



随机评估设计的样本量

Jed Friedman

世界银行**SIEF** 地区影响评估研讨班

北京, 中国

07, 2009

Adapted from slides by Esther Duflo, J-PAL

随机评估的样本量估算

- 问题:

要可靠地测量一个特定效果的水平,需要多大的样本量?

- 这里“可靠(Credibly)”的含义是;

可以有充分理由认为项目是导致项目干预组和对照组之间差异的原因。

- 随机化可以避免偏移,但是不能排除“干扰(noise)”

- 可以避免偏移,是因为样本量足够大,那么需要多大的样本量呢?

基础依据

- 在一个实验结束时，我们会比较干预组和对照组之间有趣的结果。

- 我们所感兴趣的差异：

$$\begin{aligned} & \text{干预组的均数} - \text{对照组的均数} \\ & = \text{效应量} \end{aligned}$$

- 比如：在农村通过免费分发获得蚊帐的平均人数
V.S.
在农村通过成本补偿获得蚊帐的平均人数

估算

我们所观察的是样本,而非总体.

在样本的每个村子里,有一个蚊帐数.这个数据或多或少地与全人群的实际平均数接近,因为这个数据也是各类外部因素对蚊帐数影响的结果.

我们通过计算样本的平均数来进行均数估算

如果我们只有很少的村庄,那么均数就不够精确.这样当我们比较不同样本组均数的差异时,就不能断定这个差异来自实验影响还是由别的因素导致.

估算

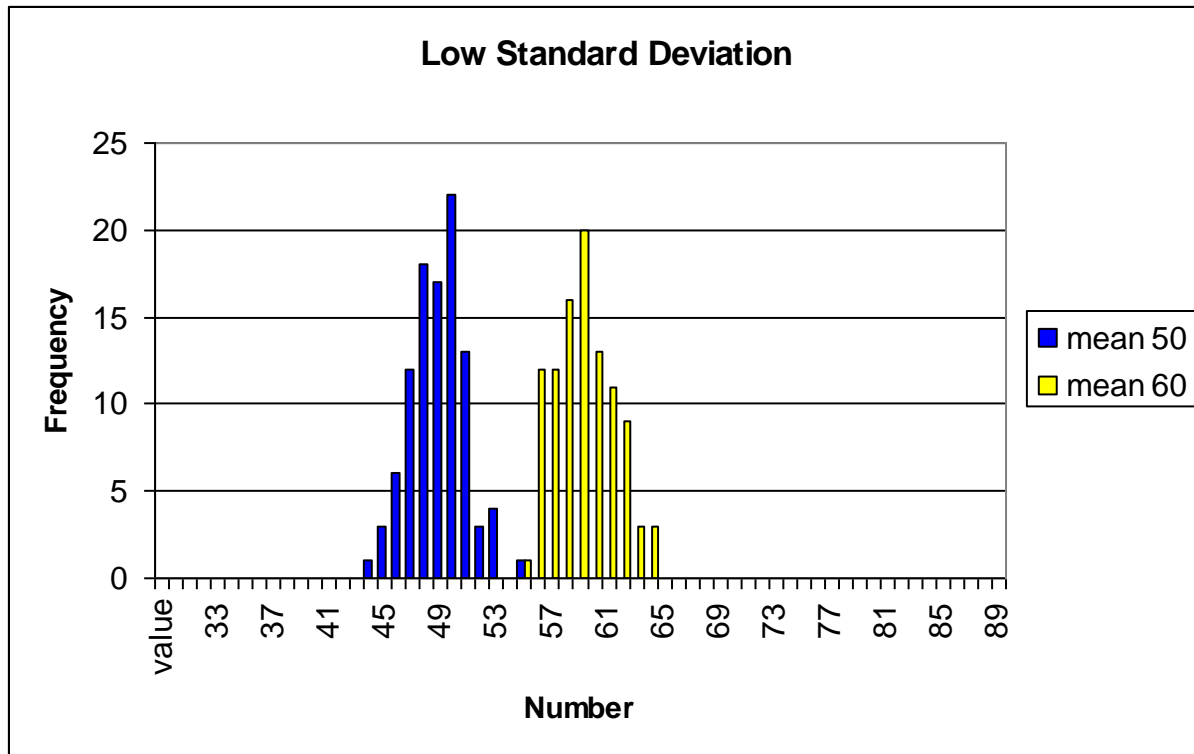
样本量:

