Workshop on Econometric Methods for Program Evaluation

Day 3: Designing randomized experiments

Institutions for Growth RPC

Kampala, Uganda
28–30 January, 2008

Outline

1. Statistical power of a randomized experiment
   - Why and how to conduct 'power calculations'
   - Imperfect compliance
   - Group-level randomization
   - Block randomization to improve power
2. Practical options for randomized experiments
   - Oversubscription
   - Phase-in
   - Within-group randomization
   - Encouragement designs

What do we want to know before conducting an experiment?

1. Experiments are costly, and we have a fixed budget. Is this going to be worth the money?
2. Suppose we believe that the policy intervention we want to study does indeed have positive effects. How big a sample do we need in order to be likely to be able to detect a "reasonable" effect?
3. How do alternative designs for our experiment affect our likely ability to detect an effect?

Statistical power of an individual-level randomization

- Following Duflo et al. (2007), consider estimating the equation

$$Y_i = \alpha + \beta T + \varepsilon_i, \tag{1.1}$$

  with perfect compliance. T takes on values of 0 and 1.
- Randomization ensures that T isn't correlated with ε_i (there's no endogeneity problem) $\Rightarrow$ simple OLS regression ok.
- Assume the ε_i independently and identically distributed, variance σ^2.
- Then the variance of the OLS estimator, $\hat{\beta}$, is given by

$$\mathrm{Var}(\hat{\beta}) = \frac{1}{P(1-P)} \frac{\sigma^2}{N} \tag{1.2}$$

  where P the proportion of treated, and N sample size.

Statistical power, cont'd

What does this variance mean? Recall that we use the distribution of $\hat{\beta}$ to test for statistical significance. Under the 'null hypothesis' that $\beta = 0$, $\hat{\beta}$ is distributed as

