ABDUL LATIF JAMEEL
Poverty Action Lab
J-PAL EUROPE

TRANSLATING RESEARCH INTO ACTION

Power calculation

J-PAL Advanced course, Paris 2012

Prepared by Marc Gurgand and Roland Rathelot

Practical issue

Anticipate if sample size is enough for the expected effects

- Interviews and treatment are expensive and you have a budget
  $\rightarrow$ You want to optimize the sample size w/r to budget
- Sample size is limited by field constraints

Uncertainty

- It is all a matter of statistics: you want to minimize the odds that something disappointing happens
- You have to set acceptable probabilities and take the risk
- Additional source of uncertainty: depends on parameters whose values are unknown

Parameters of the power calculation

There is a true effect and you compare averages in control and treatment groups

The chances that those averages are significantly different depend (primarily) on:

- How large is the true effect
- Sample sizes
- Variance of the outcomes in the population
- Significance criterion
- Design

1. Significance test: reminder
2. Computing power
3. Imperfect compliance
4. Design effect
5. How to increase power

Notations

y: outcome

D: treatment dummy

N: sample size

Compare means of treated and untreated or OLS on:

$$y = c + \beta D + u$$

Standard error

Using usual formula or computing from means:

$$V(\hat{\beta}) = \sigma^2_{\hat{\beta}} = \frac{1}{\bar{D}(1-\bar{D})} \frac{V(u)}{N}$$

NB: adding controls

$$y = c + \beta D + X\gamma + u$$

Because $D \perp X$, does not affect $\hat{\beta}$ (asymptotically)
But reduces $V(u)$, so improves precision

Significance tests

Estimator $\hat{\beta}$ asymptotically normal
with mean β and variance $\sigma^2_{\hat{\beta}}$

If $\beta = 0$, then, for a risk α (e.g. 5%) we can define $t_{\alpha/2}$ such that:

$$P\left(-t_{\alpha/2} < \frac{\hat{\beta}}{\sigma_{\hat{\beta}}} < t_{\alpha/2}\right) = 1 - \alpha$$

For $\alpha = 0.05$, $t_{\alpha/2} = 1.96$

If $|\hat{\beta}/\sigma_{\hat{\beta}}| > 1.96$, we can reject the null $\beta = 0$

The power of the experiment

If the policy has an impact, we want to be able to see it

Unless the effect is very small, we want to reject the null
but if a lot of imprecision (large estimator variance), we may fail to do so

Type II error: $\beta > 0$, but we fail to reject the null
(i.e. $\hat{\beta}/\sigma_{\hat{\beta}} < 1.96$)

This will happen sometimes, for some samples

Power= 1 minus the probability that this happens

The power of the experiment

Usual approach: set an acceptable power (often 80%)

Set a β that you feel you should be able to "see"

And figure out the sample size that ensures that power for a true effect β

Computing the power

$[\hat{\beta}/\sigma_{\hat{\beta}} < t_{\alpha/2}|\beta]$ is random

$$P\left(\frac{\hat{\beta}}{\sigma_{\hat{\beta}}} > t_{\alpha/2}|\beta\right) = \kappa$$