- 如果干预组和对照组各有一个村庄,可以得出什么推断结果?
- 如果我们在实施疟疾干预时,把两个班级分别作为干预组和对照组,比较结果如何?
- (接上) 当这个班级人数较多时?
- 有效样本量的意义是什么? 比如,干预组和对照组的单位数量(既班级数), 上述疟疾项目中,在一个教室内,这个单位指什么?

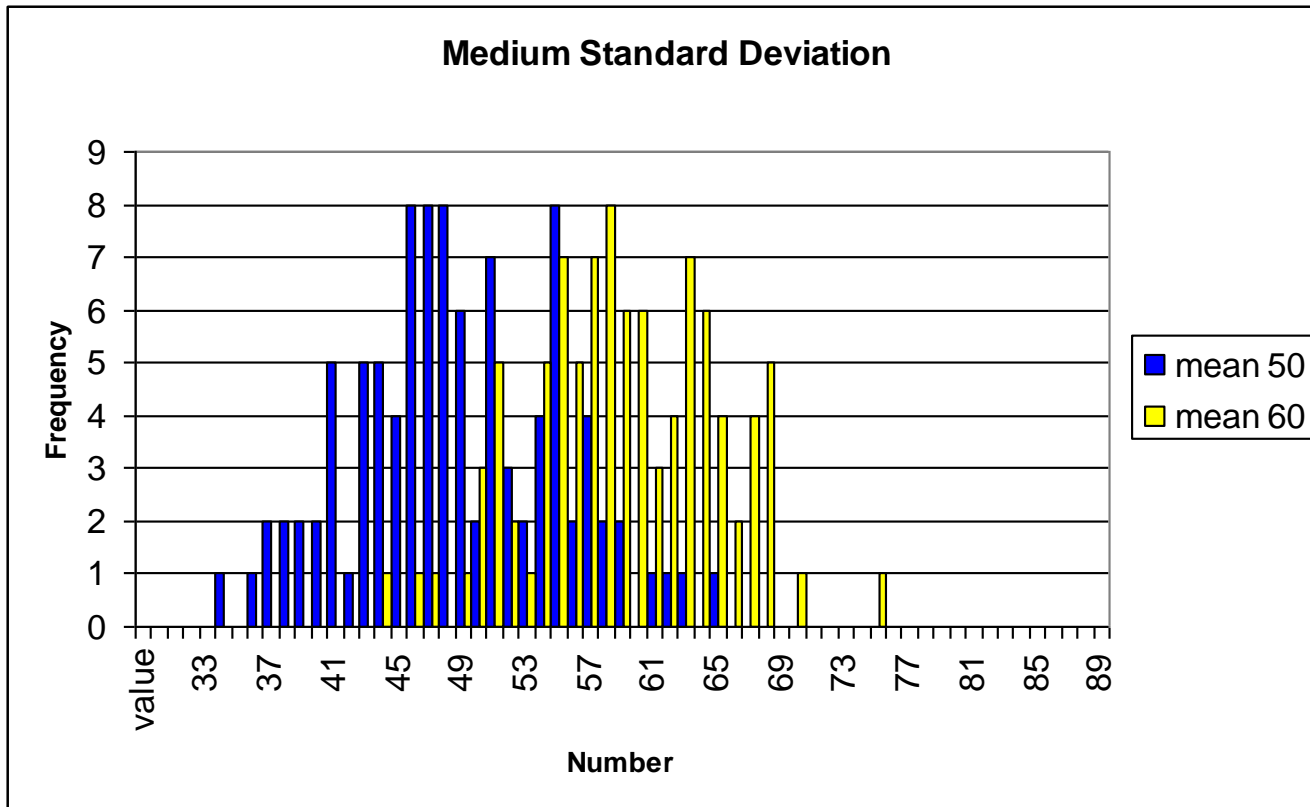
