ij} = \beta_{0A} + \beta_{1A} x_{ij} + e_{ij}$ in group A
  - $Y_{ij} = \beta_{0B} + \beta_{1B} x_{ij} + e_{ij}$ in group B

- The required number of subjects per group to detect a difference in the rate of change of $d = \beta_{1A} - \beta_{1B}$ with significance level α and power $1-\beta$ is

$$m = 2(z_{1-\alpha} + z_{\beta})^2 \sigma^2 (1-\rho) / n s_x^2 d^2$$

where
ρ is the correlation among repeated measures,
n is the number of observations per individual
and s_x^2 is the within-subject variance of the x's

$$s_x^2 = \Sigma_j (x_j - \bar{x})^2 / n$$

Diggle, Liang, Zeger. Analysis of Longitudinal Data.

Example of Sample size for Rate of Change

- Randomized trial of two weight loss regimens.
- Comparing rate of change in weight.
- Weight is measured at months 3, 6, 9.
- Guesstimate ρ to be 0.5.
- Expect range of rates of change to be from –3kg per month to 1 kg per month.
- Guesstimate σ of rate of change to be 1 kg/month.
- Clinically significant difference in change per month is determined to be 0.1kg/month.
- Significance level = .05, 80% power.

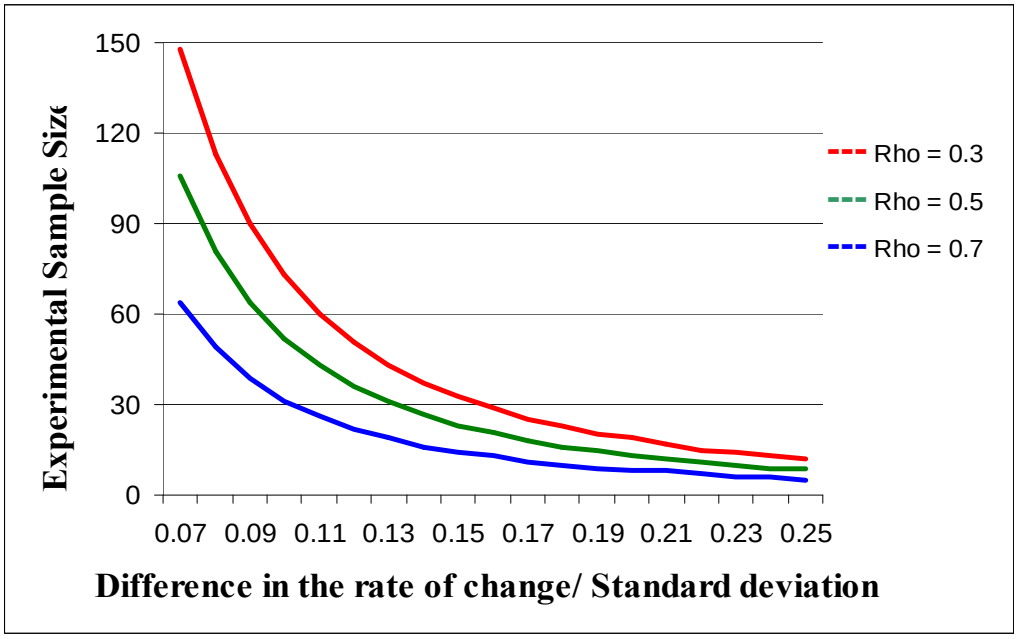
$\Rightarrow$ Sample size is 44 participants per group.

Factors Affecting Sample Size

Visits	ρ=0.3	ρ=0.5	ρ=0.7
3, 6	244	174	104
3, 6, 9	61	**44**	26
3, 6, 9, 12	24	17	10
1,2,3,4	218	157	94