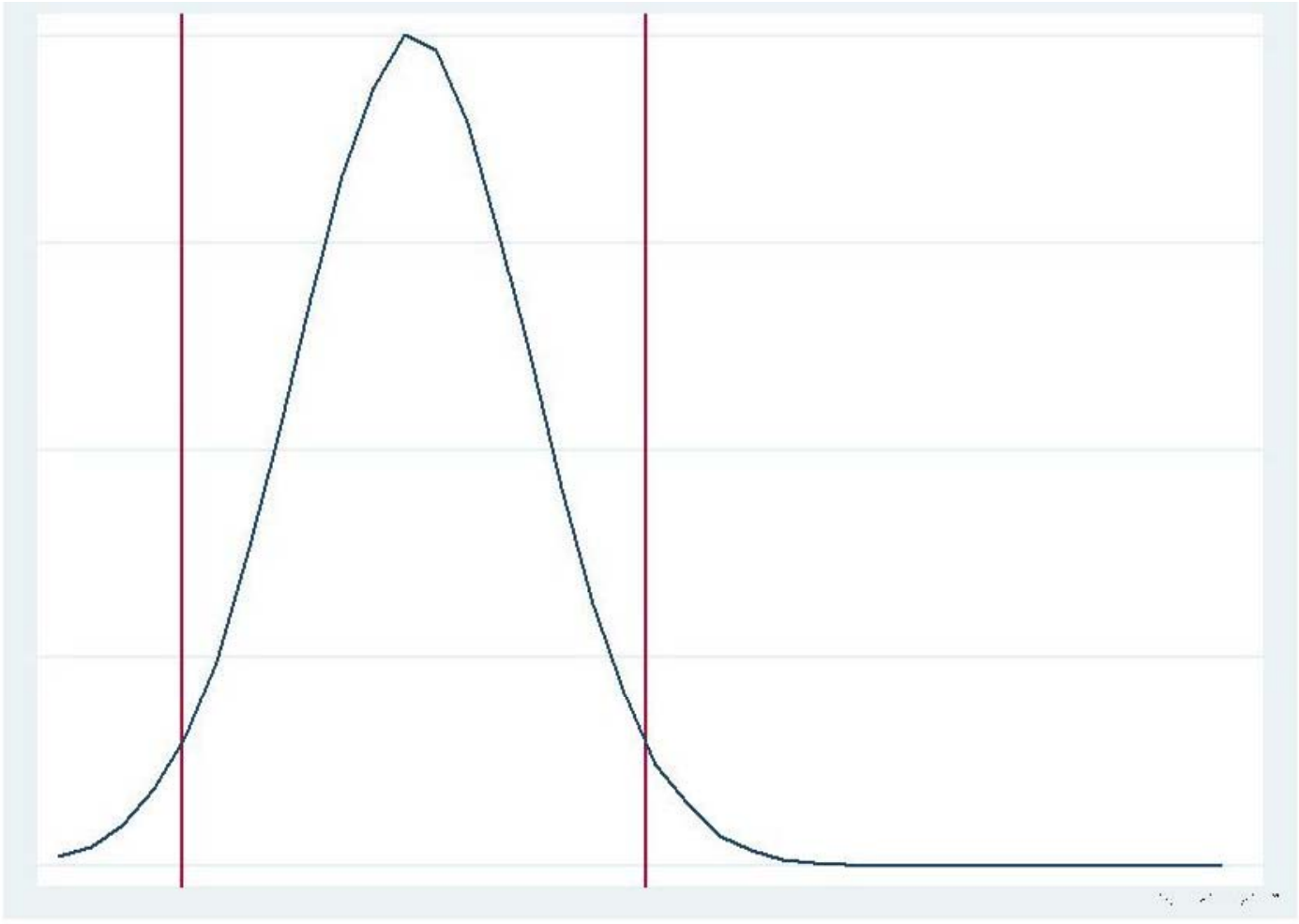
where κ is the power.

$$\begin{aligned} P\left(\frac{\hat{\beta}-\beta}{\sigma_{\hat{\beta}}} > t_{\alpha/2} - \frac{\beta}{\sigma_{\hat{\beta}}}|\beta\right) &= \kappa \\ \Phi\left(\frac{\beta}{\sigma_{\hat{\beta}}} - t_{\alpha/2}\right) &= \kappa \end{aligned}$$

