# Determining Sample Size

## How many patients do we need in our study?

# GOALS

- Review of the inputs for determining sample size
- Compare sample sizes for Parallel, Crossover and Factorial designs
- Increase understanding of the impact of assumptions and practical constraints
- Show the effect on sample size of uncertainty in the inputs and outline approaches to deal with uncertainty

# Factors that affect sample size

- Study design
  - parallel, crossover, factorial, nesting structures
- Nature of the endpoint (measured, event, event time)
  - for events, “n” might be “# of events”
- Analysis plan
- Monitoring plan
- Inherent variability of response and co-variability of responses
- Adherence and dropouts
- Goals: test size&power; CI length, pr(correct decision)...
- Balancing biologically reasonable assumptions, time, effort, expense and other practical constraints
- …….

Generally, you need to consider a range
of sample size values linked to assumptions

# Focused factors for Hypothesis Testing sample size
## (in the context of the general list)

**Goal is to:** have sufficient power to choose between two simple hypotheses

- Variability of the "building block" response ($\sigma^2$)
- Type I error ($\alpha$), significant level
- Type II error ($\beta$), power = 1- $\beta$

- Size of minimal difference considered important ($\Delta$)

# A review of hypothesis test type I and II errors and power

| Conclusion from the analysis | "Truth" | |
|---|---|---|
| | $H_o$ true | $H_a$ true |
| Reject $H_o$ | Type I error ($\alpha$) | Correct conclusion (Power=1-$\beta$) |
| Fail to reject $H_o$ | Correct conclusion | Type II error ($\beta$) |

# Hypothesis testing sample size (single group, or paired differences)

$$n = \frac{\sigma^2 \times (Z_{\alpha/2} + Z_\beta)^2}{\Delta^2} = \left(\frac{\sigma}{\Delta}\right)^2 \times (Z_{\alpha/2} + Z_\beta)^2$$

Required sample size depends inversely on the square of the effect size

**Effect size =** $\Delta$ (sometimes $\left(\frac{\Delta}{\sigma}\right)$ is referred to as the effect size)

Decreasing it by a factor of 2 increases n by a factor of 4

# Sample size formula for a two group comparison

$$n \text{ per group} = \frac{2 \times \textbf{Variability} \times [z_{\alpha/2} + z_{\beta}]^2}{\Delta^2}$$

What causes sample size ⬆ ?

- ⬆ Variability
- ⬇ Type I error
- ⬇ Type II error ⇔ ⬆ Power
- ⬇ $|\Delta|$

# Example: Prophylaxis for Toxoplasmosis*

- Primary endpoint: Toxoplasmic encephalitis (TE)
- Control (placebo) group event rate: 30% in 2.5 yrs
- Treatment (pyrimethamine) group event rate: 15% in 2.5 yrs

  ➔ treatment effect: $\Delta$ = 15 percentage point rate reduction
- Death rate for causes unrelated to TE: 33% in 2.5 years
- Dropouts
- Adherence
- Type I rate and power:

  $\alpha$ = 0.05 (2-sided)

  1-$\beta$ = 0.80
- .....

These assumptions/guesses were based on very little information

* Jacobson M, Besch C, Child C, et al. Eur J Clin Microbiol Infect Dis., 10: 195-8, 1991

# Influence of Effect Size Δ
# TOXO Sample Size

- With $\alpha = 0.05$ (2-sided) and power: $1-\beta = 0.80$

| Event Rate (%) | | Δ (%) | Percent Reduction | Sample Size |
|---|---|---|---|---|
| Placebo | Pyrimethamine | | | |
| 30 | 10 | 20 | 67 | 130 |
| 30 | 15 | 15 | 50 | 265 |
| 30 | 17 | 12 | 42 | 400 |
| 30 | 20 | 10 | 33 | 650 |

# A few sample size formulas, (there are thousands of these!)

- Assuming equal sample sizes $n_1=n_2=n$

| Population Value $\Delta$ | Estimator | Sample Size |
|---|---|---|
| $\mu_2 - \mu_1$ | $\overline{X}_2 - \overline{X}_1$ | $n = \frac{(z_{\alpha/2} + z_{\beta})^2(\sigma_1^2 + \sigma_2^2)}{\Delta^2}$ |
| $p_2 - p_1$ | $\hat{p}_2 - \hat{p}_1$ | $n = \frac{(z_{\alpha/2}\sqrt{2\bar{p}\bar{q}} + z_{\beta}\sqrt{\hat{p}_1\hat{q}_1 + \hat{p}_2\hat{q}_2})^2}{\Delta^2}$ |

# Focused factors for Confidence Interval length sample size

- Most items on the "factors" slide
- Variance of a single observation ($\sigma$)
- The maximal CI length (L)
  - some prefer to use the "margin of error," the half-length: d = L/2
- The coverage probability (1-$\alpha$) for a a two-sided interval
  - you can do a one-sided interval

# CI length sample size

$$n = 4Z^2_{\alpha/2}\left(\frac{\sigma}{L}\right)^2$$

- Required sample size is inversely related to the square of the maximal length
- Decreasing “L” by a factor of 3 increases n by a factor of 9

# USA TODAY/CNN/Gallup Poll results*

## Poll results are based on telephone interviews with "National adults" conducted April 1-2, 2005

Q26. As a result of the recent rise in gas prices, would you say you have or have not done each of the following?

| 2005 APR 1-2 (sorted by Yes, have) | Yes, have | No, have not |
|---|---|---|
| Seriously considered getting more fuel-efficient car the next time you buy a vehicle | 57 | 42 |

*For results based on the total sample of "National Adults," one can say with **95% confidence** that the margin of sampling error is **±3 percentage points**.*

**What does this mean?**

If you are conducting a survey requiring the same precision, how many people do you need to interview?

* http://www.usatoday.com/news/polls/tables/live/2005-04-03-poll.htm

# Sample size for estimating a proportion *p*

- Estimate ***p*** with 3% margin of error (**d=*0.03*, *L = .06***)

$$\mathbf{SE}(\hat{p}) = \sqrt{pq/n}$$

$$\mathbf{d} = 1.96\ \mathbf{SE}(\hat{p})$$

$$3\% = 1.96\sqrt{pq/n}$$

$$n = (1.96/0.03)^2 pq$$

**Assume $p = q = 0.5$ (most conservative)**

$$n = 1068 = (1.96/0.03)^2(0.5)(0.5)$$

# General CI length sample size

## when the Normal distribution is a good approximation

- Estimate the effect of interest using "**estimate**"
  - Difference in means
  - Regression slope
- Use this estimate for a CI by computing

  CI = **estimate** $\pm$ 1.96$\times$se(estimate)
- Find a sample size (and other features) so that:

$$\text{Var}(\text{estimate} \mid n_1, n_2, \ldots, \text{design}, \text{analysis}, \ldots) \leq \frac{L^2}{4Z^2_{\alpha/2}}$$

**"Var" depends on all aspects of the design and analysis**

# Sample size for a rare outcome or for survival analysis

- In these situations, the **number of events** needs to be sufficiently large, not simply "n"
- For example, if we need a total of 50 events, with event probability "p" we need n = 50/p to, on average, generate the required number of events:

| Event probability | Required "Sample Size" |
|---|---|
| 1/10 | 500 |
| 1/100 | 5,000 |
| 1/1000 | 50,000 |

