Practical Considerations for Sample Size Calculations in Clinical Research

Daniel J. Tancredi, Ph.D.
UC Davis School of Medicine

Graduate Group in Biostatistics Seminar
October 6, 2015

Case Study

Colleague wants to know: Does teaching HIV-infected moms in sub-Saharan Africa to flash heat their breast milk lead to better outcomes for their infants?

Asks: “How many subjects do I need?”

Answer Question with Questions

A good answer to the “How many subjects do I need?” requires knowing:

- Study objective(s)
- Response variable(s), plan for measuring, alternative instruments
- *Infinite data case:* “If we had a very, very large quantity of data of the kind under consideration, would it answer our research questions?”
- What can go wrong? Non-response rate?
- Key sources of variation?
- Time frame? Other practical constraints?

Back to Case Study: Objective

Determine appropriate sample size for a prospective evaluation of a

- Cluster-randomized intervention in a
- Two-arm, parallel group, usual-care controlled study for **superiority** where
- intervention uptake could be low and
- losses during follow-up substantial

A practical sample size strategy

- Determine the *effective sample size* needed for a prospective evaluation of
  - *individually*-randomized interventions in a trial
  - to detect effect sizes that would be meaningful (and attainable) from an Intent-to-Treat perspective
  - losses during follow-up are *negligible*.
- Apply *variance inflation factors** to the effective sample size to determine a target sample size for actual enrollment

* Hsieh et al. 2003. An overview of variance inflation factors for sample size calculation. *Eval Health Prof* 26:239-57

Variance Inflation Factor (VIF), Design Effects and Effective Sample Sizes

Given two design and analysis strategies S_0 and S_1 for estimating the parameter θ,

$$VIF(S_1\ vs.S_0) = \frac{V_1(\hat{\theta})}{V_0(\hat{\theta})}$$

Usually, S_0 is simple (i.e. easy!).

VIF generalizes *design effects** to also account for analysis features

* Kish, Leslie. 1965. *Survey Sampling*. Wiley

Variance Inflation Factor (VIF), Design Effects and Effective Sample Sizes

When variance varies inversely with sample size (as usual), translate the required **effective sample size N_0** for S_0 into the target **actual sample size** N_1 for S_1:

"Actual" N_1 = "Effective" N_0 × VIF(S_1 vs. S_0)

To power S_1, take advantage of readily available power calculations for S_0, provided VIFs are known!