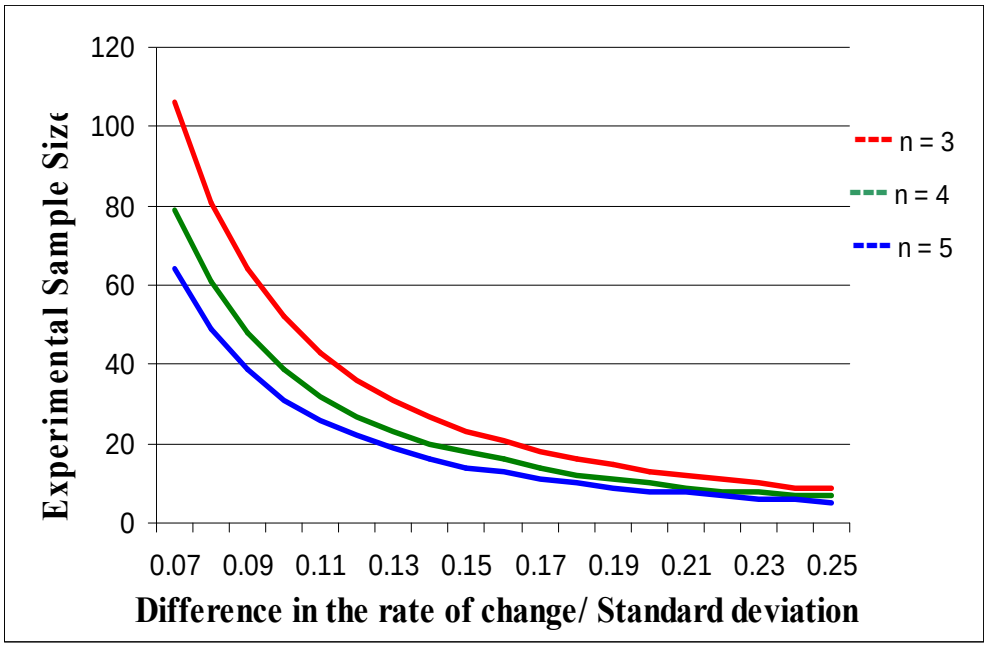
Assumes: $\alpha = 0.05$, $\beta = 0.2$, $n = 3$, $S_x^2 = 4$

Varying ρ and d/σ



Assumes: $\alpha = 0.05$, $\beta = 0.2$, $\rho = 0.5$, $S_x^2 = 4$

Varying n and d/σ



Longitudinal Outcomes (ctd)

- Comparison of **average responses**
- $Y_{ij} = \beta_0 + \beta_1 x + e_{ij}$ where β_1 is a binary treatment indicator
- The required number of subjects per group to detect a difference in the average response d with significance level α and power 1- β is

$$m = 2(z_{1-\alpha} + z_{\beta})^2 \sigma^2 \{1 + (n-1)\rho\} / nd^2$$

where

- ρ is the correlation among repeated measures within individuals,
- σ^2 is the variance of the e_{ij}'s
- n is the number of observations per individual

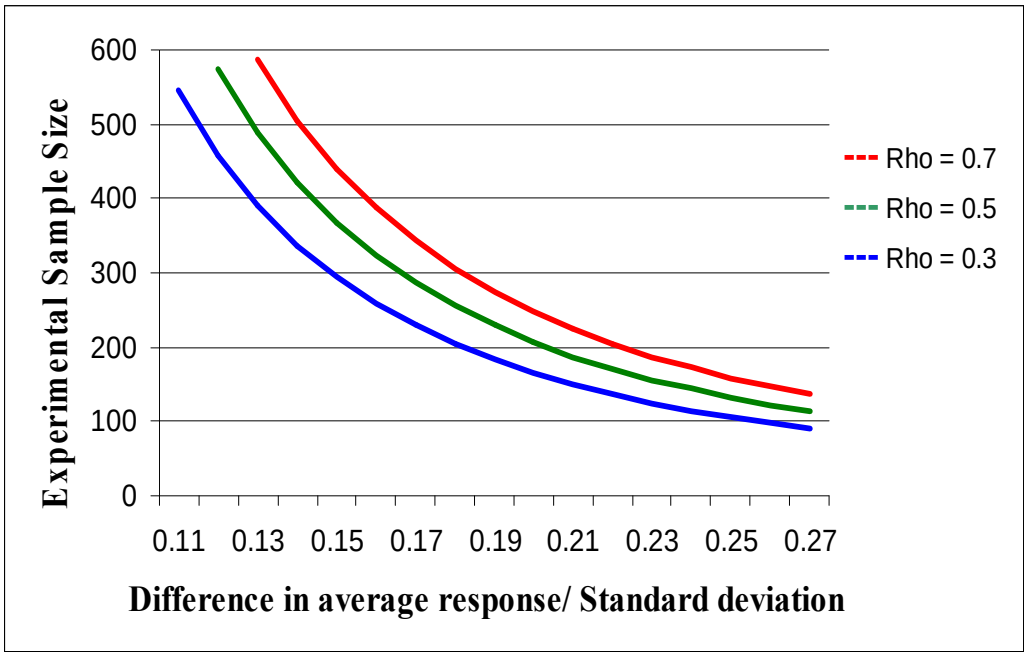
Example of Sample size for Average Response