- These sample sizes give the required **expected** number of events, but do not guarantee that the required number will occur (more on this later)

# Sample size for a rare outcome or for survival analysis

## (Optional, technical details)

$$H_o : \frac{p_2}{p_1} = 1 \quad \text{or } p_2 - p_1 = 0$$

$$H_a : \frac{p_2}{p_1} = R \quad \text{or } p_2 - p_1 = \Delta = p_1(R-1)$$

so,

$$n_1 = n_2 \approx \frac{2p_1(1-p_1)[z_{\alpha/2} + z_\beta]^2}{p_1^2(R-1)^2}$$

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

# Sample size for a rare outcome or for survival analysis

## (Optional technical details, continued)

$p_1$ is near 0, so

$$n_1 = n_2 \approx \frac{2[z_{\alpha/2} + z_\beta]^2}{p_1(\mathrm{R}-1)^2}$$

"bring over" $p_1$

$$(n_1 + n_2)p_1 \approx \frac{4[z_{\alpha/2} + z_\beta]^2}{(\mathrm{R}-1)^2}$$

Expected events = constant $\left(= \frac{4[z_{\alpha/2} + z_\beta]^2}{(\mathrm{R}-1)^2}\right)$

# Sample Size in a Factorial Design

# Example of the efficiency of a factorial design

## Aspirin, sulfinpyrazone, or both in unstable angina. Results of a Canadian multicenter trial*

- A randomized trial of 555 patients, hospitalized in coronary care units with unstable angina
- Primary outcome was cardiac death or nonfatal myocardial infarction
- Patients received one of the four treatment combinations: aspirin, sulfinpyrazone, both or neither
  - Aspirin was included only when the study statistician, at the last minute, promoted the factorial design

* Cairns, J. et al., N Engl J Med. 1985;313:1369-75

# Sample Size for a Factorial Design

## Results from the Canadian Aspirin Study

**Number of cardiac deaths and nonfatal MIs in two years** (number of patients)

| | Placebo | ASA | Total |
|---|---|---|---|
| **Placebo** | 18 (139) | 8 (139) | 26 (278) |
| **Sulfinpyrazone** | 18 (140) | 9 (137) | 27 (277) |
| **Total** | 36 (279) | 17 (276) | 17 (555) |

- Suppose we are designing a parallel study to detect a 50% reduction in the primary outcome with $\alpha$=0.05 and (1-$\beta$)=0.8
- Assume $p_1$ = 15% (observed rate = 18/139 = 13%)
- $\Delta = p_1 - p_2 = 15\% - 7.5\% = 7.5\%$
- A parallel design requires 277 patients for each group; a total of 554 patients to evaluate a single treatment
- The factorial design delivers "2 for the price of 1"
  - Assuming assumptions are satisfied!!!

# Sample Size in a Crossover Design

## A crossover study comparing bronchodilators

**Peak expiratory flow (L/min) 8 hours after treatment**

| PatientID | Formoterol | Salbutamol | Difference (Formoterol-Salbutamol) |
|---|---|---|---|
| 1 | 310 | 270 | 40 |
| 4 | 310 | 260 | 50 |
| 6 | 370 | 300 | 70 |
| 7 | 410 | 390 | 20 |
| 10 | 250 | 210 | 40 |
| 11 | 380 | 350 | 30 |
| 14 | 330 | 365 | -35 |
| 2 | 385 | 370 | 15 |
| 3 | 400 | 310 | 90 |
| 5 | 410 | 380 | 30 |
| 9 | 320 | 290 | 30 |
| 12 | 340 | 260 | 80 |
| 13 | 220 | 90 | 130 |
| **Variance** | **3559** | **6866** | **1648** |

# Sample Size in Crossover Design (cont'd)

**Crossover: $n$ patients in the study**

$$\text{Var}_{\text{xover}}(\hat{\beta}) = \frac{\sigma^2_{\text{diff}}}{n}$$

**Parallel: $n$ patients in each treatment group**

**(a total of $2n$ patients in the study)**

$$\text{Var}_{\text{para}}(\hat{\beta}) = \frac{\sigma^2_{\text{for}} + \sigma^2_{\text{sal}}}{n}$$

$$\frac{\sigma^2_{\text{for}} + \sigma^2_{\text{sal}}}{\sigma^2_{\text{diff}}} = \frac{3559 + 6866}{1648} = 6.3$$

- The variance of the estimated treatment effect in a parallel design is 6 times larger than that in a crossover study
  - Parallel study needs $\boldsymbol{n}$ to be 6 times larger
  - Parallel study needs 12.6 times more total patients to detect the same effect with the same size and power
- But, the crossover design needs two measurements for each patient, whereas parallel design needs only one

# Factors that influence Variability

## Biological variability

- Depends on the target population
- Depends on the choice of outcomes
  - Outcome may have high variability even within a homogenous target population
  - Number of primary events (duration of follow-up & competing events)
- Missing data (Losses to follow-up)
- Measurement variability
  - All factors identified in the measurement module
  - Measurement error

# Choice of $\alpha$

## One-sided vs two sided test

- Often $\alpha$ is specified at 0.05
  - For a two-sided test, $z_{\alpha/2} = 1.96$
  - For a one-sided test, $z_{\alpha} = 1.64$
- If we conduct a one-sided test,
  - for the same sample size, power increases
  - the required sample size to attain the same power decreases
- If we are only interested in the positive treatment effect, why waste $\alpha$ on the "other side?"

# Why two-sided?

- Protects the study against the unexpected
- Even if you aren't interested in the "other side" and design for a one-sided test (or CI), there is a power penalty in that someone wanting to conduct a two-sided analysis will have reduced power
  - They would have needed a larger study to maintain an overall 0.05 and desired power, but are stuck with what you give them

**Reality test: Can you honestly say that if results are strongly in the "other direction" you are going to ignore them and not report them?**

# Cardiac Arrhythmia Suppression Trial (CAST)

- CAST demonstrated that ventricular arrhythmia suppression is a failed surrogate for death from arrhythmia

BUT

- Many thought it was unethical to do the study and "subject" participants to the placebo
- There was strong pressure to design CAST as a one-tailed test
- The DSMB statistician argued for two-sided, but eventually settled for one-sided at $\alpha = 0.025$
- "Lets require very strong evidence" of the beneficial effect
  - not initially designed to report that the treatment can cause harm
- Of course, the other 0.025 was held in reserve to have some available for the harm conclusion
- Results showed that patients treated with active drug had a higher rate of death from arrhythmias than those taking placebo

# Considerations in Specifying the Treatment Effect ($\Delta$)

- Smallest difference of clinical significance/importance
- Stage of research
- Realistic estimates based on:
  - Previous research
  - Expected event rate
  - Expected non-compliance rate
  - Expected switchover rate
- "Instrumentation" variability
- Implications on sample size
- .....

# Factors which influence "Realized Δ"

- Due to the squared effect of Δ on sample size, it is important to control in the design and/or incorporate in the sample size assessment:
  - Measurement variance
  - Non-compliance
  - Switchover from the assigned treatment regime
  - Lag time for the treatment effect
  - .....