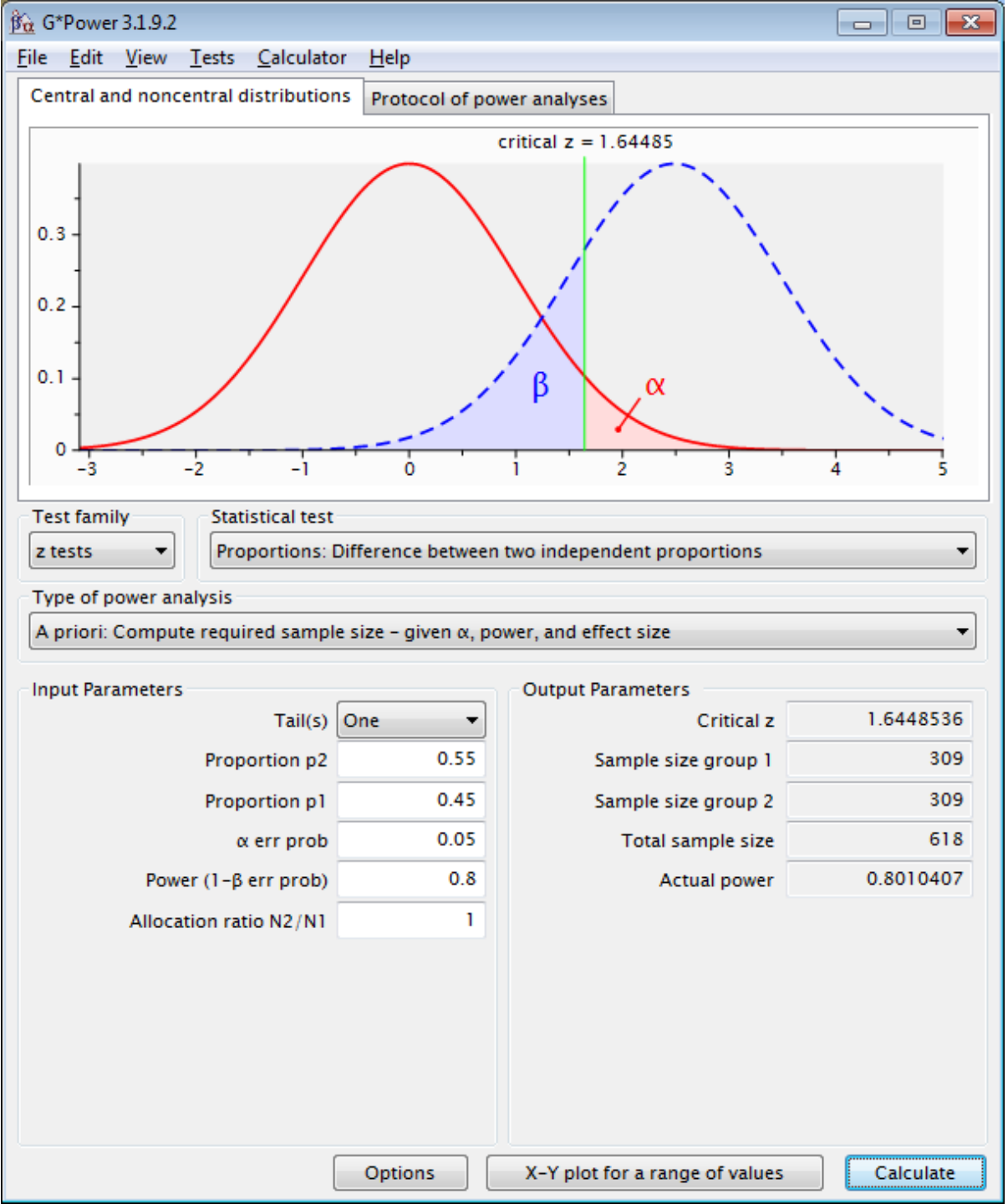
Example:

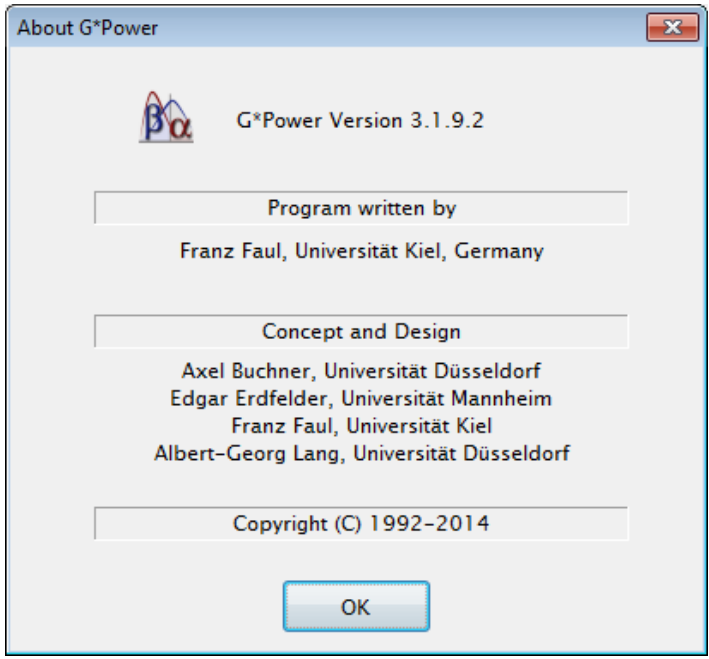
- N_0 of 100 is needed for SRS
- Clustered survey VIF for a given outcome is 1.2.
- Hence, use N_1=120 for the Clustered survey.