Thus

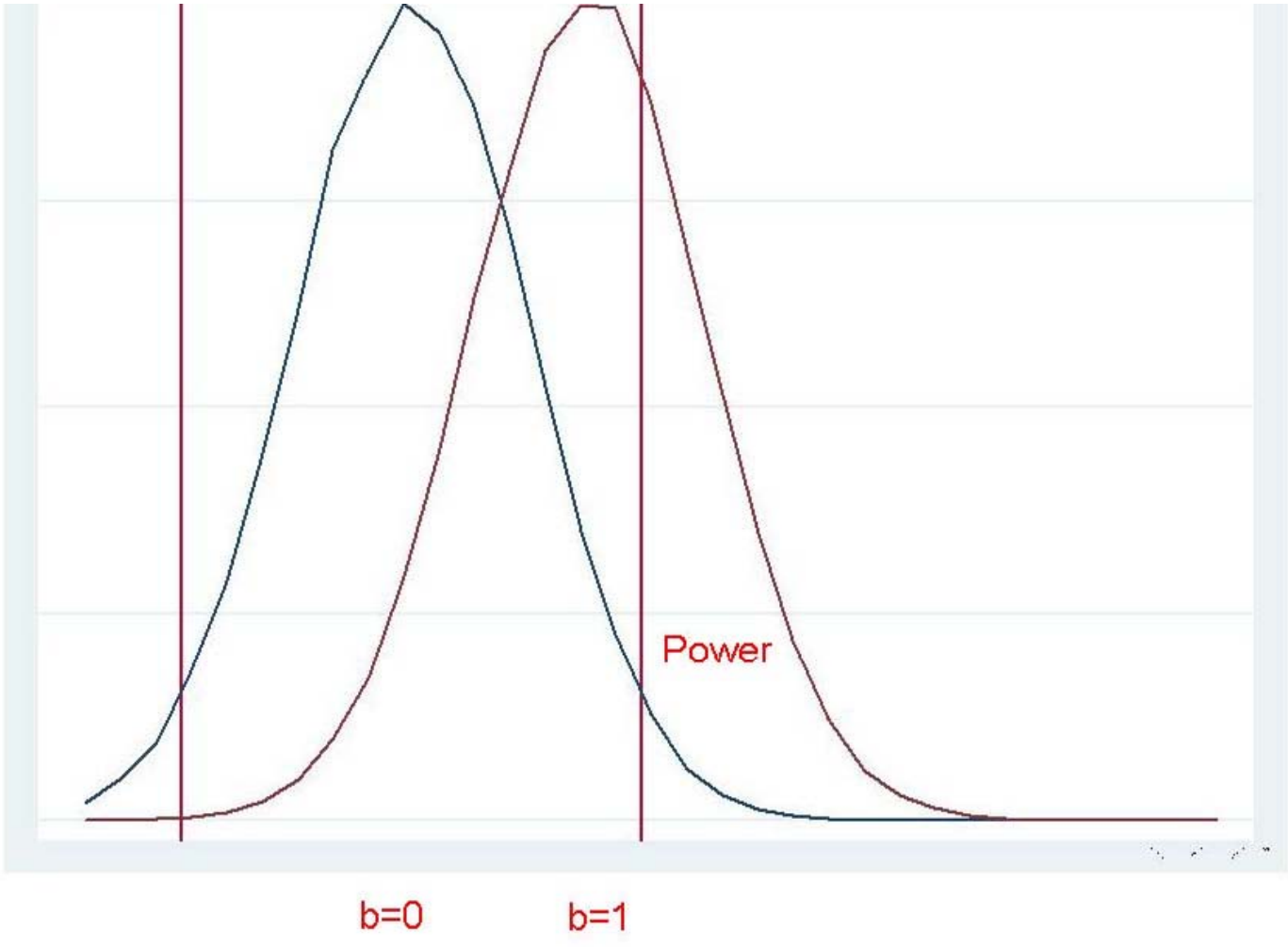
$$\frac{\beta}{\sigma_{\hat{\beta}}} - t_{\alpha/2} = t_{1-\kappa}$$