**Identify factors you can control and control them**

**Design to deal with those you can't control**

# Choice of (1-β): Power of detecting a difference

$$1 - \beta = \mathbf{P}_{H_a}(\textbf{Reject the null})$$

- Power of a test is the probability of detecting a true, underlying difference
  - In practice, one that is both worth detecting, but biologically reasonable
- $\alpha$ is usually set to 0.05,
- Power (1-β) is often set to 0.8
- Sample size needs to be selected to ensure the desired power for the scientifically relevant difference of interest

# Influence of 1-β
# TOXO Sample Size

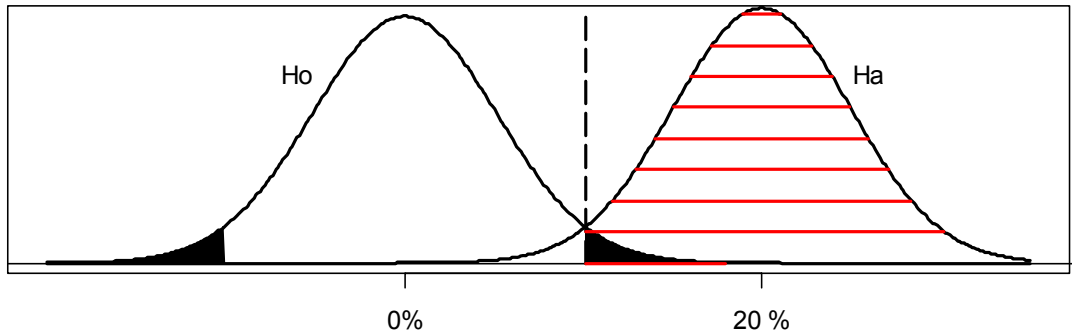
- With $\alpha$ =0.05 (2-sided)

| Event Rate (%) Placebo | Event Rate (%) Trt | Δ (%) | Percent Reduction | Power (1–$\beta$) | Sample Size per group |
|---|---|---|---|---|---|
| 30 | 15 | 15 | 50 | **0.90** | **161** |
| | | | | **0.80** | **120** |
| | | | | **0.70** | **95** |
| 30 | 20 | 10 | 33.3 | **0.90** | **392** |
| | | | | **0.80** | **293** |
| | | | | **0.70** | **231** |

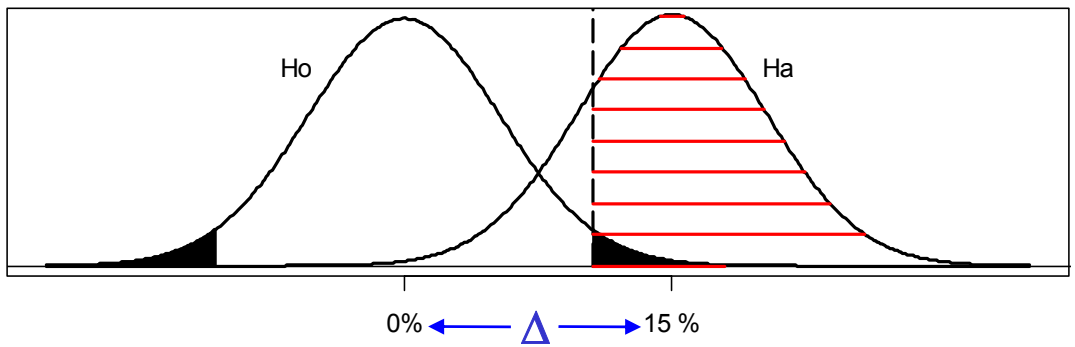
# Power and Effect size Trade-off

- Power depends on the alternative hypothesis defined by the effect size $\Delta$
- When sample size *n* is limited, for any fixed $\alpha$ we can trade-off Power and Effect Size
  - This trade-off is represented by the Operating Characteristic curve
- We can always increase the effect size to reach a certain power, but it is very important to check if the assumed effect size is biologically plausible

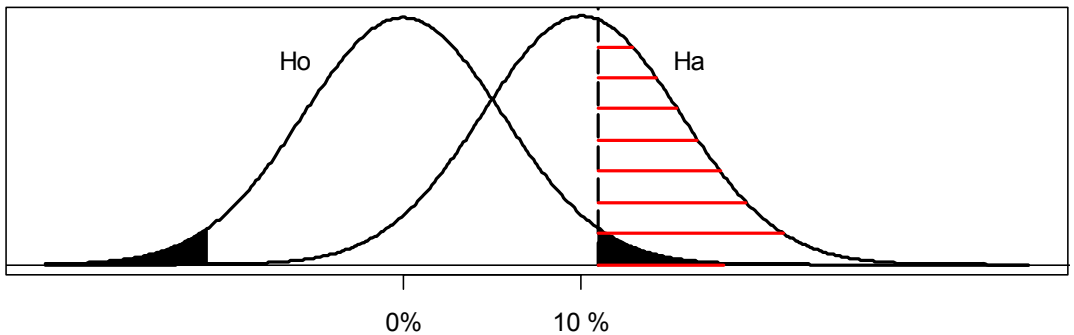
# Power and Effect size trade-off (cont'd)