- Again, randomized trial of two weight loss regimens.
- Comparing average weight over the follow-up period.
- Weight is measured at months 3, 6, 9.
- Guesstimate ρ to be 0.5.
- Expect average weights to range from 60 to 90 kg.
- Guesstimate σ to be 15 kg.
- Clinically significant difference in average weight is determined to be 5 kg.
- Significance level = .05, 80% power.

$\Rightarrow$ Sample size is 94 participants per group.

# of Visits	ρ = 0.3	ρ = 0.5	ρ = 0.7
2	92	106	120
3	75	94	113
4	67	88	109
10	52	78	103

Assumes: $\alpha = 0.05$, $\beta = 0.2$, $n = 3$
Varying ρ and d/σ



Longitudinal Binary Outcomes

- Comparison of mean proportions
- $\Pr(Y_{ij} = 1) = p_A$ in group A and p_B in group B
- The number of subjects **per group** required to detect a difference $d = p_A - p_B$ with significance level α and power $1-\beta$ is:

$$m = [z_{1-\alpha}\{2\bar{p}\bar{q}(1+(n-1)\rho)\}^{1/2} + z_{\beta}\{(1+(n-1)\rho)(p_A q_A + p_B q_B)\}^{1/2}]^2 / nd^2$$

where $\bar{p} = (p_A + p_B)/2$, $\bar{q} = 1 - \bar{p}$ and ρ is the correlation among repeat observations within individuals