- VIFs are available for many features, tend to be (asymptotic) approximations, and can be multiplied together

Basic inputs

- Objective and Comparison: *Flash heat arm vs. Usual Care arm, for superiority*
- Binary outcome: Infant is *alive, HIV - and normal sized at 18 months of age.*
- One-sided testing, Prob{ Type-1 error } = 5%
- 80% power
- A tough-to-spot 10 percentage point difference (e.g. 45% vs. 55%). *Assumes 50% uptake of an intervention that truly produces a 20 percentage point difference.*





http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/

Effective Sample Size

- Objective and Comparison: *Flash heat arm vs. Usual Care arm, for superiority*
- Binary outcome: Infant is *alive, HIV - and normal sized at 18 months of age.*
- One-sided testing, Prob{ Type-1 error } = 5%
- 80% power
- A tough-to-spot 10 percentage point difference (e.g. 45% vs. 55%). *Assumes 50% uptake of an intervention that truly produces a 20 percentage point difference.*

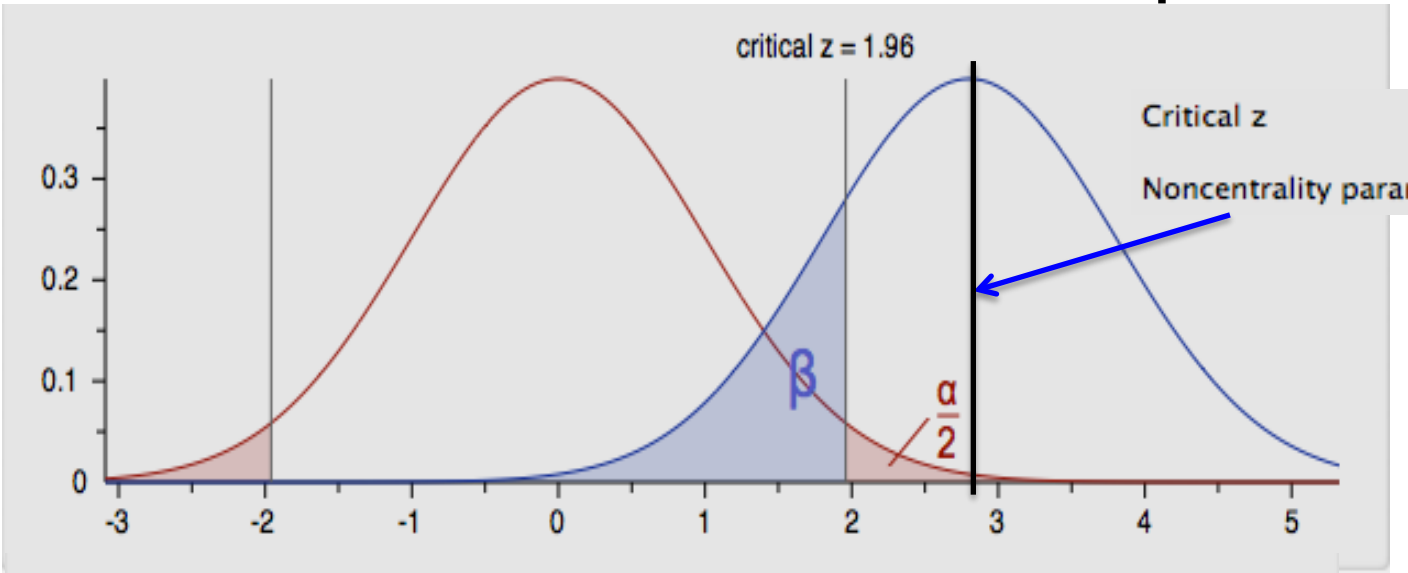
- **Effective sample size of 309 mother/child dyads per arm, 618 dyads total**