**67% reduction**

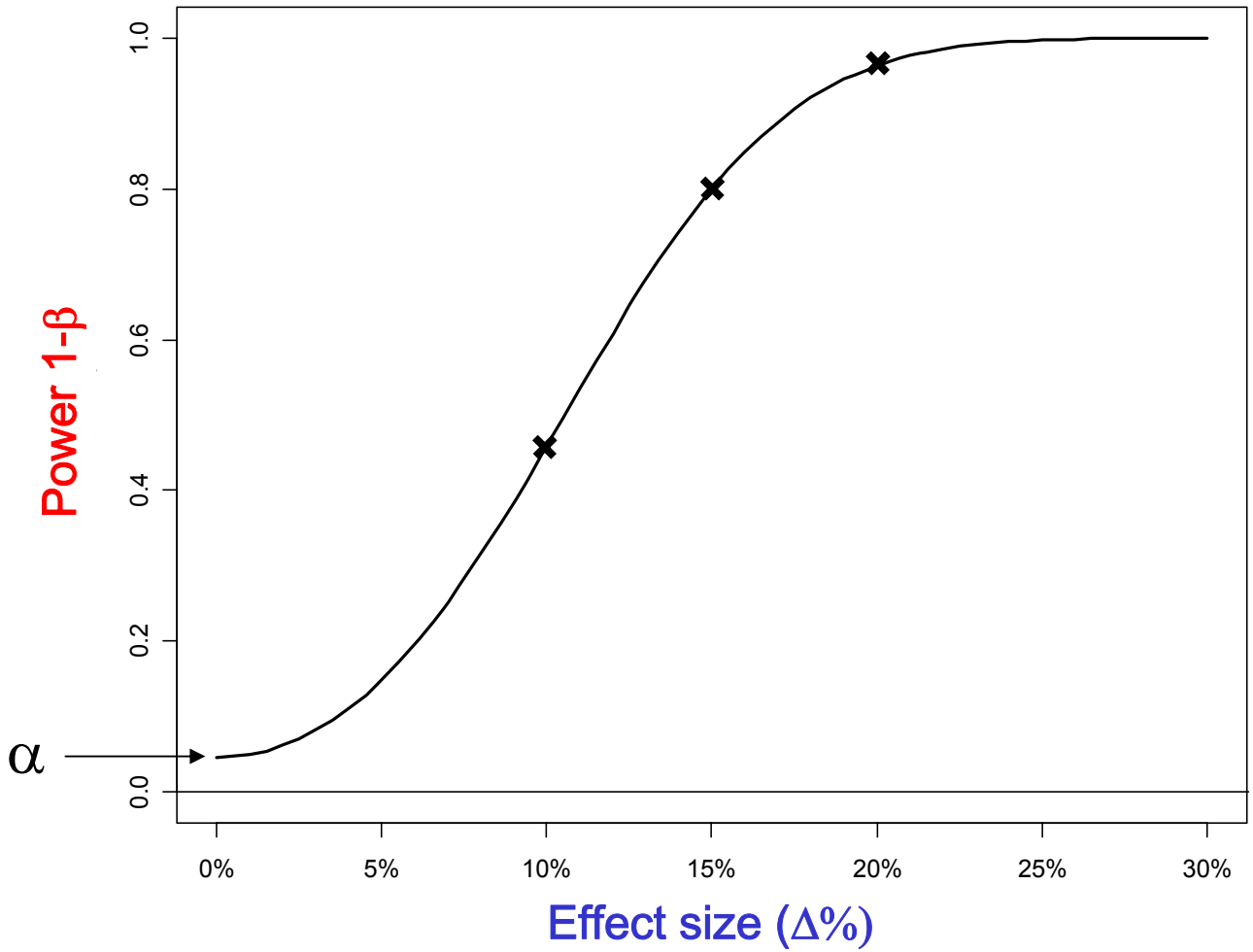



**50% reduction**




**33% reduction**

# Operating Characteristic curve



# Sometimes (frequently) we cannot get the "required" sample size

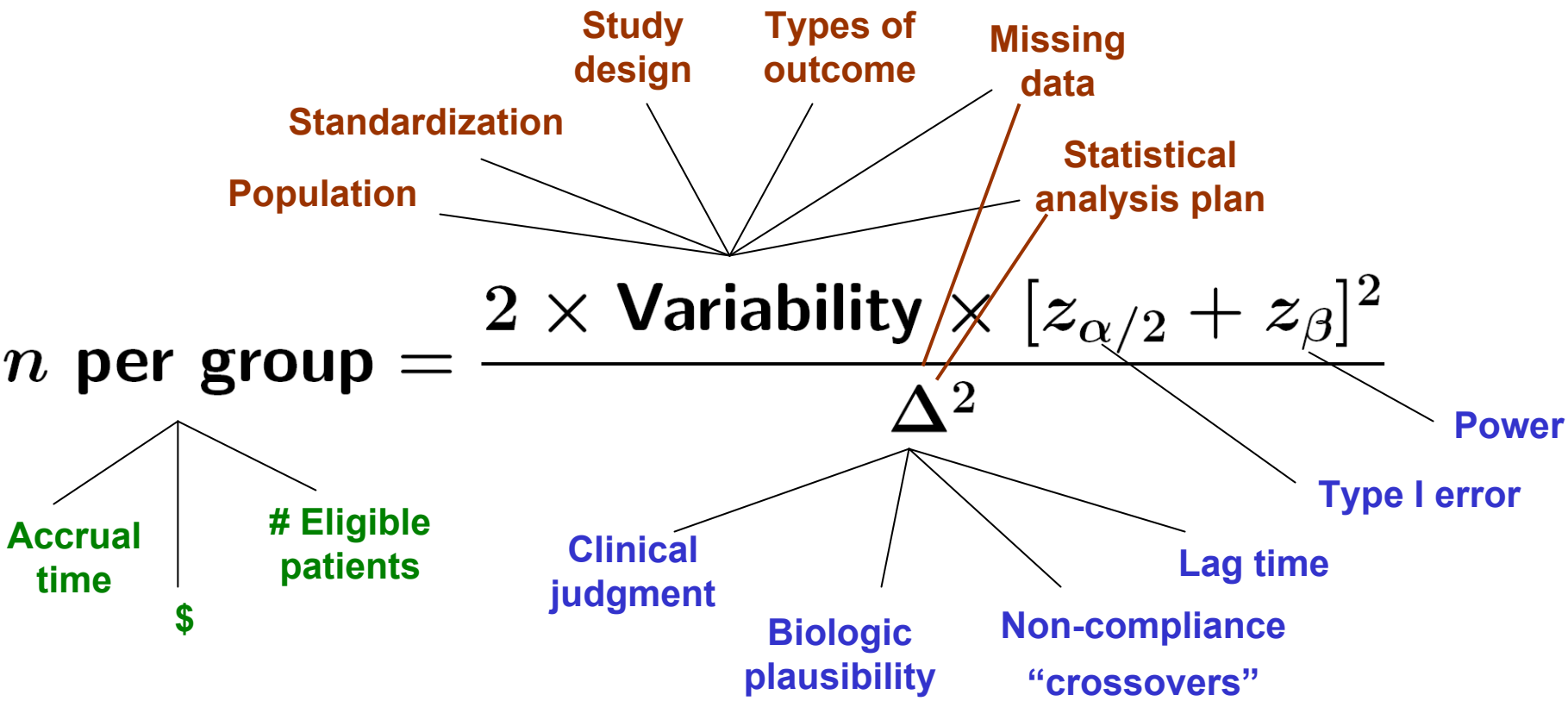
- The total number of available patients may be limited by factors such as money, size of target population and accrual time
- If the sample size is limited, we can try to reduce $\sigma$ (more precise measurements), change $\Delta$, adjust power
- But, need to be realistic!
- In situations where conducting a high power test is impossible, it may still be worthwhile to conduct a high-quality study as an input to a meta-analysis, **or as a pilot, phase II study**
- But, ....

# Sample Size and Statistical Significance

- A large sample size can produce statistical significance when $\Delta$ is very small (not of practical, clinical or public health importance)
- A small sample size may fail to detect a difference, even when $\Delta$ is of practical, clinical or public health importance
- The literature is populated by false positives and under represents the potentially true positives

Freiman et al., “The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials.” *NEJM* 1978;299:690-4

# Things to consider in determining sample size (testing template)



# Computer programs and approaches

- Sample size for a desired type I error, power and effect size
- Power for a given sample size, type I error and effect size
- Minimal detectable effect size for a given sample size, type I error and power
- Similar for CIs

# Summary

- Sample size (or the monitoring plan) should be specified before conducting the study
- Inputs for the sample size calculation should be based on results from other studies or reasonable judgment
- Interim analysis provides an opportunity to check these assumptions and adapt
- Study design, the statistical analysis plan, missing data, .... all play an important role in determining the sample size

**Do a comprehensive assessment and be realistic**

# Bayesian Experimental Design

- We are all Bayesians in the design phase
- We use previous information and our opinions to determine goals and inputs
- Generally, there is considerable uncertainty regarding the inputs
- So, use distributions and produce effective designs
- Bayesian design for either Bayesian or frequentist (traditional) goals

# Bringing in uncertainty in determining Sample Size
## (to control CI length)

1. When you know the formula and can do the math
2. When you know the formula, but can't do the math
3. When you don't know the formula or there isn't a formula