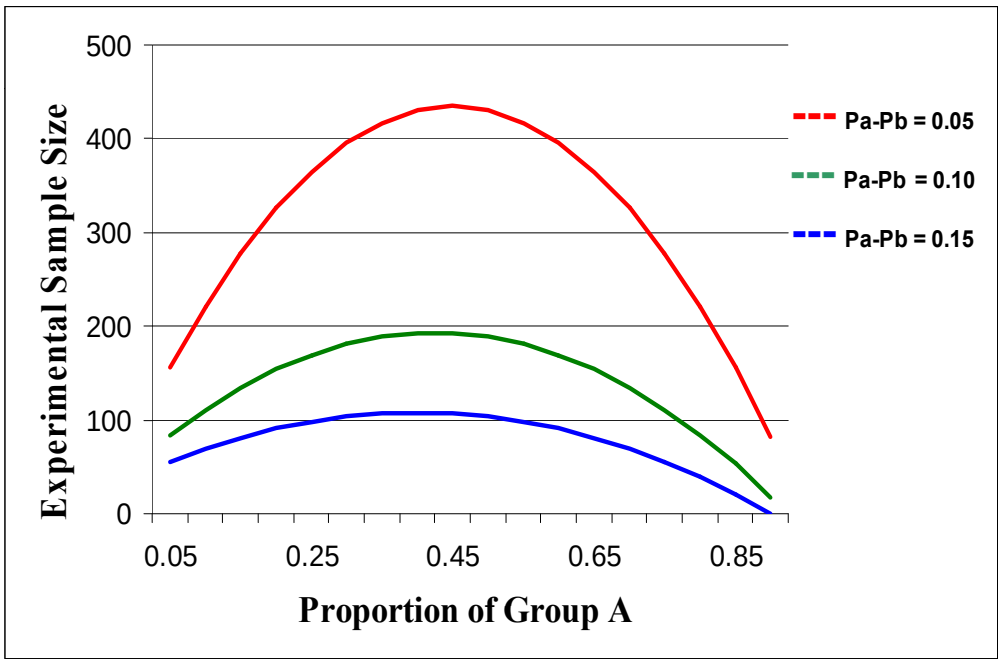
Example of Binary Longitudinal Outcome

- Randomized trial of two smoking cessation programs.
- Outcome = average proportion of smokers during the follow-up period.
- Clinically important difference = 10% (from 60% to 70%).
- Assume correlation among visits = .5.
- n = # follow-up visits =5.
- Significance level =.05, 80% power.
- Sample size per group = 214.

Longitudinal Binary Sample Size

Detectable Difference p_A-p_B	# of Visits	ρ = 0.3	ρ = 0.5	ρ = 0.7
.1	3	189	237	285
.1	5	157	214	271
.1	10	132	196	260
.15	3	81	101	121
.15	5	67	91	115
.15	10	56	84	111

Assumes: $\alpha = 0.05$, $\beta = 0.2$, $n = 3$
Varying P_A and $(P_A - P_B)$



Website for Sample Size Calculation for Longitudinal Designs

- www.healthstats.org
- "RMASS" under "Projects"
- Allows
  - Unequal group sizes
  - Attrition, which can vary over time
  - Different correlation structures
    - Exchangeable, stationary AR1, toeplitz/banded
  - Clustering of centers

Sample Size for Joint Modeling of Longitudinal and Survival Data

- Chen LM, Ibrahim JG, Chu H. Sample size and power determination in joint modeling of longitudinal and survival data. Statistics in Medicine. 2011; 30: 2295-2309.
- Assumes
  - survival follows exponential distribution
  - Longitudinal outcome follows a general polynomial model
- Closed form sample size formula

Crossover Trial

- A two period crossover trial is one in which patients receive both treatments but are randomized to the order in which they receive the treatments. Patients either receive A, then B or B, then A.
- The primary outcome is the within-patient difference in the outcome between the two treatment periods.

Crossover trial sample size

The variance, σ^2, has two components:

$$\sigma^2 = \sigma_s^2 + \sigma_e^2$$

Where σ_s^2 is the subject-to-subject variability in the level of response and

σ_e^2 measures variability due to random measurement errors and other intra-subject factors.

The number of subjects required for a crossover trial, n*, is related to the number of subjects required for a parallel group trial, n, as follows:

$$n^* = n\,(1-R)/2$$

where $R = \sigma_s^2/(\sigma_s^2 + \sigma_e^2)$

Cross-over trial - efficiency

- Even if R is as low as 0.5, the total number of subjects required for a cross-over trial is ¼ of the number of subjects required for a parallel group study.
- Because each subject is measured twice in a two period crossover study, the total number of measurements on the 2n* subjects in it is

$$2 \times 2n^* = 2n(1-R),$$

A fraction (1-R) of the total 2n measurements in the parallel group study.

Factorial Design

- Factorial designs can be used to evaluate the effects of more than one intervention at the same time.
- If you are interested in comparing interventions A and B with placebo, a 2x2 factorial design may be more efficient than a 3 arm trial.
- In a factorial design, participants are assigned to interventions as follows:

		Intervention A	
		Yes	No
Intervention B	Yes		
	No		

Sample Size Calculations for a Factorial Design

The most common procedure is to calculate the sample size required to detect the effect of interest for each of the interventions as if it were to be compared to a control group in a parallel group design. The sample size is then the larger of the two sample sizes.

Montgomery AA, Peters TJ, Little P. Design, analysis and presentation of factorial randomized controlled trials. BMC Medical Research Methodology 2003; 3:26.

Factorial Design

- Eg. We are interested in the effects of two interventions to improve adherence to medication, we might use a factorial design.
- If the effect size of interest for intervention A is .35 SD then 243 pts per group are required to compare A to placebo.
- If the effect size of interest for intervention B is .3 SD, then 332 pts per group are required to compare B to placebo.
- The sample size for a 3 arm trial would be 243+332+332 = 907.
- For a factorial design, the sample size required is 332+332 = 664.