在结果变量中,我们试图测量变化程度

- 如果有很多其他没有被考虑到的因素能够解释我们的结果,就很难说我们采取的实验因素是真实有效的.

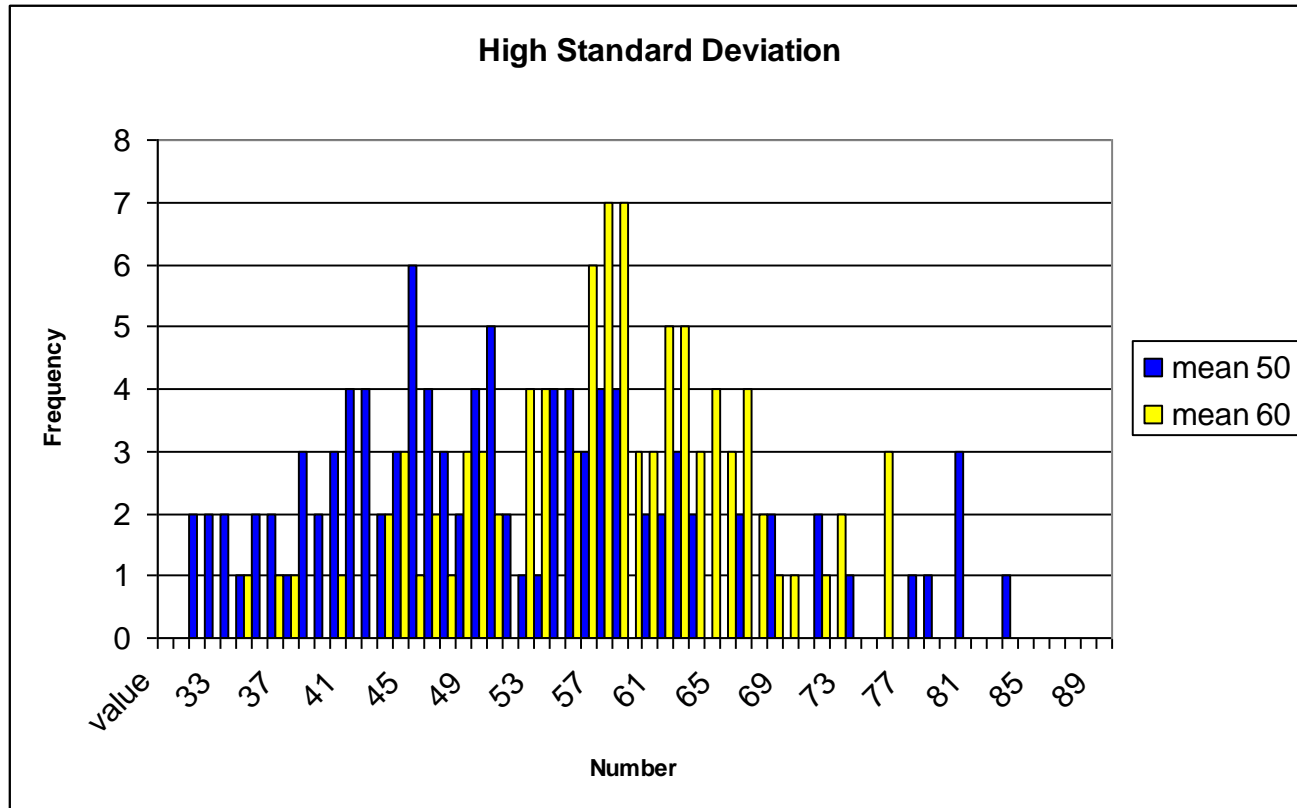
当结果很精确时标准差小



不够精确时标准差处于中等水平



能下结论吗?