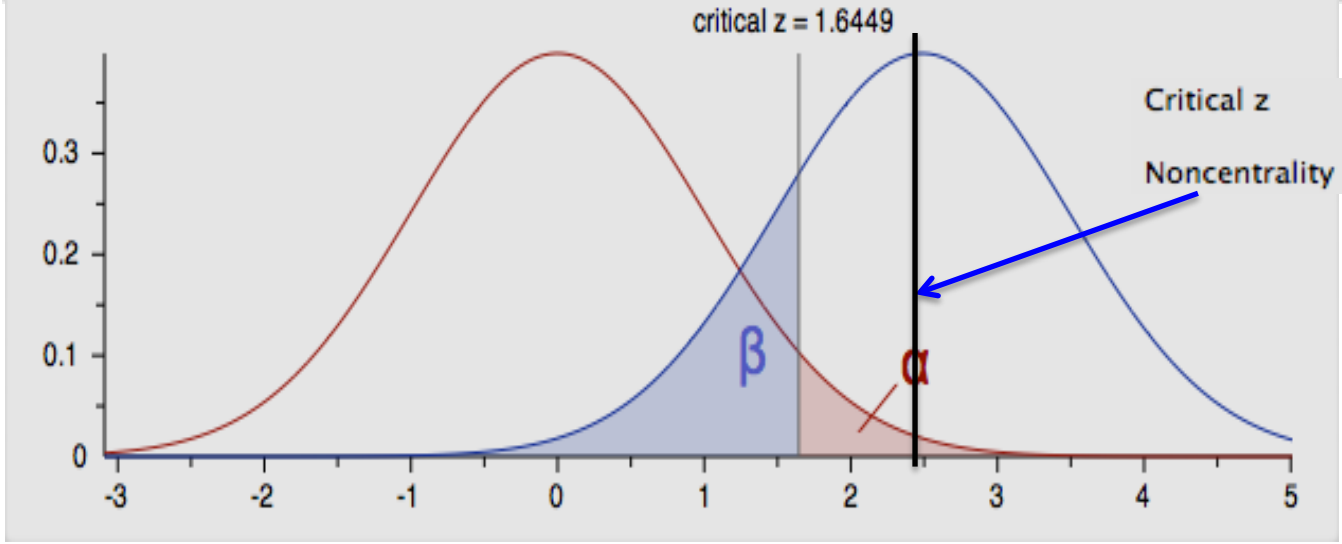
Question for Students: Sample size impacts of basic inputs

How much does a one-sided versus a two-sided test impact the required effective sample size?

- (a) About a 5% reduction?
- (b) About a 20% reduction?
- (c) About a 40% reduction?

Noncentrality parameters for Z-test 2-tailed vs. 1-tailed α=5%, β=20%





Noncentrality parameter for Z-test = Critical Z + $Z_{1-\beta}$. Recall that $Z_{0.80} \approx 0.84$.

Noncentrality parameter (NCP) for Z-test

To test null hypothesis of equivalence:

- One-sided α =5%, β = 20%: **2.5**
- Two-sided α=5%, β = 20%: **2.8**

➔A one-sided test requires only

(2.5 / 2.8) ** 2 ≈ 80%

of the effective sample size as a two-sided test, under standard testing conditions (α=5%, β=20%).

Question for Students: Sample size impacts of basic inputs

To cut the minimum detectable effect size in half, how much do you have to increase the required effective sample size?

- (a) Make it 1.5 times as big?
- (b) Make it 2 times as big?
- (c) Make it 4 times as big?

Minimum Detectable Effect and Noncentrality Parameter for Z

MDE = NCP × SE(of Effect Estimator)

(e.g. MDE = 2.8 SE)

Asymptotic Standard Error of θ-hat

= 1 / √(Fisher Info for θ) = Φ / √N

➔ Need to quadruple N to cut the SE (and hence the MDE) in half!