Power illustration

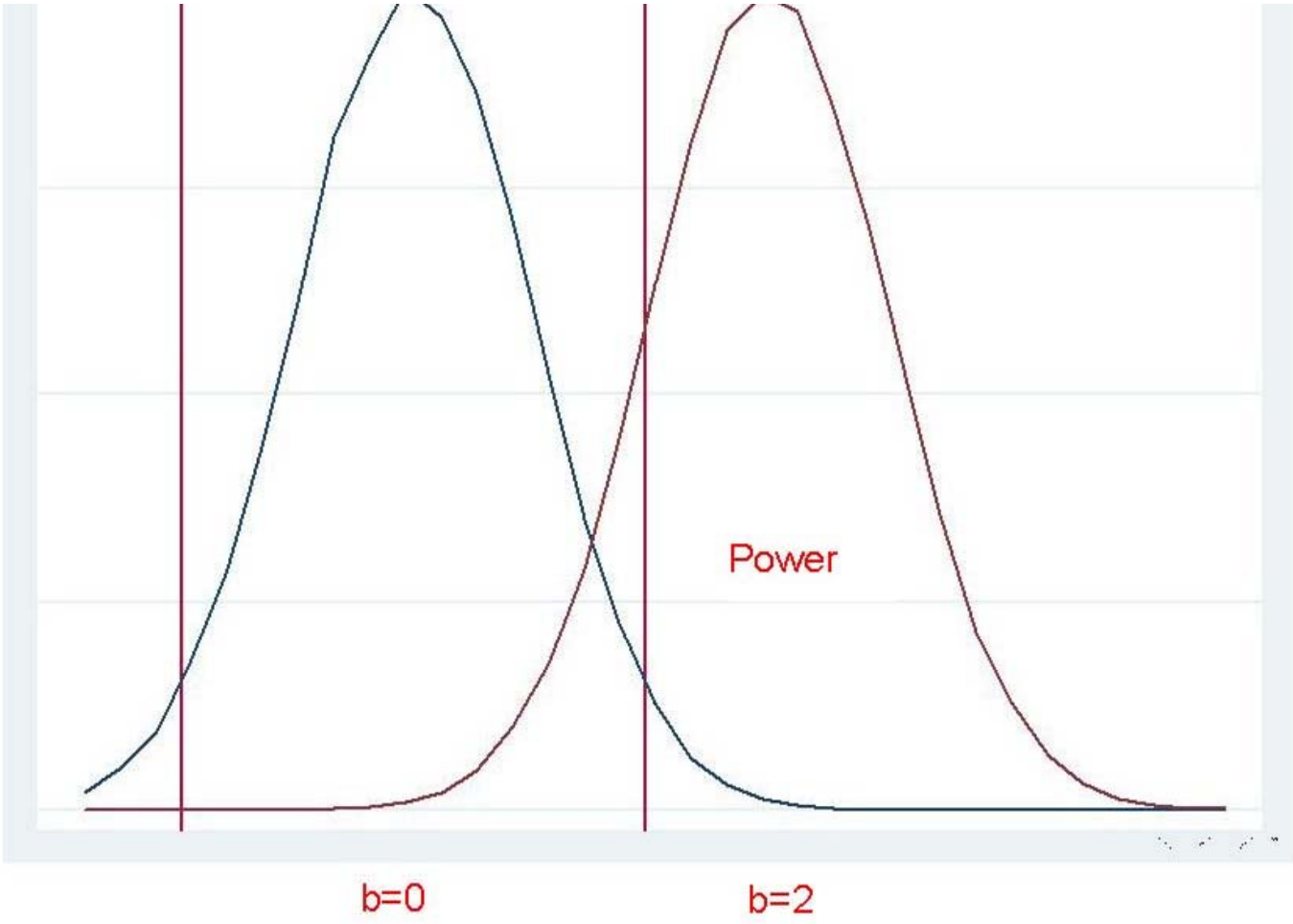


b=0

Power illustration



Power illustration



Minimum detectable effect

The β that will be "significant" 80% of the time (at 5% level) is such that:

$$\frac{\beta}{\sigma_{\hat{\beta}}} - t_{\alpha/2} = t_{1-\kappa}$$

or

$$\beta_{\text{MDE}} = (t_{\alpha/2} + t_{1-\kappa})\sigma_{\hat{\beta}}$$

with $t_{\alpha/2} = 1.96$ if $\alpha = 0.05$
and $t_{1-\kappa} = 0.84$ if $\kappa = 0.80$

$(t_{\alpha/2} + t_{1-\kappa})\sigma_{\hat{\beta}}$ is the **minimum detectable effet** (MDE)

Minimum detectable effect

Remember that

$$\sigma^2_{\hat{\beta}} = \frac{1}{\overline{D}(1-\overline{D})}\frac{V(u)}{N}$$

Thus

$$\beta_{\text{MDE}} = (t_{\alpha/2} + t_{1-\kappa})\sqrt{\frac{1}{\overline{D}(1-\overline{D})}\frac{V(u)}{N}}$$

MDE: main ingredients

$$\beta_{\text{MDE}} = (t_{\alpha/2} + t_{1-\kappa}) \sqrt{\frac{1}{\overline{D}(1-\overline{D})} \frac{V(u)}{N}}$$

1. Sample sizes
2. Variance of the outcomes in the population

   More homogenous = small differences unlikely in absence of treatment effect
3. Significance criterion
4. Design