可信区间

- 每次估算的效应量（样本均数间的差异）都只适用于该次抽样。每一次的抽样都会给出一个略有不同的答案。我们如何使用样本来估算总体呢？
- 效应量的95%可信区间是指，对于能够从同一总体中抽取得所有样本而言，95%的样本的效应量会落在这个区间里。
- 标准误（SE）同时涵盖了样本的大小和其结果的变异（在一个小样本中,SE会较大,它是可变的）
- 经验法则：95%可信区间大约坐落于均数的正负两个标准误之间。

假设检验

我们经常对检验效应量(effect size)是否为零感兴趣.目的是想否定没有实际效果的政策或项目.

这时,我们试图检验的假设是:

$$H_0 : \text{Effect size} = 0$$

对立假设是:

$$H_a : \text{Effect size} \neq 0$$

两类错误

- 第一类误差：认为有效果,但事实上没有效果 (H_0 事实上成立, 但被错误地拒绝了)

检验水准即错误地认为项目有效果但实际上没有效果的概率。

所以对于**5%**的检验水准，可以对结论的效度有**95%**的信心，即这个项目有**95%**的可能确实有效果。

对于政策目标，如果希望给出的答案更准确：这个水准可以设置的更低。

常用的检验水准有：5%, 10%, 1%.

可信区间的应用

- 如果零没有落在我们测算效应量的**95%**可信区间内，那么我们可以有**95%**的把握认为效应量不是零。
- 所以经验法则告诉我们：如果效应量大于标准误的两倍，可以有大于**95%**的把握认为项目或政策是有效的。

两类错误

第二类错误: 当项目事实有效果时, 没能够拒绝该项目没有效果的假设(H_0 假设项目没效果, 但在项目有效果时, 统计检验结果没有能够拒绝该错误假设)

- 检验效能即如果确实存在某种效应, 在我的实验中能够发现这种显著性效应的概率 (效能越高越好, 因为我更愿意报告一个真实的效应)。
- 检验效能是设计一项研究的工具。它告诉我对于一个特定样本量, 能够发现显著性效应的可能性有多大。
- 如果检验效能是负值, 则表明项目令人失望的可能性大小。

检验效能

- 设计评价时，通过一些初步的调查结果可以测算需要的最低样本量：
 - 检验一个预先指定的假设：效应是零或者不是零；
 - 一个预先制定的检验水准（如5%）；
 - 给出预先指定的效应量（即你认为这个项目将要做什么）；
 - 达到一定的检验效能水平。
- 80%效能的含义是：如果总体中确实存在某种效应，按某一确定的样本量进行抽样实验，80%的实验里可以检测出这种效应。
- 样本量越大，效能越高。

常用的效能为：80%，90%

在一个简单研究中计算检验效能的要素

所需要素	来源
显著性水平	通常设置为5%。该值越低，基于给定效能所需的样本量就越大。
两组间结果变量均数和变异	-与前期调研结果相似的设置 -变异越大，基于给定效能的样本量越大。
预期的检验效应量	可推动政策响应的最小效应是什么？ 我们希望检测出的效应量绝对值越小，基于给定效能的样本量需求越大。

选择效应量

- 为了推动项目被政府采纳,选择最小效应量应该按照以下方法:
 - 项目成本V.S.它所带来的效益
 - 项目成本V.S. 同量资金可能的别的用处
- 如果效应量小于上述要求,甚至是零.那么,一般人对这种检验一个非常小的效应与零不同的项目兴趣不会大.
- 相反,如果效应量大于这个最小效应量,则可以证明项目有价值可被采纳.这时,我们希望能够进一步检验区分它和零之间的不同。
- 共同危险: 效应量的选择过于乐观——样本量可能太小!

