

# Sample size calculation in the planning of a cluster randomized trial: analytical and simulation based approaches

Henrik Støvring

hstovring@health.sdu.dk  
Research Unit for General Practice  
University of Southern Denmark

PAFP: National dissemination meeting  
Beijing, December 2007

# Outline

Ordinary power calculations

Statistical effects of clustering

Power in cluster randomized studies

Summary and perspectives

# Setting

- ▶ Compare two packages intended to improve family planning (avoid unwanted pregnancies)
- ▶ Packages are to be implemented at abortion clinics
- ▶ Only one package at each center
- ▶ **How large should the study be?**

## Naive strategies

1. Deliberately choose a number
2. Take as many as time and money allows
3. Include patients until difference is statistically significant
4. ...
  - ▶ Problems with above approaches
    - ▶ Likely too costly or too small
    - ▶ May introduce bias in estimates
    - ▶ May underestimate uncertainty

## Objective of power calculations

- ▶ Minimize burden on patients
- ▶ Minimize administrative burden
- ▶ Minimize cost
- ▶ Maximize scientific value, i.e. maximize precision

## Key concepts

- ▶ Based on
  - Significance level: Probability of declaring a difference **when there is none** ( $\alpha$ )
  - Power level: Probability of declaring a difference **when there is one** ( $1 - \beta$ )
- ▶ Related concepts
  - Type I error: declaring difference when there is none
  - Type II error: declaring no difference when there truly is one
- ▶ Typical choices:  $\alpha = 5\%$  and  $1 - \beta = 80\%$

## Input needed for power calculations

1. Outcome measure (Example: blood pressure, pregnancy status)
2. Statistical model (Examples: Normal distribution, Binomial distribution)
3. Minimal, clinically relevant difference in outcome (Example: 5 mmHg difference in blood pressure)
4. Measure of variability if data are continuous
5. Levels of significance and power
6. Assume all observations to be **independent**

## Example

- ▶ Test new treatment for hypertension
- ▶ Average, expected decrease in control group: 8 mmHg
- ▶ Average, expected decrease in intervention group: 10 mmHg
- ▶ Standard deviation of individual change in both groups: 7.5 mmHg
- ▶ Decreases are normally distributed and independent
- ▶ Significance level = 5%, power = 80%
- ▶ Equal sized groups

## Example (continued)

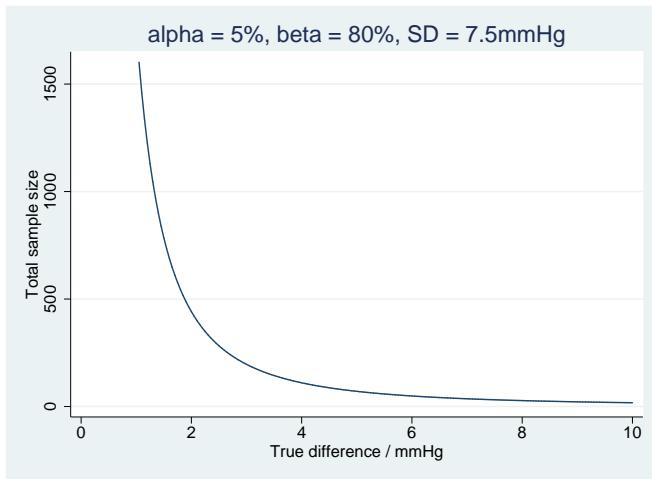
- ▶ Test statistic

$$z = \frac{\text{Av. Diff}}{\sqrt{\frac{2}{n}}SD}$$

- ▶ Total sample size is given by

$$\begin{aligned}n &= \frac{(\Phi^{-1}(0.975) + \Phi^{-1}(0.8))^2 \times SD^2}{(\text{True diff})^2} \\ &= \frac{(1.96 + 0.84)^2 (7.5\text{mmHg})^2}{(10\text{mmHg} - 5\text{mmHg})^2} \\ &\approx 441\end{aligned}$$