Sample size impacts of cluster randomization

- Should we make any adjustments to the sample size calculation to account for cluster randomization?
  - (a) No
  - (b) Maybe/Yes

Cluster randomized trials have pitfalls, at both study design and analysis

Lit Review	Types of studies reviewed	OK Power Calcs (%)	OK Statistical Methods (%)
Donner et al. (1990) [*PMID: 2084005*]	16 CRTs published 1979-89	19	50
Rooney & Murray (1996) [*882241*]	131 School-based smoking prevention studies publ. 1975-1991	<17 were OK on both	
Simpson et al (1995) [*7573621*]	21 Primary prevention CRTs in AJPH & PM 1990-93	19	57
Varnell et a (2004) [*14998802*]	60 Primary prevention CRTs in AJPH & PM 1998-2002	16	55
Murray et al (2008) [*18364501*]	75 Cancer prevention and control. CRTs 2002-06	24	45

Variance Inflation Factors for Cluster Randomized Trial

Cornfield penalties: Two important and underappreciated adjustments required for a proper (classical) analysis of cluster-randomized trials

1. Mean square error for clusters, not individuals
2. Degrees of freedom for T-test reference distribution based on number of clusters, not individuals

Further reading:

Cornfield. AJE 1978;108:100-102;
Murray DM. Eval Rev. 1996 Jun;20(3):313-37 ;
Eccles et al (2008) [PMID: 12571345]

Variance Inflation Factors for CRT

For cluster randomization with equal cluster sizes **m**,

VIF_C1 = { 1 + (**m** – 1)**ρ** }, where **ρ** is the intracluster correlation (ICC).

ICC is outcome dependent. When **ρ** = 0, there is no penalty for cluster randomization, when **ρ** > 0, the effective sample size is less than the actual sample size because less and less information comes from each additional observation from within the same cluster.[1,2]

For now, let's assume that when **m ≈ 30,** VIF_C1 **≤ 1.80**, for our outcomes**.**[3]

Notes:

1. Rule of thumb: keep the cluster size **m ≤** 1 / **ρ,** the information bound.
2. Killip et al, 2004. What Is an Intracluster Correlation Coefficient? Crucial Concepts for Primary Care Researchers. *Ann Fam Med* 2:204-208.
3. Taljaard et al. 2008. Intracluster correlation coefficients from the 2005 WHO Global Survey on Maternal and Perinatal Health: implications for implementation research. *Paediatric and Perinatal Epidemiology,* **22**, 117–125

What about VIF for degrees of freedom at randomization level?

- $\sqrt{VIF\_C2} = (t_{1-\alpha,df} + t_{1-\beta,df}) \div (z_{1-\alpha} + z_{1-\beta})$
- Big concern when df < 10, negligible concern when df > 40
- For df = 20:
  - $t_{0.975,\ 20} = 2.09$
  - $t_{0.80,\ 20} = 0.86$
  - VIF_C2 ≈ 1.10

Adjustments for CRT

Combining VIFs for cluster randomization approximately doubles target enrollment:

Target Enrollment of 618 x 1.8 x 1.1, rounds up to 1,224. With about 30 mother/infant dyads per clinic, this rounds up to 42 clinics.

Variance Inflation Factors For Within-cluster losses

- For inclusion of HIV-neg. moms:

$$5/4 = 1.25$$

- For children becoming HIV+ before 6 m:

$$1 / 85\% \approx 1.18$$

- For loss to follow-up:

$$1 / 80\% = 1.25$$

$$\rightarrow VIF\_W = 1.25 * 1.18 * 1.25 \approx 1.85$$

So, to get 30 “useful” dyads from a clinic we need to enroll about 56 dyads.

Target Enrollment Calculation

Combining these variance inflation factors, we find that the total effective sample size of 618 would be achieved by an actual enrollment sample size of 56 dyads in each of 42 clinics, a target enrollment of 2,352.

Question for Students: Sample size impact of unbalanced allocation

How much would a 2:1 allocation of clinics to treatment condition impact the required effective sample size?

- (a) Increase it about 10 to 15%
- (b) Increase it about 25 to 35%

Unbalanced allocation is less of a problem than “everyone” thinks

It can easily be shown that the VIF for an unbalanced k:1 allocation in a two-group parallel study is

$$(k + 1)^2 / 4k$$

For k=2, VIF=9/8

For k=3, VIF=4/3

Case-Study Conclusions

- Our target sample size is about 4 times larger than a naïve “basic” sample size calculation would recommend!
- But we account for key features of the study design, including
  - cluster randomization,
  - moderate expected losses during follow-up, and
  - moderate intervention uptake.
- Because sample sizes can be so amplified, the following refinements can have a big payoff:
  - Use of covariates to reduce residual between-cluster and within-cluster variance components
  - Improvements in intervention uptake.