标准化效应量

- 基于给定的样本，可检测的效应量取决于结果的变异程度。
 - 例如：如果在没有实施项目时,所有的儿童都有一个非常相似的学习水平，那么即使一个很小的影响也很容易被监测到。
- 标准差可反映结果的变异。变异越大，标准差越高。
- 标准化效应量即效应量除以标准差。
 - $d = \text{效应量} / \text{标准差}$
- 常用的效应量：

$d=0.20$ (小) $d=0.40$ (中) $d=0.50$ (大)

影响效能的设计因素

- 随机程度
- 基线数据的可得性
- 控制变量及分层次的可得性
- 被检验的假设类型

整群随机抽样

整群随机抽样是以社会单元或群体而不是以个人的形式随机地分配到实验各组之中的一类实验。

比如:

条件转移支付	自然村
蚊帐分发	健康诊所
疟疾控制	学校
社会支持	家庭

采取整群随机抽样方法的原因

■ 需要尽量减少或消除“污染”

- 例如：在驱虫项目的研究中，由于蠕虫具有传染性，学校作为一个单位被抽样。

■ 基本的可行性因素

- 例如：如果对一个村庄的一些家庭进行了抽样，而另一些家庭没有抽中，那么 **PROGRESA** 项目就不具有政策上的可行性。

■ 唯一的必然选择

- 例如：影响整个课堂的任何教育干预（如翻转图，教师培训等）

整群随机抽样的影响

- 在一个抽样单位内所有个体的结果可能相关
 - 所有的村民暴露在同样的天气环境下；
 - 所有的病人有着共同的保健医生；
 - 所有的学生共有一位校长；
 - 村民们彼此相互影响；
- 样本量需要根据这种相关性进行调整
- 结果之间的相关性越大，我们需要调整的标准误差越大。

举例：群组乘数效应

组内 相关	随机分组大小			
	10	50	100	200
0.00	1.00	1.00	1.00	1.00
0.02	1.09	1.41	1.73	2.23
0.05	1.20	1.86	2.44	3.31
0.10	1.38	2.43	3.30	4.57

提示

- 随机抽取/分配足量的群组是极为重要的。
- 通常，组中独立个体数的影响小于组数的影响；
- 只有当群组数随机增加时，才考虑“大数定律”；
- 在地区层面不能保持随机化,比如各随机选择一个实验地区和一个非实验地区!!!!

基线的可及性

■ 基线有三个主要的作用:

- 可以检验在实施处理前，对照组和处理组的相同点和不同点；
- 减少样本量需求，但是需要在干预开始前做一个调查：涉及到成本问题；
- 可以用于分层，形成亚群组。

■ 计算基线的效能:

- 需要了解随后两次测量结果之间的相关性（如：两年的消耗量测算）
- 相关性越强，增益越大
- 足够大的增益需要持续获得结果，需大量人员参与。

可控变量

如果我们有额外的相关变量（如农村人口数，农村设置的街区数等），我们也能够控制这些变量。

对于效能有重要影响的,是控制了这些变量后的残差。

如果可控变量解释了大部分的变异，
那么精确度将会增加，且样本量的需求会减少。

警告：可控变量必须只包括那些不受项目措施影响的变量：
通常在干预前收集这些变量。

分层抽样

- 分层:通过可控变量值建立区组，在每个区组内随机化。
- 分层能确保通过这些可控变量平衡处理组和对照组。
- 分组可减小变异的两个原因：
 - 将减小每个层中结果的变异；
 - 群内各单元的相关性。
- 例如：如果对喷雾灭蚊计划按地区进行分层
 - 需要控制农业气候和相关流行病学因素；
 - 分层后，“同类地区政府效应”消失。