## Example (continued)



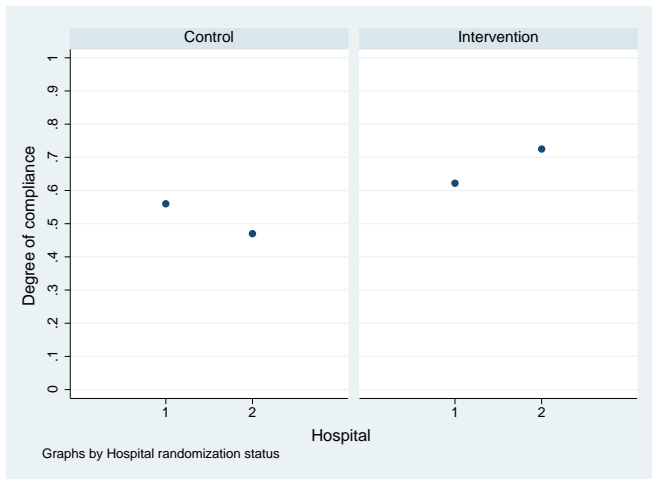
# Summary

- ▶ Power calculations are useful
- ▶ Involve complicated formulas. . .
- ▶ ... but can be by-passed in modern statistical computer packages
- ▶ Standard formulas only work with **independent** data

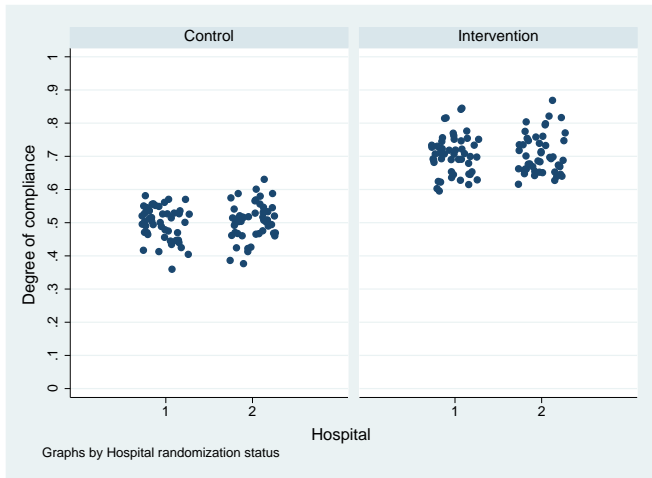
## Example

- ▶ Imagine 4 hospitals, each with 50 eligible patients
- ▶ Randomize hospitals into two groups, two in each group
- ▶ Suppose outcome is compliance for a specific medication, rated from 0% to 100%
- ▶ Results:
  - ▶ Intervention group: 70% compliance ( $SD = 5\%$ )
  - ▶ Control group: 50% compliance ( $SD = 5\%$ )
- ▶ Naive  $p$ -value is (very nearly) zero

## Example (continued): Clustered



## Example (continued): Independence



## Summary on Example

- ▶ Conclusions depend on scenario
- ▶ If no variation within hospital  
→ Effective sample size is 4
- ▶ If all variation is due to patients alone  
→ Effective sample size is  $4 \times 50 = 200$
- ▶ Both are extreme cases, truth is often somewhere in-between

## Causes for dependence

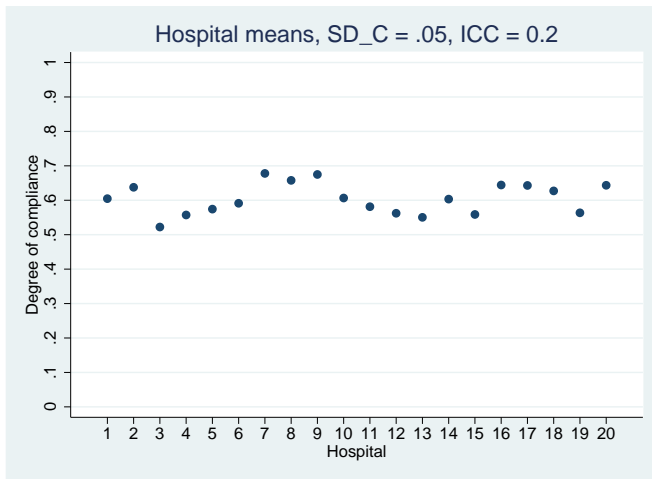
- ▶ Patients sharing physician:
  - ▶ Treatment habits of physicians dominate patient symptoms
  - ▶ Different physicians attract different types of patients
  - ▶ Diagnostic ability varies from physician to physician
- ▶ Patients sharing neighborhoods
- ▶ Patients sharing hospitals
- ▶ ...
- ▶ Often impossible to identify cause for dependence
- ▶ More important: Dependence/clustering can rarely be ruled out *a priori*