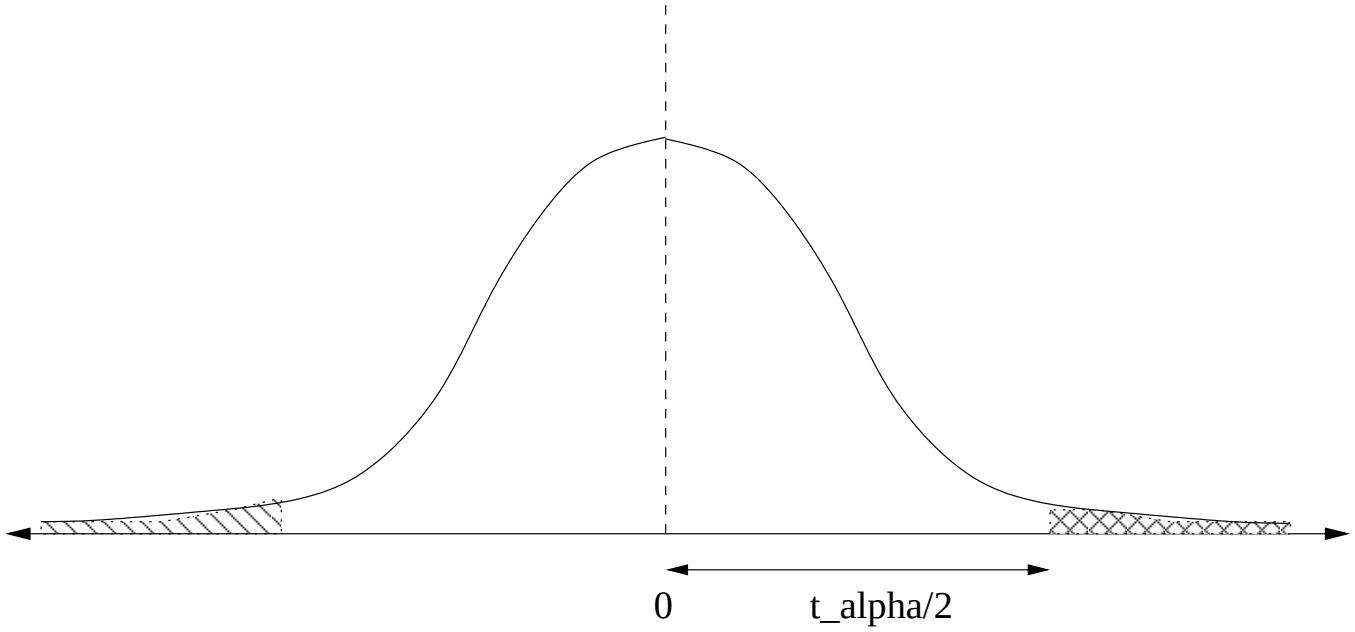


Figure: Testing the hypothesis that $\beta = 0$, at confidence level α

Statistical power, cont'd

Suppose that we *know* that the true coefficient is, say, 2. Then if you run the experiment many times, $\hat{\beta}$ will actually be distributed as

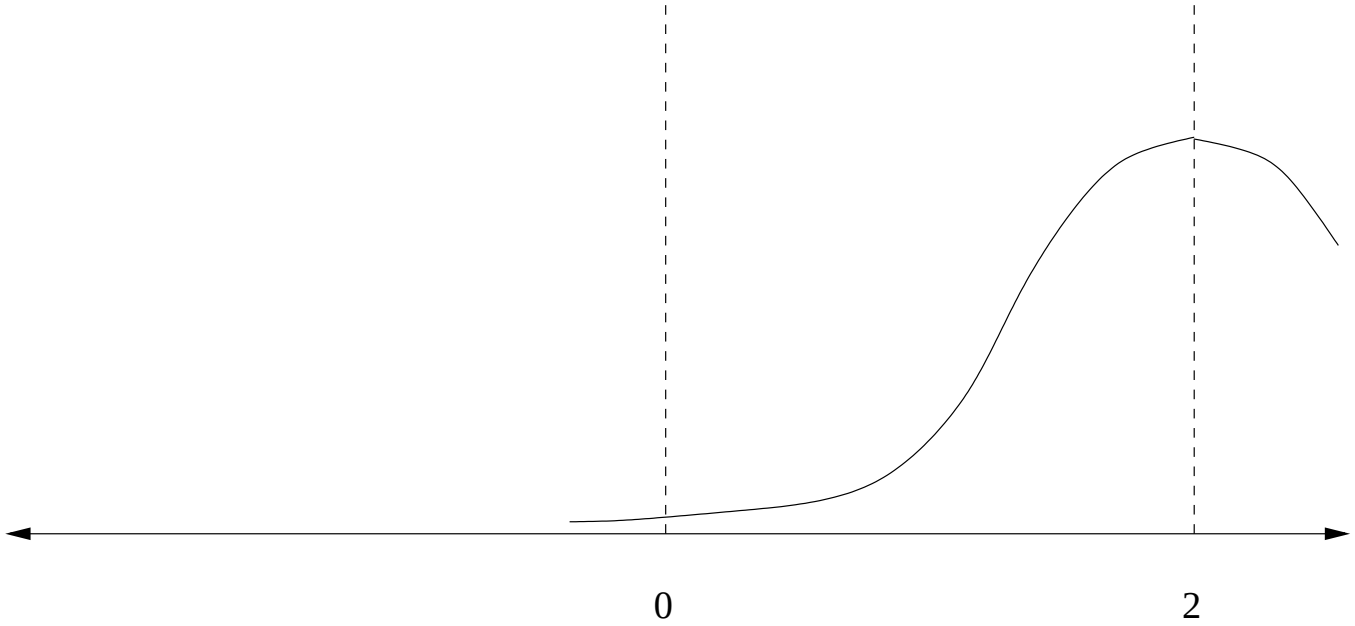


Figure: Distribution of $\hat{\beta}$ when truth is $\beta = 2$

Statistical power, cont'd

So the question we want to ask is: If the truth is β, for a given sample size, what is the likelihood that we will estimate a $\hat{\beta}$ big enough to reject the hypothesis that $\beta = 0$?