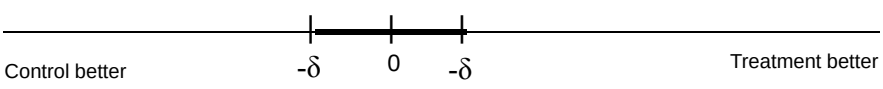
Factorial Design

- **This procedure assumes that the effects of the two interventions are independent**.
- Not always a reasonable assumption.
- If you wish to detect an interaction of the same magnitude as the main effect, a 4-fold increase in the sample size is required.
- With no increase in sample size, the interaction needs to be twice as large as the main effect in order to be detected with the same power.

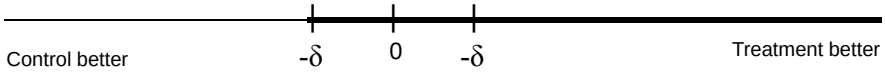
Superiority, Equivalence and Noninferiority Trials

Superiority

Equivalence



Non inferior



Hypothesis Tests

- Superiority Trials
  - H_0: $\mu_A = \mu_B$
  - H_A: $\mu_A \neq \mu_B$
- Equivalence Trials
  - H_0: $\mu_A \neq \mu_B$
  - H_A: $\mu_A = \mu_B$
- Non-inferiority Trials
  - H_0: $\mu_A - \mu_B \leq -d$
  - H_A: $\mu_A - \mu_B \geq -d$

  where d is the largest difference that is clinically acceptable.

Motivation for Equivalence/ Non-Inferiority Trials

- Situations where there is a currently available treatment and we want to compare a new treatment that is cheaper/easier to take/has fewer side effects but want to be sure that the new treatment is "as effective" as the standard of care.
- Used when placebo-controlled trials are not appropriate.

Equivalence Trial

- It is impossible to prove complete equivalence.
- Usually a predefined clinically important difference of treatment effects, Δ, is specified.
- Equivalence is implied if the difference between treatments is less than or equal to Δ.
- The null hypothesis is then $H_o: |\mu_T - \mu_S| > \Delta$ and the alternative hypothesis is $H_A: |\mu_T - \mu_S| <= \Delta$
- In equivalence trials, it is common to specify power of 90%.
- Note that we are now doing a one-sided test.

Equivalence Trial Sample Size Formulae

Comparison of proportions:

$$n = (z_{1-\alpha} + z_{1-\beta})^2 [p_A(1-p_A) + p_B(1-p_B)] / \Delta^2$$

Where p_A and p_B are the response rates of standard and new treatments, respectively, and Δ is the hypothesized difference.

Frequently, it is assumed that $p_A = p_B$.

Comparison of means:

$$n = 2(z_{1-\alpha} + z_{1-\beta})^2 \sigma^2 / \Delta^2$$

Where σ is the standard deviation of treatment effect, and Δ is the hypothesized difference.

Example of Equivalence Trial Sample Size Calculation

- We are comparing two treatments for HIV.
- The outcome is the proportion of patients with viral load < level of detection
- We assume that the proportion is 80% in both groups and want to show that the difference is not more than 10%.

$$n = (z_{1-\alpha} + z_{1-\beta})^2[p_A(1-p_A) + p_B(1-p_B)]/(|p_A - p_B| - \Delta)^2$$

$$N = (1.645 + 1.28)^2[.8*.2 + .8*.2]/(.1)^2 = 274$$

Under the usual setting, $N=(1.96+.84)^2[.8*.2+.7*.3]/(.1)^2 = 290$ or $N=(1.96+.84)^2[.8*.2+.9*.1]/(.1)^2 = 196$

Non-inferiority trial

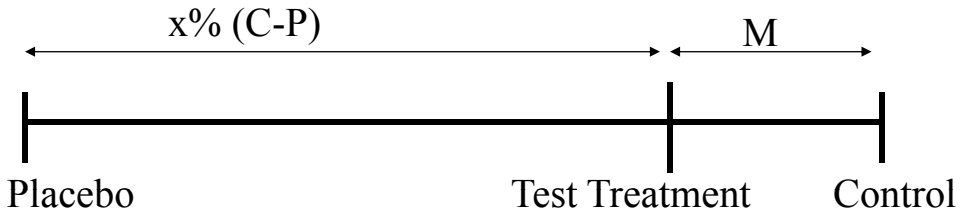
- The goal of a non-inferiority trial is to show that the new treatment is **not worse than** the standard of care by more than a pre-specified small amount.
- Two results of interest from the clinical trial
  - The point estimate of the effect.
  - The lower limit of the 95% CI.