影响效能的设计因素

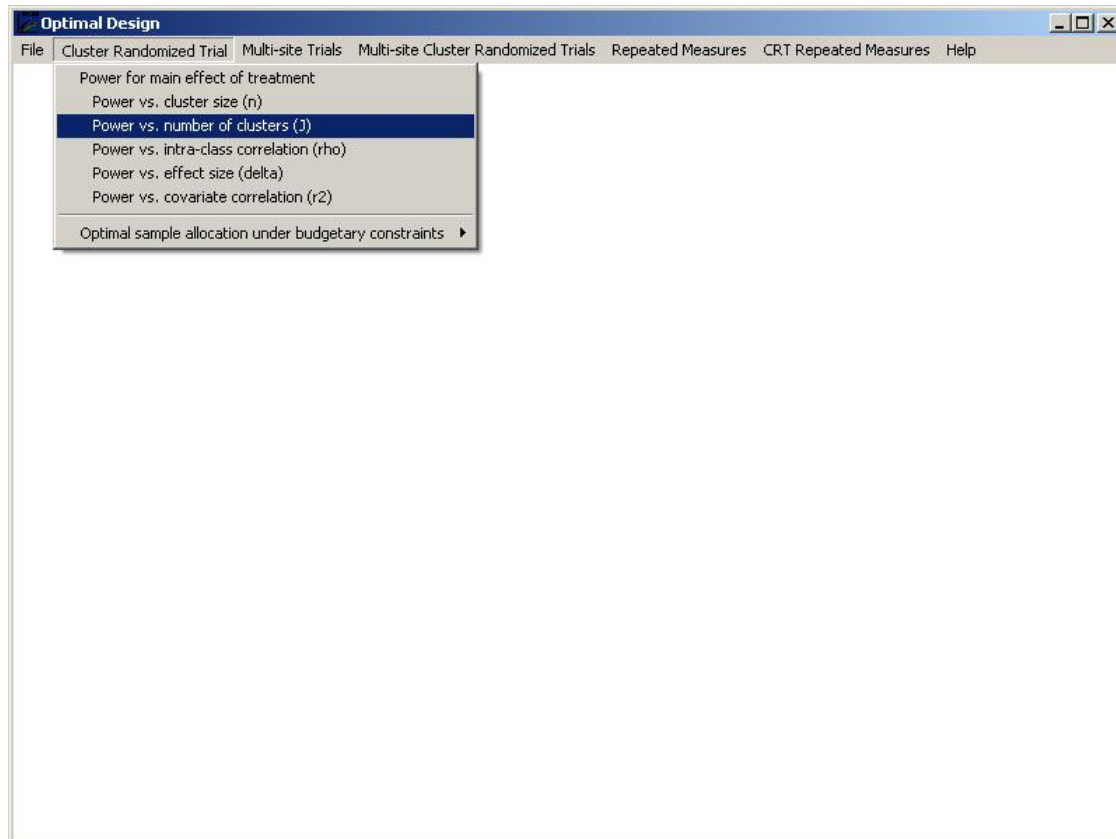
- 整群抽样设计
- 基线的可及性
- 可控变量和分层的可及性
- 假设检验的类型

检验假设

- 对处理组之间差异与处理组和对照组差异是否有同样的兴趣？
- 对处理组之间的交互作用是否感兴趣？
- 对于不同亚群间效应相同与否是否感兴趣？
- 设计是否仅仅包含了部分依从性？（例如激励设计）