# When you know the formula and can do the math

$$n = 4Z^2_{\alpha/2}\left(\frac{\sigma}{L}\right)^2$$

You pick these

# What if you don't know $\sigma^2$

- Using some background information on $\sigma^2$ , do some "what ifs" or pick a conservative value
  - e.g., assume background data indicate that $\sigma^2$ has a log-normal distribution, with mean "avg" and variance $C^2 \times (avg)^2$
  - C is the coefficient of variation
- To control expected CI length, use

$$n = 4Z^2_{\alpha/2}\left(\frac{avg}{L^2}\right)$$

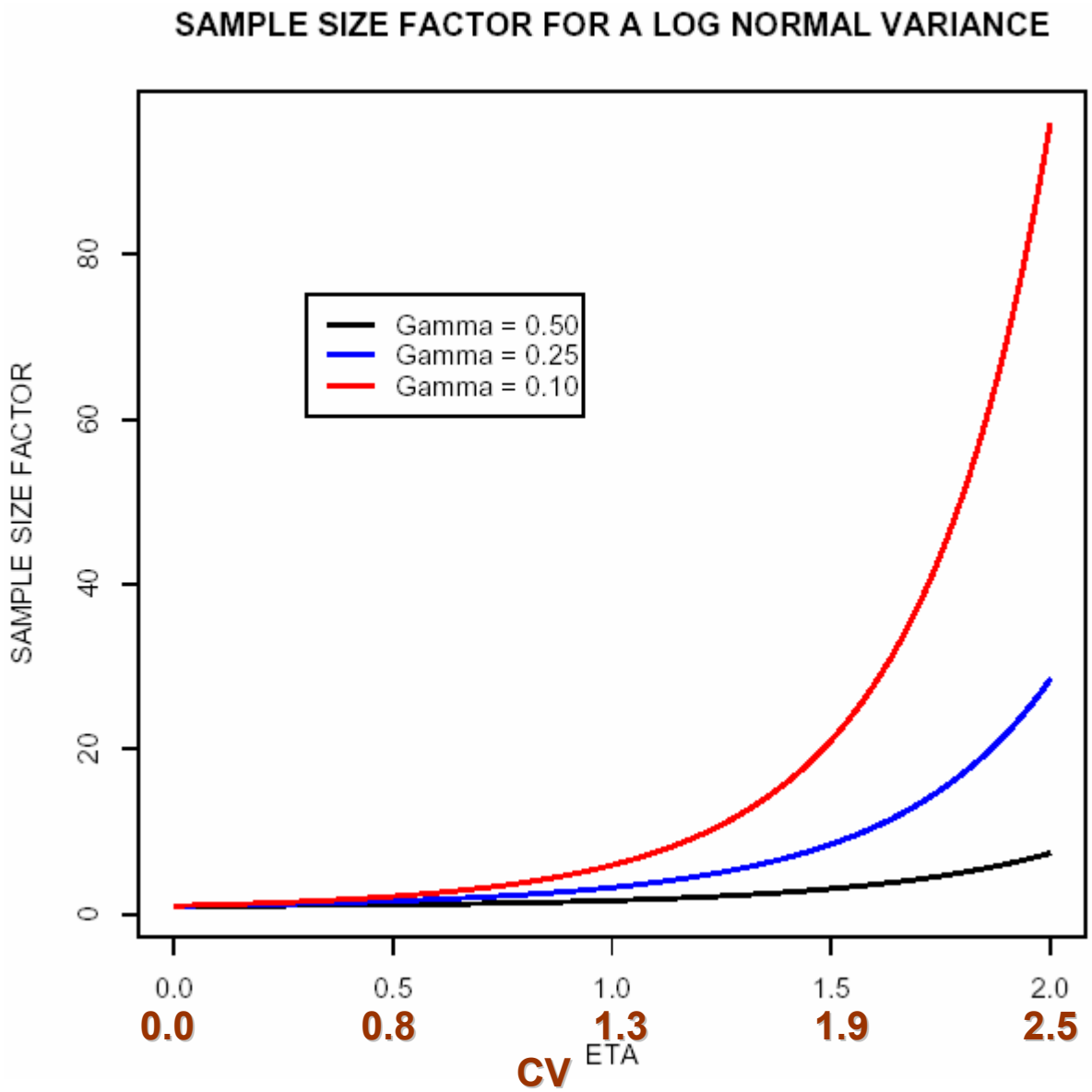
- This is just like using a "best guess" for $\sigma^2$

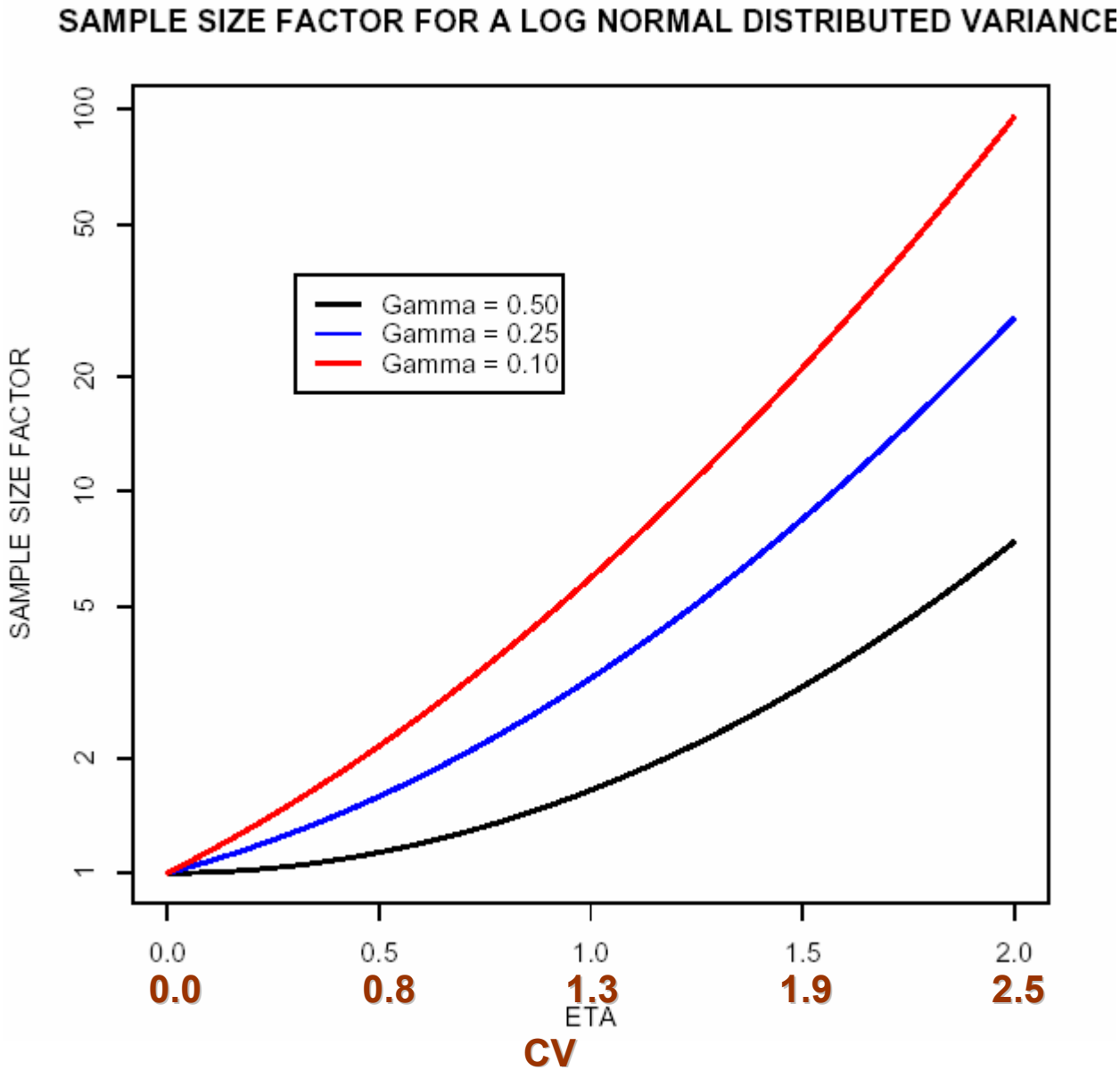
# "On average" length control can leave you with a too-wide CI