The “Only Formula” Your Consultees Need To Understand

$$\text{Confidence} = \frac{\text{Signal}}{\text{Noise}} \times \sqrt{\text{Sample size}}$$

Source: CMAJ. 2001 Oct 30;165(9):1226-37.
“Why randomized controlled trials fail but needn't:
2. Failure to employ physiological statistics, or the only formula a clinician-trialist is ever likely to need (or understand!)” by DL Sackett
PMID: 11706914

Get louder signals

1. Selectively enroll high-risk patients.

2. Selectively enroll highly responsive patients.

3. Use potent treatments and give them a chance to exert their effect.

4. Avoid sloppy ascertainment.

$$\text{Confidence} = \frac{\text{Signal}}{\text{Noise}} \times \sqrt{\text{Sample size}}$$

Source: CMAJ. 2001 Oct 30;165(9):1226-37.

Minimize noise

1. Apply multiple treatments to each patient (e.g. cross-over designs).

2. Reduce uncontrolled patient heterogeneity in risks & responsiveness.

3. Achieve high adherence with treatments.

4. Objective and repeated measurements

$$\text{Confidence} = \frac{\text{Signal}}{\text{Noise}} \times \sqrt{\text{Sample size}}$$

Source: CMAJ. 2001 Oct 30;165(9):1226-37.

Question for Students: Sample size impact of unbalanced allocation

For a RCT for which a pre-test measure with 50% correlation with the post-test measure is available, what would be the difference in required effective sample sizes for an ANCOVA vs. a Change Score Analysis

- (a) ANCOVA needs ~25% less than CS
- (b) ANCOVA needs the same
- (c) ANCOVA needs ~25% more than CS

VIF for two pretest strategies

ANCOVA: $E(Y_{post} | Y_{pre}, T) = \beta_0 + \beta_1 Y_{pre} + \beta_T T$

Change Score: $E(Y_{post} - Y_{pre} | T) = \beta'_0 + \beta'_T T$

Let $R := \text{Correlation}(Y_{post}, Y_{pre})$

$$VIF_{ancova} \approx 1 - R^2 = (1 + R) * (1 - R)$$

$$VIF_{change} \approx 2 * (1 - R)$$

VIF for Covariates

Covariates (W) have both **good effects** (reduced noise in outcome) and **bad effects** (from collinearity) on precision for regression coefficient for key exposure variable X:

$$\frac{VIF(Y|X,W)}{VIF(Y|X)} = \frac{1 - R^2_{Y|X,W}}{1 - R^2_{Y|X}}$$

$$VIF(X|W) = \frac{1}{1 - R^2_{X|W}}$$

A BRIEF ASIDE: USING FISHER SCORES TO DEVELOP INTUITION

Practical considerations for sample size calculations in clinical research

Fisher Information

Information is the (co-)variance of scores

$$I(\theta) = E\left(\frac{\partial \log f}{\partial \theta_i}\,\frac{\partial \log f}{\partial \theta_j}\right)$$

The inverse of the Fisher Information is the asymptotic variance of MLE estimators (in regular models). So the more variance in the scores, the more information, the smaller the variance.

Inverse of partitioned matrix

(Aitken-) Block diagonalization of a non-negative definite partitioned matrix

$$A = \begin{pmatrix} \boldsymbol{A_{11}} & \boldsymbol{A_{12}} \\ \boldsymbol{A_{21}} & \boldsymbol{A_{22}} \end{pmatrix}$$

$$A = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{A_{21}A_{11}^{=}} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{A_{11}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A_{22} - A_{21}A_{11}^{-}A_{12}} \end{pmatrix} \begin{pmatrix} \boldsymbol{I} & \boldsymbol{A_{11}^{\sim}A_{12}} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix}$$, where

$\boldsymbol{A_{11}^{=}, A_{11}^{-}}$ & $\boldsymbol{A_{11}^{\sim}}$ *are arbitrary generalized inverses of* $\boldsymbol{A_{11}}$.