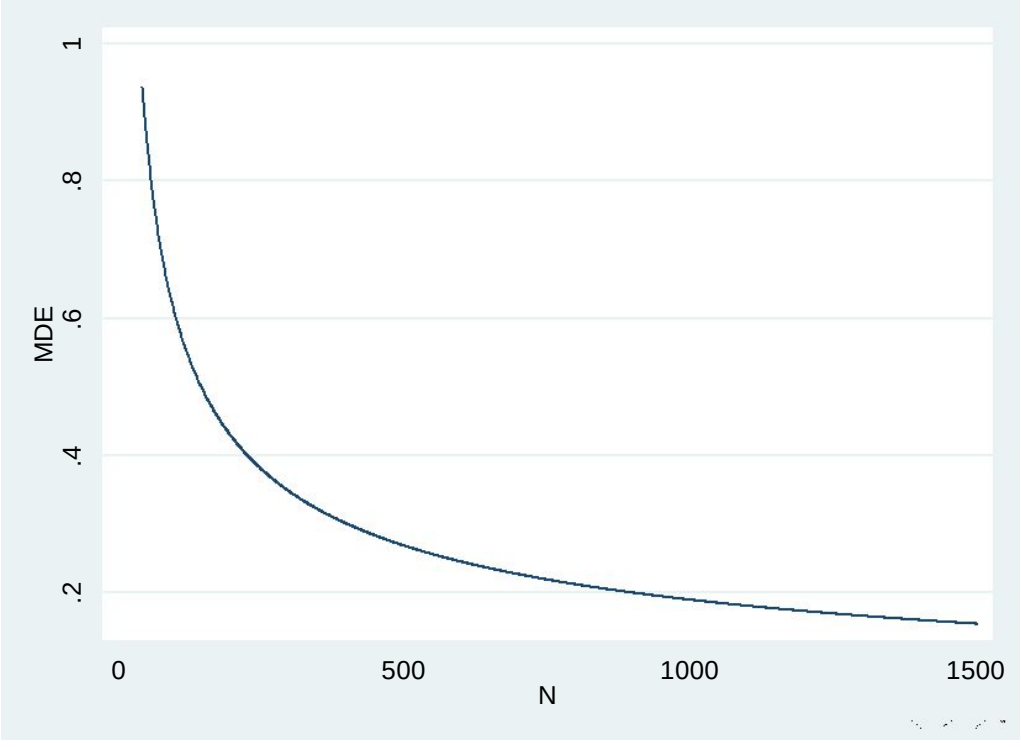
   A 50% treated is optimal. But more design issues to come

Effect-size

A usefull metric for MDE is the s.d. of the outcome

$$\frac{\beta_{\text{MDE}}}{\sigma_u} = (t_{\alpha/2} + t_{1-\kappa})\sqrt{\frac{1}{\overline{D}(1-\overline{D})}\frac{1}{N}}$$

Effect-size as a function of sample size in standard case



Effect-size

An effect size of…	is considered…	…and it means that…	Required N (under 50% treatment)
0.2	Modest	The median member of the treatment group had a better outcome than the 58th percentile of the control group	900
0.5	Large	The median member of the treatment group had a better outcome than the 69th percentile of the control group	144
0.8	VERY Large	The median member of the treatment group had a better outcome than the 79th percentile of the control group	56

Imperfect compliance

Whenever treatment assignment is not identical to treatment

- Encouragement design
- Some do not take up the program
- Unassigned do take up the program