- So, consider controlling the probability that the CI is too wide
- That is, find an "n" so that using the log-normal distribution,

**pr(CI length > L | n) < $\gamma$ (= .10, for example)**

- With this sample size we have only a 10% chance of obtaining a too-wide interval
- Could do a sequential study, continuing until L falls below the maximal length
  - This guarantees control, but pays by not knowing the sample size up front


SAMPLE SIZE FACTOR FOR A LOG NORMAL VARIANCE
SAMPLE SIZE FACTOR
0
20
40
60
80
Gamma = 0.50
Gamma = 0.25
Gamma = 0.10
0.0
0.5
1.0
1.5
2.0
0.0
0.8
1.3
1.9
2.5
CV
ETA


SAMPLE SIZE FACTOR FOR A LOG NORMAL DISTRIBUTED VARIANCE
SAMPLE SIZE FACTOR
1
2
5
10
20
50
100
Gamma = 0.50
Gamma = 0.25
Gamma = 0.10
0.0
0.5
1.0
1.5
2.0
0.0
0.8
1.3
1.9
2.5
ETA
CV

## WHAT IF YOU KNOW THE FORMULA, BUT CAN'T DO THE MATH? SIMULATE

1. Generate "$NREPS$," $\sigma^2$ values from its distribution

**Number of simulation replications**

2. For each $\sigma^2$ value, compute CI length using the formula
   - In our basic case a multiple of $\frac{\sigma}{\sqrt{n}}$
3. From these $NREPS$ CI lengths, find the fraction that are $>$ L
   - If this fraction is $\geq \gamma$, increase $n$
   - If this fraction is $< \gamma$, decrease $n$

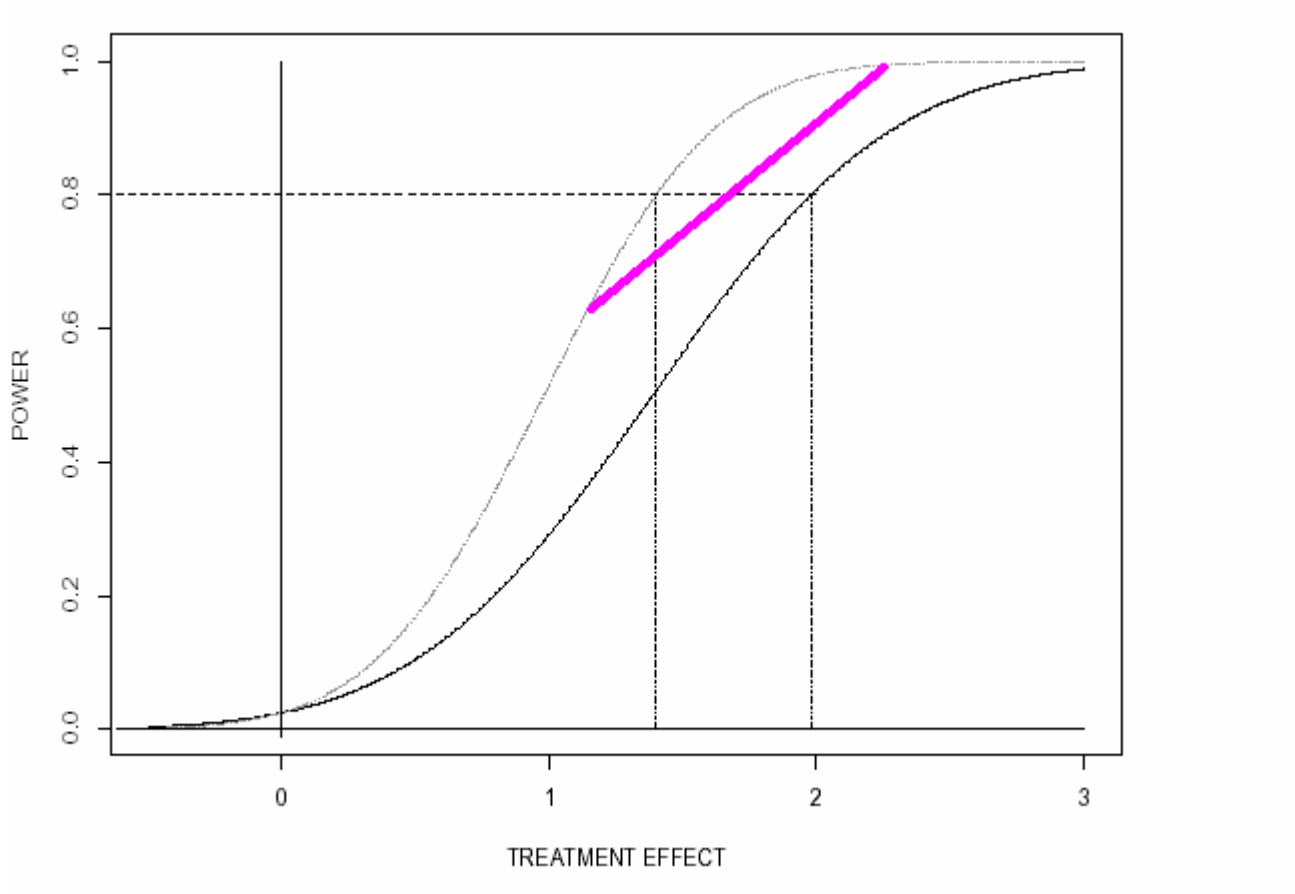
## WHAT IF YOU DON'T HAVE A FORMULA?