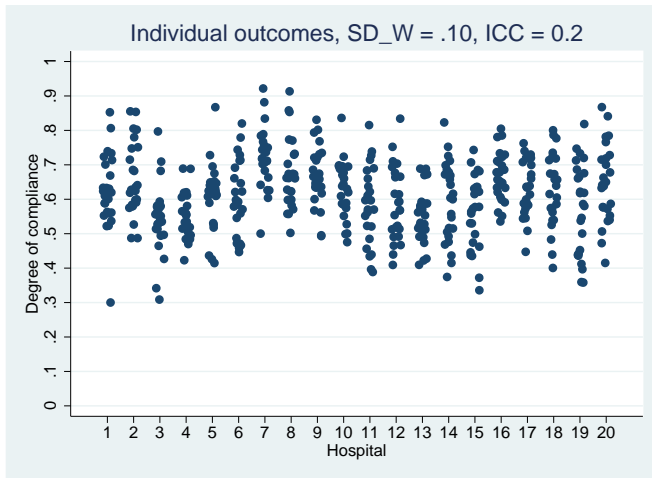
## Intra cluster correlation

- ▶ Variation between cluster means:  $SD_C$
- ▶ Variation within clusters:  $SD_W$
- ▶ Intra-cluster correlation coefficient (ICC)

$$ICC = \frac{SD_C^2}{SD_C^2 + SD_W^2}$$

- ▶ Zero if no variation between clusters
- ▶ One if no variation within clusters, but only between





# Statistical analysis of clustered data

- ▶ Invalid approach
  - ▶ Ignore clustering, assume independence
- ▶ Valid approaches
  - ▶ Stratify on cluster (simple, generally least efficient)
  - ▶ Random effects model (complex, generally most efficient)
  - ▶ Robust variance estimation (often best compromise)

## The setting

- ▶ Randomize in clusters (physician, hospital, neighborhood)
- ▶ Assume we have determined the usual quantities (power, significance level, clinically relevant difference)
- ▶ Further: know variation between and within clusters

$SD_C$  : Standard deviation of cluster means

$SD_W$  : Standard deviation within clusters

- ▶ Or equivalently

$ICC$  : Intra-cluster correlation coefficient

$SD_W$  : Standard deviation within clusters

- ▶ Challenge: What sample size do we then need?

## Correction with Design Effect (DEFF)

- ▶ Simple two step procedure
  1. Estimate sample size with the usual formulas
  2. Correct sample size with

$$N_{\text{cluster}} = \text{DEFF} \times N_{\text{ordinary}}$$

where

$$\text{DEFF} = 1 + (m - 1)ICC$$

where  $m$  is number of individuals in each cluster

## Restrictions of analytical approach

- ▶ Formula only valid when
  - ▶ outcome is normally distributed
  - ▶ compare two groups
- ▶ Extensions exist, but not for complex designs

## Review of power

- ▶ Definition of power

**Power level:** Probability of declaring a difference **when there is one**

- ▶ Can be interpreted directly if study is repeated, say 1,000 times
- ▶ Estimated power is proportion of simulated datasets yielding statistical significance

## Idea for estimating power by simulation

- ▶ Idea: Repeat the following  $M$  times<sup>1</sup>:
  1. Generate a dataset
  2. Test hypothesis of interest
  3. Count if dataset yielded statistical significance ( $m_S$ )
- ▶ Power is then estimated as

$$\widehat{1 - \beta} = \frac{m_S}{M}$$

---

<sup>1</sup>  $M$  can be determined from sample size considerations, ie. how precise do we want to estimate power?

## Remarks

- ▶ No direct estimation of sample size
- ▶ Forces joint consideration of
  - ▶ Sampling mechanism
  - ▶ Dependence structure
  - ▶ Distributional shape
  - ▶ Model for analysis
- ▶ Key point: Is extremely flexible

## Setting

- ▶ Study objective: Improve persistence with anti-psychotic medication
- ▶ Two treatment arms, intervention is re-organization and motivational patient interviews
- ▶ Randomize ten hospital wards
- ▶ Main outcome is sum of a ten item scale, each item scored 0/1
- ▶ Study is a before-after study, six months follow-up
- ▶ Not interested in clustering effect  
→ linear regression with robust variance estimation

## Key quantities

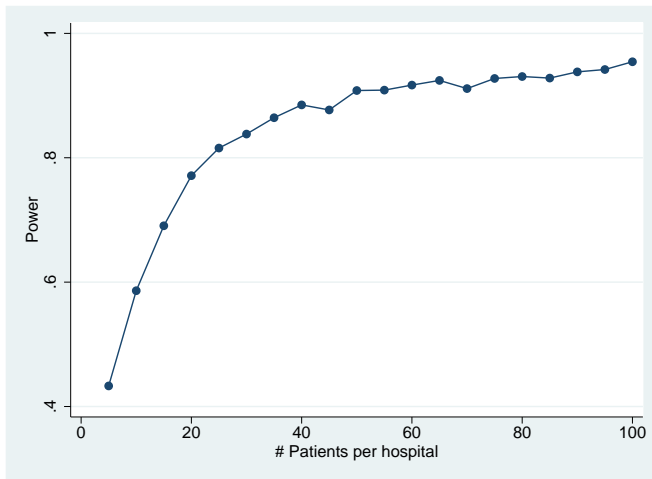
- ▶ Minimal relevant difference in mean change: .5 point
- ▶ Mean change in control group: .75 point
- ▶ Standard deviation of individual baseline scores: 2point
- ▶ Standard deviation of individual changes: 1 point
- ▶ Baseline mean score: 3
- ▶ Standard deviation of hospitals baseline score: 0.25 points
- ▶ Standard deviation of hospitals mean change: 0.2 points