Statistics in Medicine 2003; **22**: 169-186. Non-inferiority trials: design concepts and issues – the encounters of academic consultants in statistics.

Criteria to Show that an Experimental Treatment is non-inferior to Standard Treatment

- There must be convincing prior evidence that the standard treatment is better than placebo or no treatment.
- It must be clear that the standard treatment is effective in the current trial.
- The current study must be capable of distinguishing between an effective and an ineffective treatment.
  - Adequate sample size
  - No excessive dropout, treatment crossover, etc

Snapinn S. Noninferiority Trials. Curr Control Trials Cardiovasc Med 2000; 1:19-21



- We want to be sure that the test treatment is within a non-inferiority margin, M, of the control treatment.
- We also want to preserve a specified percentage of the control vs placebo effect (say 80% or 50%).
- There is concern about "biocreep", ie. that if we lose a proportion of the treatment effect with each subsequent clinical trial, eventually the new treatment may not be any better than placebo.

Inferiority Trial: Example

- Randomized trial of two antibiotics to treat afebrile neutropenia.
- Outcome = clearance of fever by 4 days.
- Success rate = 50% with standard of care.
- Company wants to show that their new product is "equivalent" to the standard of care.
- Can infer that they want to show the new product is not worse than the standard of care.
- Want to show that success rate is at least 40%.

Sample size

$$n = (z_{1-\alpha} + z_{1-\beta})^2 [p_A(1-p_A) + p_B(1-p_B)] / \Delta^2$$

$$N = (1.645 + 1.28)^2[.5*.5 + .5*.5]/(.1)^2 = 428$$

The usual formula for a two sample test of proportions is:

$$n = \frac{[p_1(1-p_1) + p_2(1-p_2)](z_{1-\alpha/2} + z_{1-\beta})^2}{(p_1 - p_2)^2}$$

$$= [.5*.5 + .4*.6](1.96 + .84)^2/.1^2 = 384$$

Proc POWER in SAS

Calculates sample size and power for*:*

- *t* tests for means
- equivalence tests for means
- confidence intervals for means
- tests of binomial proportions
- multiple regression
- tests of correlation and partial correlation
- one-way analysis of variance
- rank tests for comparing two survival curves
- Logistic regression with binary response
- Wilcoxon rank sum test

PROC POWER
sample size for two sample t test

```
ods html;
ods graphics on;
proc power;
  twosamplemeans
  groupmeans = (300 320) (300 330) (300 340)
  stddev = 20 25 30 35
  ntotal = .
  power = 0.8;
  run;
ods graphics off;
ods html close;
```

The POWER Procedure
Two-sample t Test for Mean Difference

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Nominal Power	0.8
Number of Sides	2
Null Difference	0
Alpha	0.05
Group 1 Weight	1
Group 2 Weight	1

Computed N Total

Index	Mean1	Mean2	Std Dev	Actual Power	N Total
1	300	320	20	0.807	34
2	300	320	25	0.807	52
3	300	320	30	0.808	74
4	300	320	35	0.808	100
5	300	330	20	0.848	18
6	300	330	25	0.802	24
7	300	330	30	0.807	34
8	300	330	35	0.811	46
9	300	340	20	0.876	12
10	300	340	25	0.845	16
11	300	340	30	0.805	20
12	300	340	35	0.829	28

PROC POWER
Sample Size for Two Survival Curves

```
ods html;
ods graphics on;
proc power;
     twosamplesurvival test=logrank
        groupmedsurvtimes = (5 6) (5 7)
        accrualtime = 1 2
        followuptime = 3
        groupmedlosstimes = (10 20) (10 15)
        power = 0.8
        npergroup = .;
  run;
ods graphics off;
ods html close;
```