1. Generate "$NREPS$," $\sigma^2$ values from its distribution
2. For each $\sigma^2$ value:
   (a) Simulate data for a study of size $n$
   (b) Feed these data into an analysis package
   SAS, WinBUGS, STATA, MLWin, ...
   (c) Compute CI length using the computer output
3. From these $NREPS$ CI lengths, find the fraction that are $>$ L
   - If this fraction is $\geq \gamma$, increase $n$
   - If this fraction is $< \gamma$, decrease $n$

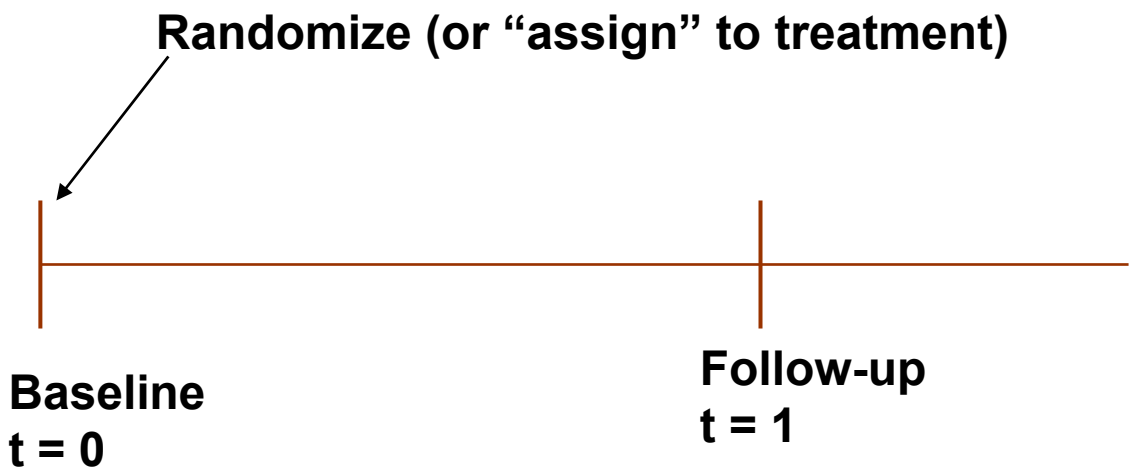
- Can require mucho CPU, so need to be efficient
  - Start with an initial set of "$n - values$"
  - Then home-in
  - Use theory to help get started and to home in

**Number of simulation replications**

# For hypothesis testing: In the region of "usual" powers, the average power is less than the power of the average



# A baseline/follow-up design



# Sample size for a baseline/follow-up design

$Y_{ijt}$ = measurement for person i in group j (j = 1 or 2)
at time t ( t = 0 or 1)

$G_j$ = 0, if group 1; $G_j$ = 1, if group 2

**Model**

$$Y_{ijt} = \mu + \alpha t + \gamma G_j + \beta t G_j + e_{ijt}$$

- $cor(e_{ij0}, e_{ij1}) = \rho$
- $\alpha$ is the time trend; $\gamma$ is the group main effect
- $\beta$ is the treatment effect
  - **The treatment by time interaction**

# The transparent analysis (a difference of differences)

$$(Y_{ij1} - Y_{ij0}) = \alpha + \beta G_j + (e_{ij1} - e_{ij0})$$

With $\bullet$ indicating averaging

$$\hat{\beta} = (Y_{\bullet 21} - Y_{\bullet 20}) - (Y_{\bullet 11} - Y_{\bullet 10}) = \beta + \text{residuals}$$

- The parameter $\gamma$ disappears
- With n per treatment group:

$$\text{Var}(\hat{\beta}) = \frac{4\sigma^2(1-\rho)}{n}$$

# The "optimal" analysis (a difference of adjusted differences)

Use $(Y_{ij1} - \rho Y_{ij0})$ as the building block

With $\bullet$ indicating averaging

$$\hat{\beta} = (Y_{\bullet 21} - \rho Y_{\bullet 20}) - (Y_{\bullet 11} - \rho Y_{\bullet 10}) = (1- \rho)\,\gamma + \beta + \text{residuals}$$

- The parameter $\gamma$ does not disappear
- So, this is a biased estimate unless $\gamma = 0$
  - i.e., need a randomized study so that groups are comparable at baseline
- With n per treatment group:

$$\text{Var}(\hat{\beta}) = \frac{2\sigma^2(1-\rho)^2}{n} < \frac{4\sigma^2(1-\rho)}{n}$$

# Applications & Comments

- The optimal analysis is dangerous unless the study is randomized so that the group effect is 0 at baseline
- So, need the “transparent” (FU – BL) analysis for a non-randomized study or a crossover study
  - The non-randomized is still dangerous due to regression to the mean
  - The crossover study must use the transparent analysis because, though randomized, it is to the sequences AB and BA, so not comparable at baseline
- If $\rho = 0$, the optimal analysis compares the follow-up measurements only and so is really a “two group” comparison

# Example for a weight loss study

- Goal is to reduce BMI among overweight/obese individuals
- Baseline BMI: mean = 33.1; $\sigma$ = 6.0
- Correlation between baseline and 6 month follow-up: $\rho$ = .90 or .95
- Effect size of interest: $\Delta$ = 1.0 is important and possible
- Let's try some sample sizes, $\sigma$ and $\rho$ values and see what $\Delta$s emerge as "detectable" with acceptable power

# Demonstration with the “R” program

# Bonus Slides

## Five steps to obtain the sample size formula

$\gamma$ is a generic quantity of interest

# 1. Create Two simple hypothesis

$H_o: \gamma = 0$

$H_a: \gamma = \Delta$

# 2. Find an estimator & calculate its properties

$H_o: \gamma = 0$ | $H_a: \gamma = \Delta$

Unbiased Estimator

$$\hat{\gamma}[data]$$

$$E(\hat{\gamma}[data]) = \gamma = 0 \qquad E(\hat{\gamma}[data]) = \gamma = \Delta$$

$$\mathbf{SE}(\hat{\gamma}[data]) = \mathbf{SE_o} \qquad \mathbf{SE}(\hat{\gamma}[data]) = \mathbf{SE_a}$$

# 3. Create a Test Statistic & find its distribution

$H_o: \gamma = 0$ | $H_a: \gamma = \Delta$

Test Statistic

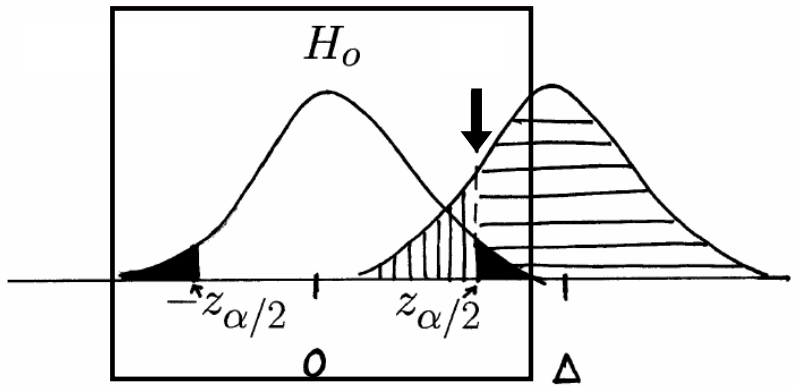
$$Z = \frac{\hat{\gamma}[data]}{SE_o}$$

$$Z \sim N(0,1) \qquad Z \sim N\left(\frac{\Delta}{SE_o}, \left\{\frac{SE_a}{SE_o}\right\}^2\right)$$