Then randomization status can serve as an instrument for actual treatment

$\rightarrow$ Loss of power of the IV estimator

Power with instrumental variable

$$y = c + \beta T + u$$

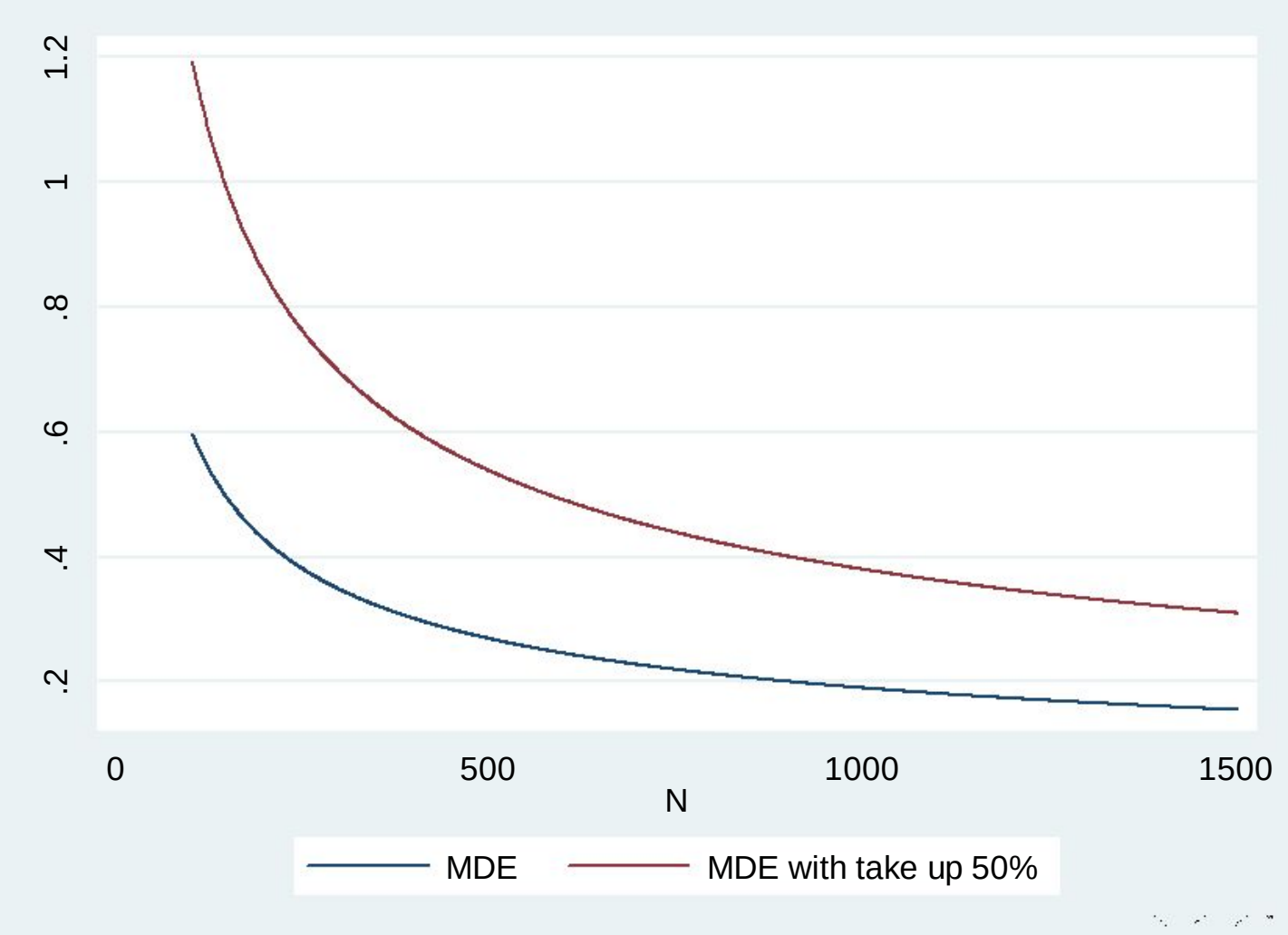
with T instrumented by D

Call π the net take up ($\pi = E(T|D = 1) - E(T|D = 0)$)