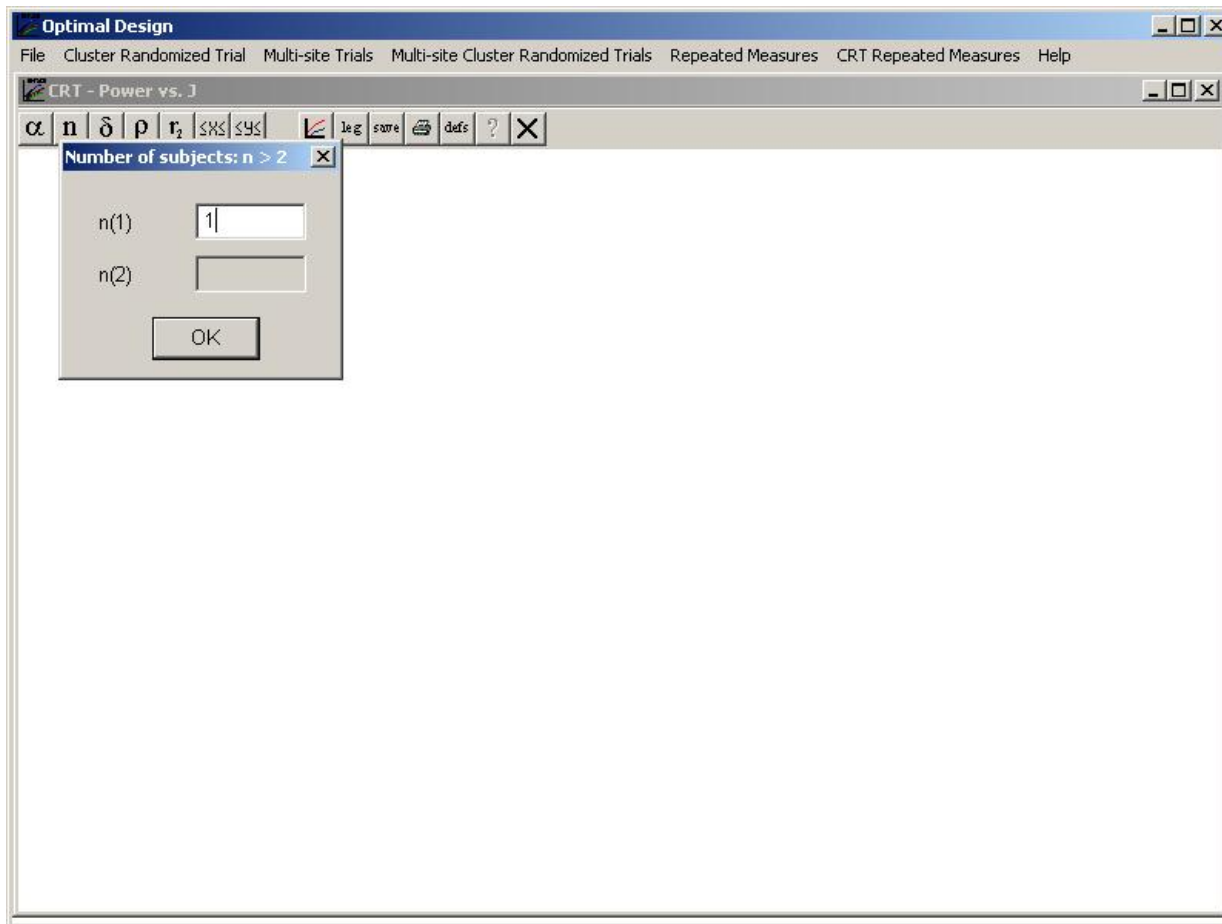
利用OD软件计算效能

- 在菜单“clustered randomized trials”，选择“Power v. number of clusters”