# 4. Setup the rejection rule

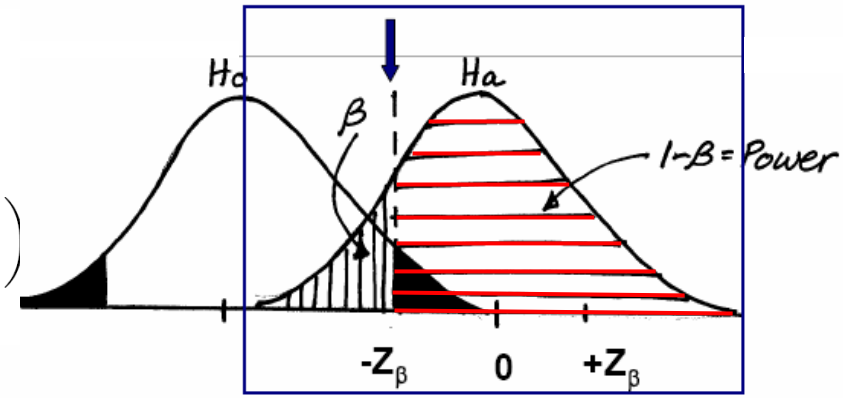
$$\alpha = \mathsf{P}_{H_o}(\text{Reject the null})$$

$$= \mathsf{P}_{H_o}(|Z| > z_{\alpha/2})$$



# 5. Calculate the power of the test

$$
\begin{aligned}
1-\beta &= P_{H_a}(\text{Reject the null}) \\
&= P_{H_a}(|Z| > z_{\alpha/2}) \approx P_{H_a}\left(Z > z_{\alpha/2}\right) \\
&= P_{H_a}\left(\left(Z - \frac{\Delta}{SE_o}\right)\left\{\frac{SE_o}{SE_a}\right\} > \left(z_{\alpha/2} - \frac{\Delta}{SE_o}\right)\left\{\frac{SE_o}{SE_a}\right\}\right)
\end{aligned}
$$

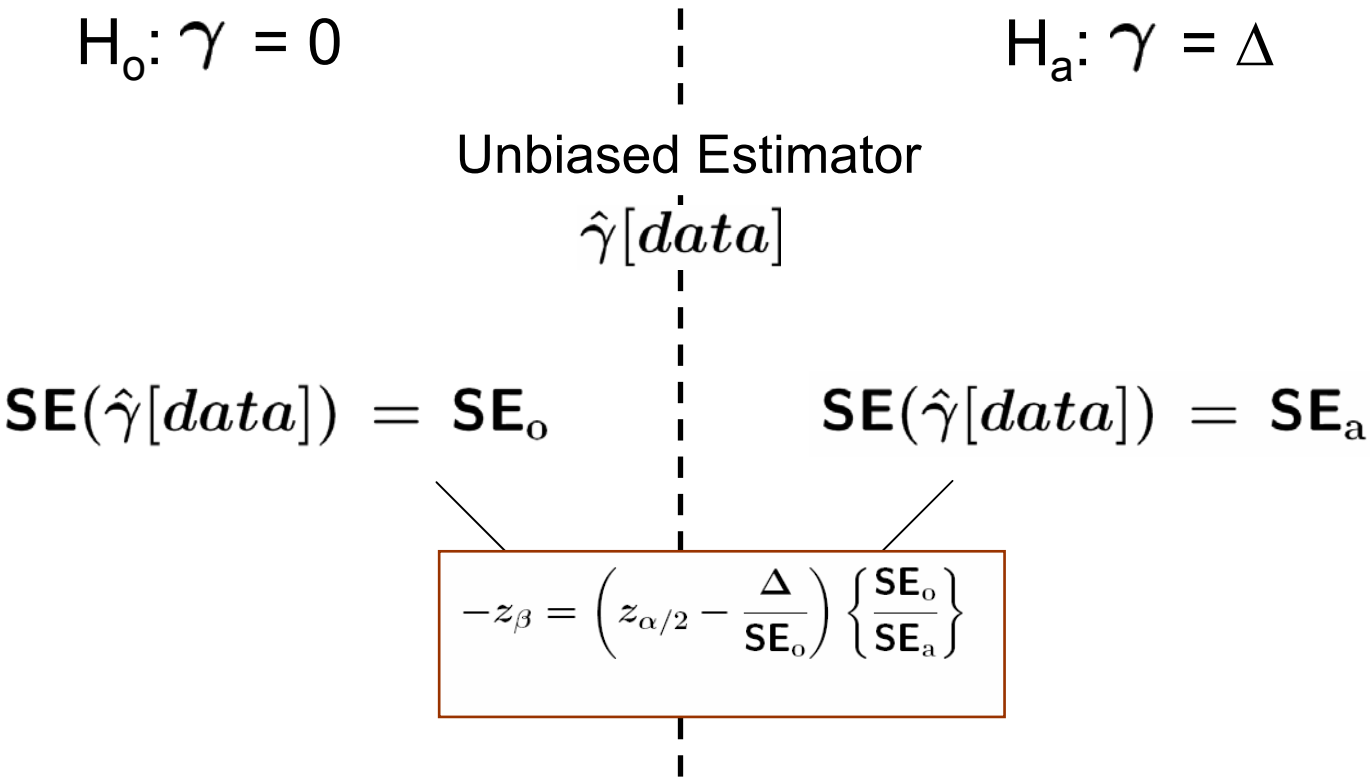



$$N(0,1)$$

$$-z_\beta = \left(z_{\alpha/2} - \frac{\Delta}{SE_o}\right)\left\{\frac{SE_o}{SE_a}\right\}$$

Standardization

# Use the standard error formula from step 2 to calculate the required sample size *n*



# Sample size for estimating $\Delta$

- Goal: Estimate $\Delta$ with a certain precision ("narrowness of CI")
- Let $\hat{\gamma}[data]$ be an unbiased estimator: i.e.

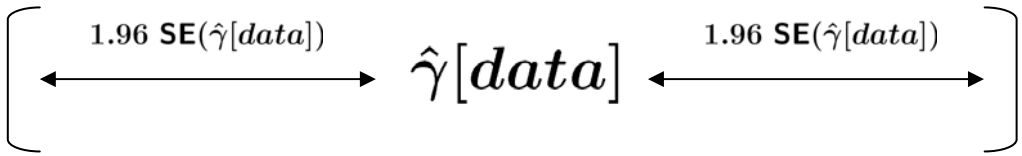
$$E(\hat{\gamma}[data]) = \gamma$$

- Standardization (assuming Normality – usually by CLT):

$$Z = \frac{\hat{\gamma}[data] - \gamma}{\mathbf{SE}(\hat{\gamma}[data])} \sim N(0,1)$$

- 95% C.I. for $\gamma$ is: $\hat{\gamma}[data] \pm 1.96\ \mathbf{SE}(\hat{\gamma}[data])$

# Sample size for estimating $\gamma$ (cont'd)

- The 95% C.I. is: 


- To control the half width (**d**) of the CI, it requires a sample size $n$ such that

$$1.96\ \mathbf{SE}(\hat{\gamma}[data]) = \mathbf{d}$$

- Where is the sample size in the formula?
  - Inside $\mathbf{SE}(\hat{\gamma}[data])$ !

- E.g. $\hat{\gamma}[data] = \overline{X}$ and $\mathbf{SE}(\hat{\gamma}[data]) = \sigma/\sqrt{n}$

# Sample Size required to detect an effect size $\Delta$ between two proportions

$H_o: p_2 - p_1 = 0$ | $H_a: p_2 - p_1 = \Delta$

Unbiased Estimator

$$\hat{\gamma}[data] = \hat{p}_2 - \hat{p}_1$$

$$SE_o = \sqrt{\frac{2\bar{p}\bar{q}}{n}} \qquad SE_a = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n} + \frac{\hat{p}_2\hat{q}_2}{n}}$$

$$n = \frac{(z_{\alpha/2}\sqrt{2\bar{p}\bar{q}} + z_{\beta}\sqrt{\hat{p}_1\hat{q}_1 + \hat{p}_2\hat{q}_2})^2}{\Delta^2}$$