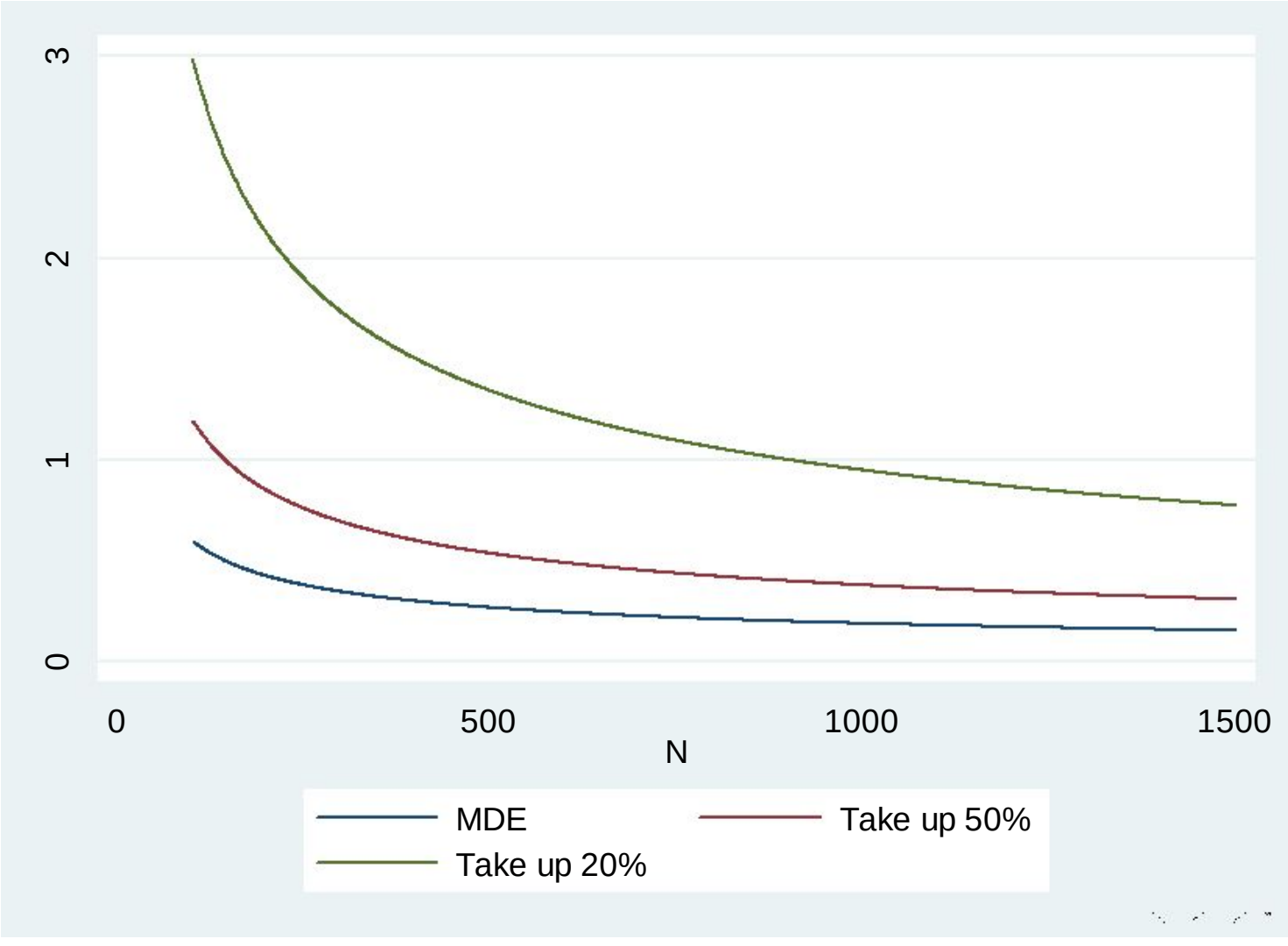
$$\sigma^2_{\hat{\beta}} = \frac{1}{\overline{D}(1 - \overline{D})} \frac{V(u)}{N} \frac{1}{\pi^2}$$

If take up is 50%, N must increase 4 times

Impact of imperfect compliance: 50% take-up



Impact of imperfect compliance: 20% take-up



Group-level randomization

Up to now, randomization at individual level
But we may want to randomize at higher group level:
Village, school, district, family

Individuals within each treatment group get the same treatment

May be required by field constraints or theoretical reasons

Impact of group-level randomization

- Outcomes within each group may be correlated (teacher effect, local event,)
- Each additional individual does not bring entirely new information
- At the limit, imagine all outcomes within a strata are exactly the same: effective sample size is number of strata, not number of individuals
- Precision will depend on the 2 sample sizes and the within correlation

Design effect

Decompose residual into group and individual component

$$y_{ig} = c + \beta D_{ig} + \mu_g + u_{ig}$$

$$\beta_{\text{MDE}} = (t_{\alpha/2} + t_{1-\kappa}) \sqrt{\frac{1}{\overline{D}(1-\overline{D})}} \sqrt{\frac{V(\mu)}{N_g} + \frac{V(u)}{N}}$$

where N is total and N_g is number of groups

Design effect

Call $\rho = V(\mu)/[V(\mu) + V(u)]$ the intracluster correlation

The higher ρ, the less additional information there is in an additional individual within each group

As compared to individual randomization, the loss in precision (thus the increase in MDE) is proportional to

$$D = \sqrt{1 + (n-1)\rho}$$

where $n = N/N_g$, the size of groups

Exemple of values of ρ in education

Madagascar Math + Language	0.5
Busia, Kenya Math + Language	0.22
Udaipur, India Math + Language	0.23
Mumbai, India Math + Language	0.29
Vadodara, India Math + Language	0.28
Busia, Kenya Math	0.62

How bad can it be?