Source: Chapter 13 “Block-diagonalization and the Schur Complement” of S. Puntanen et al. *Matrix Tricks for Linear Statistical Models: Our personal top twenty*. Springer (2011).

Fisher Information for OLS Regression Coefficients

Example: $f(Y; x, w, \mu) = \frac{1}{\sqrt{2\pi}\sigma} exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$,

$$\mu = \boldsymbol{w}'\boldsymbol{\beta_w} + \beta_x x$$

$$I_n(\boldsymbol{\beta_w}, \beta_x) = \frac{1}{\sigma^2}\begin{pmatrix} \boldsymbol{W}'\boldsymbol{W} & \boldsymbol{W}'\boldsymbol{x} \\ \boldsymbol{x}'\boldsymbol{W} & \boldsymbol{x}'\boldsymbol{x} \end{pmatrix}$$

$$AVar(\hat{\beta}_x) = \sigma^2 \Big/ \{\boldsymbol{x}'\boldsymbol{x} - \boldsymbol{x}'\boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{x}\}$$

Hence, here VIF depends on how much of X is explained by W.

- In other settings (e.g. incomplete data) considering variances of “effective scores” gives good insight.

END OF BRIEF ASIDE

Increase sample size

1. Make it easy for collaborators to approach & enroll potential recruits. Use simple forms.

2. Minimize eligibility criteria, relying mainly on the "uncertainty principle", which prioritizes participants that matter to investigators most directly engaged.

3. Provide research assistants.

4-11. [See Source Reference for details]

$$\text{Confidence} = \frac{\text{Signal}}{\text{Noise}} \times \sqrt{\text{Sample size}}$$

Source: CMAJ. 2001 Oct 30;165(9):1226-37.

Noncentrality parameters for Z-test for other study objectives

Often, especially in survey research, **precision-of-estimation** is the objective:

Desire a suitably narrow “error margin”

a.k.a. (1-α)*100% CI

In this case, only α matters. The noncentrality parameter is just the critical value corresponding to α (e.g. 1.96 for a 95% CI).

You can use a power calculator to get a recommended sample size: just set β to 50% and α as above.

Noncentrality parameters for Z-test for other study objectives

For a **non-inferiority study**, the null hypothesis of inferiority is expressed with a tolerance margin:

$$H0: \mu_A - \mu_B \leq -d$$

$$vs. \quad H1: \mu_A - \mu_B > -d$$

One-sided α=5%, β=20%: NCP = **2.5**

But, the tolerance margin d would typically be a fraction of a minimum clinically significant difference, so sample size requirements could be large!

Noncentrality parameters for Z-test for other study objectives

For a **equivalence trial**:

$$H0: \mu_A - \mu_B \leq -d \textbf{ OR } \mu_A - \mu_B \geq d$$

$$vs. \quad H1: -d < \mu_A - \mu_B < d$$

- Null hypothesis test:
  - intersection-union-test (Berger & Hsu, 1996. Stat Sci.),
  - Two One-sided Test (TOST) (Jones et al , 1996. BMJ (313): 36-39
- Testing symmetric tolerance margins with two-sided α=5% and β=20% similar to using one-sided α=5% and β=10% (Senn, 2001. Stat Med (20): 2787-99): NCP=**2.93**

Another practical strategy: Exemplary Dataset Approach

- Using example datasets (pseudo-data) to estimate noncentrality parameters, which can then be used to determine power and sample size requirements
- Especially useful for GLM models, for longitudinal data or for tricky study designs
- Technique is incorporated into SAS and G*Power software and recommended by authoritative texts on longitudinal data analysis

Sample Code

Handouts demonstrate

- R code for logistic regression models
- SAS code for
  - case-control studies of Gene * Environment interaction
  - Comparison of slopes in longitudinal data

Remember

- As biostatisticians, we get asked a lot for sample size recommendations
- Have simple, practical methods available
- Build on these to account properly for complexity
- Be ready to help consultees improve the efficiency of their proposed strategies!