Note: Two (different!) clustering effects: Baseline and change

## Algorithm for generating datasets

1. Set number of patients
2. Generate randomization status for each hospital,  
 $T_j \sim \text{Bernoulli}(0.5)$
3. Generate mean baseline score for each hospital,  
 $Y_j \sim \mathcal{N}(3, 0.25)$
4. Generate mean change for each hospital,  
 $\Delta_j \sim \mathcal{N}(.75 + T_j \times .5, 0.2)$
5. Generate individual baseline scores,  $X_{ij} \sim \mathcal{N}(Y_j, 2)$
6. Generate individual changes,  $\delta_{ij} \sim \mathcal{N}(\Delta_j, 1)$
7. Truncate to interval  $[0; 10]$
8. Apply integer rounding to all individual scores and changes

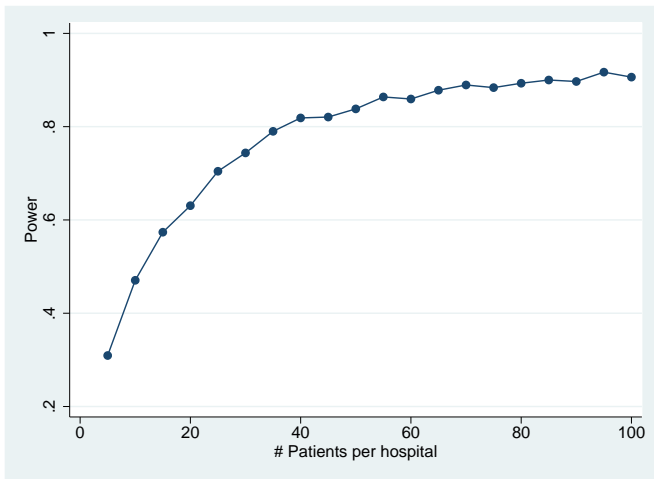
## Result



## Flexibility: Loss to follow-up

- ▶ What happens if people drop out?
- ▶ Assume it is not related to treatment
- ▶ Assume it varies by hospital
  - ▶ Variation in follow-up is normal distributed
  - ▶ On average 60% are followed until end
  - ▶ Standard deviation between hospitals is 5%

## Result









## Take home messages

- ▶ Valuable to do power calculations
- ▶ Clustering should not be ignored—neither in design, nor in analysis
- ▶ In simple settings: Use formula based on Design Effect
- ▶ More complex settings: Simulation approach
  - ▶ More time consuming
  - ▶ Estimates power, not sample size
  - ▶ General purpose
  - ▶ Much more flexible

## Remember

Sample size calculations are always wrong

-  Donner, A., K. S. Brown, and P. Brasher (1990, December).  
A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979-1989.  
*Int. J. Epidemiol.* 19, 795–800.
-  Kerry, S. M. and J. M. Bland (1998, February).  
Statistics notes: Sample size in cluster randomisation.  
*BMJ* 316, 549.
-  Klar, A. and N. S. Donner (2000).  
*Design and Analysis of Cluster Randomisation Trials in Health Research.*  
A Hodder Arnold Publication.

-  Reading, R., I. Harvey, and M. Mclean (2000).  
Cluster randomised trials in maternal and child health:  
implications for power and sample size.  
*Arch Dis Child* 82, 79–83.
-  Ukoumunne, O. C., M. C. Gulliford, S. Chinn, J. A. C.  
Sterne, P. G. J. Burney, and A. Donner (1999, August).  
Methods in health service research: Evaluation of health  
interventions at area and organisation level.  
*BMJ* 319, 376–379.
-  Wears, R. L. (2002, April).  
Advanced statistics: Statistical methods for analyzing  
cluster and cluster-randomized data.  
*Acad Emerg Med* 9, 330–341.

Thank you for the invitation  
Thank you for your attention!

Slides prepared with L<sup>A</sup>T<sub>E</sub>X and Beamer