Take $\rho = 0.2$

Think of an experiment with 50 schools and 30 pupils per school

$N_g = 50$, $n = 30$ and $N = 1500$

$D = 2.61$
Standardized MDE with individual randomization: 0.15
Standardized MDE with design effect: 0.15*2.61=0.39

Move from so-called small to so-called large

How to increase power

Asymptotically, randomization balances all characteristics between treatment and control

But at finite distance, not exactly true

We should force balancing at least against observed variables:
Ensure that T and C are balanced at least in terms of some x's

Efficiency gains

Intuitively: if you don't force balancing, x and D will be somewhat correlated

Introduces uncertainty as to whether differences are due to D or to x (even if you do control for x)

Efficiency gains: formally

Assume true model:

$$y_i = c + \beta D_i + \gamma x_i + u_i$$

Rewrite in difference to the mean to simplify the algebra

$$\begin{aligned} y_i - \bar{y} &= \beta(D_i - \bar{D}) + \gamma(x_i - \bar{x}) + u_i' \\ &= \beta \tilde{D}_i + \gamma \tilde{x}_i + u_i' \end{aligned}$$

Efficiency gains: formally

$$V(\beta, \gamma) = V(u')(M'M)^{-1}$$

with

$$M = (\tilde{D}, \tilde{x})$$

$$M'M = \begin{pmatrix} N\bar{D}(1-\bar{D}) & \sum \tilde{D}_i \tilde{x}_i \\ \sum \tilde{D}_i \tilde{x}_i & \sum \tilde{x}_i^2 \end{pmatrix}$$

NB:

$$\sum \tilde{D}_i^2 = N\bar{D}(1-\bar{D})$$

Efficiency gains: formally

$$M'M = \begin{pmatrix} N\bar{D}(1-\bar{D}) & \sum \tilde{D}_i\tilde{x}_i \\ \sum \tilde{D}_i\tilde{x}_i & \sum \tilde{x}_i^2 \end{pmatrix}$$

Asymptotically or if orthogonality is controlled,

$$\sum \tilde{D}_i\tilde{x}_i = 0$$

Efficiency gains: formally

Inverting $M'M$,

$$V(\beta) = V(u') \frac{\sum \tilde{x}_i^2}{N\bar{D}(1-\bar{D}) \times \sum \tilde{x}_i^2 - (\sum \tilde{D}_i \tilde{x}_i)^2}$$

- When $\sum \tilde{D}_i \tilde{x}_i = 0$, back to previous formula
- When $\sum \tilde{D}_i \tilde{x}_i \neq 0$, variance is larger

How to balance randomization

1. Stratification
2. Pair-wise matching
3. Re-randomization

Stratification

Randomize within strata

Simple, but limited number of variables can be used

$\rightarrow$ use variables that are strongly related to the outcome or for which subgroup analysis is desired

Pair-wise matching

Form pairs (one T, one C) using matching methods

Can rely on large number of variables

If implementation issues or non-response, can drop the affected pairs (external validity issues)

Re-randomization

A. Make a draw and test balance on observed variables
If one equality test rejected, redraw

B. Make 1,000 draws and keep the one that minimizes the maximum t-stat

Simulations

D. McKenzie and M. Bruhn, "In Pursuit of Balance: Randomization in Practice in Development Field Experiments", American Economic Journal: Applied Economics, 1(4): 200-32, 2009.

Use several real data and allocate an artificial treatment; compare randomization methods

- Below $N = 300$, balancing methods increase balance, including on future outcomes, and raise power
- Stratification and pair-wise better at reducing extreme imbalances
- Recommend to control for strata or pair dummies or x (if re-randomization)

Designs that imply stratification 1

Randomize classes within schools: for ex. half treated / half control

Schools act as strata:

- Balanced over all school characteristics (but not class characteristics)
- Include school dummies in the specifications

Designs that imply stratification 2

Randomize entire schools under a rotating design:

First graders in group A schools are treated date 1 and untreated date 2
First graders in group B schools are treated date 2 and untreated date 1

Political value: all schools benefit to the same level

If there is no dynamics (A is not affected by treatment in date 2)
T: (A,1) and (B,2) and C: (A,2) and (B,1)

- Balanced over all school and date characteristics
- Include school dummies and date dummies in the specifications