整群规模

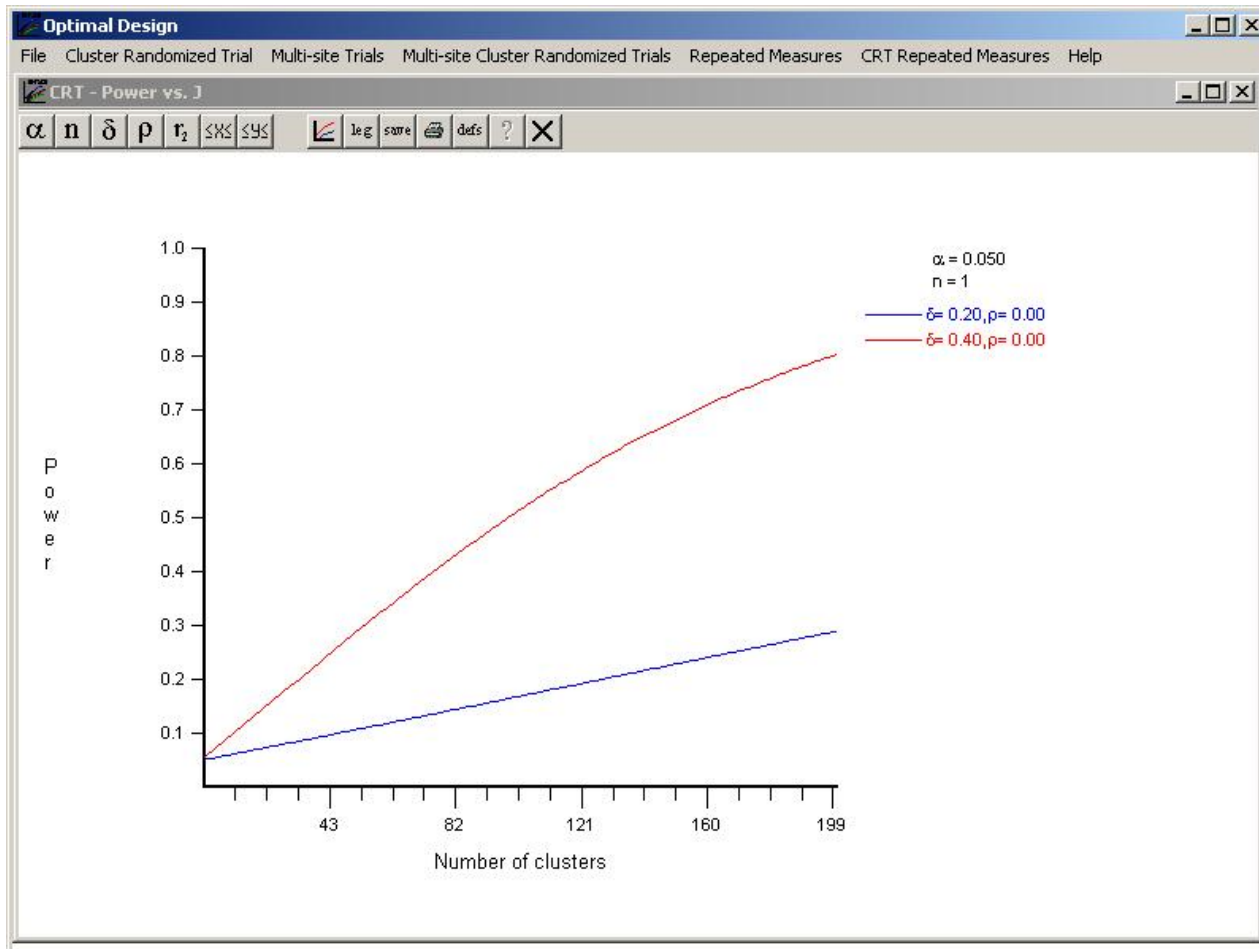
■ 选择整群规模



选择显著性水准，处理组效应及相关性

- 选择一个水准：
 - 通常选用 0.05
- 选择效应量d：
 - 可用0.20进行试验
- 选择组内相关性 (ρ)
- 获取显示样本量效能的结果图

效能和样本量



结论：实践中的效能计算

- 效能计算包含着一些推测。
- 有时缺乏准确运算所需信息
- 但是，在这些方面付出努力非常重要：
 - 避免实施无效能的研究：浪费时间和资金；
 - 将合适的资源投入到决定要做的研究中。