The POWER Procedure
Log-Rank Test for Two Survival Curves

Fixed Scenario Elements	
Method	Lakatos normal approximation
Form of Survival Curve 1	Exponential
Form of Survival Curve 2	Exponential
Follow-up Time	3
Nominal Power	0.8
Number of Sides	2
Number of Time Sub-Intervals	12
Alpha	0.05

Computed N Per Group							
Index	Accrual Time	Med Surv Time 1	Med Surv Time 2	Median Loss Time 1	Median Loss Time 2	Actual Power	N Per Group
1	1	5	6	10	12	0.800	1432
2	1	5	6	10	15	0.800	1418
3	1	5	7	10	12	0.800	448
4	1	5	7	10	15	0.801	444
5	2	5	6	10	12	0.800	1316
6	2	5	6	10	15	0.800	1302
7	2	5	7	10	12	0.800	411
8	2	5	7	10	15	0.801	407

PROC POWER
Analysis of Variance

```
ods html;
ods graphics on;
proc power;
     onewayanova test=overall
       groupmeans = (3 7 8) (3 6 8) (4 7 8) (4 7 8)
       stddev = 4 5 6
       npergroup = .
       power = .80;
  run;
ods graphics off;
ods html close;
```

Computed N Per Group						
Index	Means			Std Dev	Actual Power	N Per Group
1	3	7	8	4	0.833	13
2	3	7	8	5	0.818	19
3	3	7	8	6	0.803	26
4	3	6	8	4	0.825	14
5	3	6	8	5	0.820	21
6	3	6	8	6	0.809	29
7	4	7	8	4	0.804	19
8	4	7	8	5	0.803	29
9	4	7	8	6	0.810	42
10	4	7	8	4	0.804	19
11	4	7	8	5	0.803	29
12	4	7	8	6	0.810	42

Statement for Grant

- "A sample size of XX patients per group is required to detect a difference of XX (from XX to XX) between two groups in the primary outcome XXX with an XXX statistical test, assuming a significance level of .05, 80% power and a standard deviation of XX."
- "The sample size was further adjusted to account for an XX% expected dropout rate."
- "The sample size was further adjusted to account for an XX% expected contamination rate."

Feasibility

- **Rule of Thumb**: Recruitment is always slower than you expect.
- Need to demonstrate feasibility for CIHR/others by counting numbers of patients currently available/presenting annually to the clinic(s) who meet inclusion and exclusion criteria.
- Need to assume the number of patients expected to be willing to participate.
- Time, inconvenience, invasiveness, degree of illness, cultural issues may be barriers to participation.

Pilot Studies

- Small study conducted prior to ultimate trial
- Goals of pilot study are different from those of the ultimate trial
- Don't do a formal sample size calculation
- Analysis of a pilot study should be descriptive
- Not powered to draw conclusions

Potential Goals of Pilot Studies

- Collect data on primary outcome for determination of sample size
- Assess feasibility of recruitment
- Inclusion/exclusion criteria
- Pilot case report forms and
- Test randomization procedures
- Storage and testing of equipment
- Training of staff
- Assess acceptability of the intervention

External vs Internal Pilot Studies

- External studies – data not included in analysis of ultimate trial – especially if study design changed in any way
- Internal studies – data is included in analysis of ultimate trial

Grant Application for Pilot Study

- CIHR often requires results of pilot study before funding ultimate trial
- Grant application should state objectives of pilot study as well as objectives of ultimate trial
- Don't need a formal sample size calculation but do need to justify your choice of number of participants

Results of a Pilot Study

- Should be published
- Should be identified as a pilot study
- Should focus on objectives of pilot study, not statistical significance of outcome for ultimate trial

References

- Thabane et al. A tutorial on pilot studies: the what, why and how. BMC Medical Research Methodology. 2010, 10:1.
- Lancaster et al. Design and analysis of pilot studies: recommendations for good practice. J Evaluation Clinical Practice. 2004; 10,2,307-12.