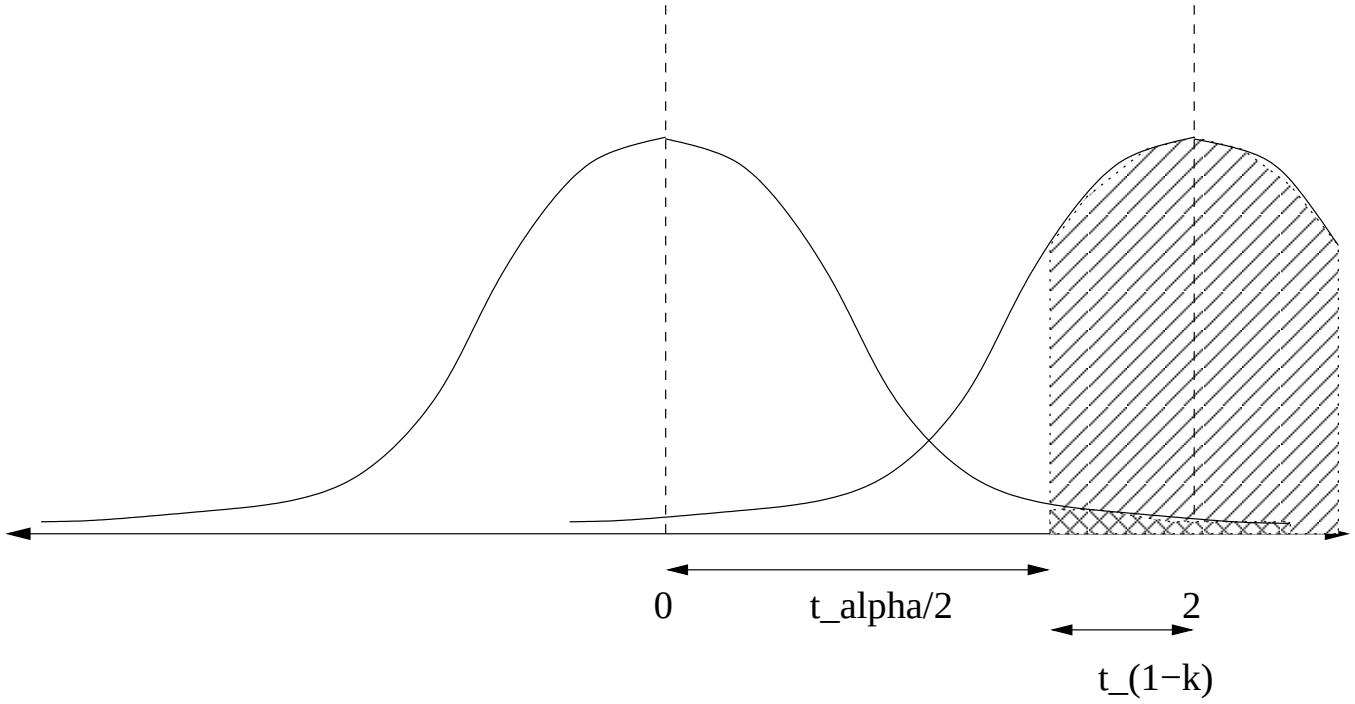


Figure: Probability of rejecting null hypothesis of no treatment effect

Question 1: For a given design, what is the smallest 'true' effect that one could expect to find?

- The last diagram illustrates that to achieve a power of κ at a confidence level of α, the *true* value of β must be at least

$$\beta > (t_{1-\kappa} + t_{\alpha/2})SE(\hat{\beta}) \tag{1.3}$$

  The smallest possible value of β that satisfies this is called the **minimum detectable effect** (MDE).
- Taking the square root of our earlier equation (1.2) for $\mathrm{Var}(\hat{\beta})$,

$$\beta_{MDE} = (t_{1-\kappa} + t_{\alpha/2})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}} \tag{1.4}$$

- So if we have data available to estimate σ^2, we can plug in our design parameters to solve for the MDE.

Question 1, continued

Key points about the MDE,

$$\beta_{MDE} = (t_{1-\kappa} + t_{\alpha/2})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}} : \tag{1.5}$$

- The more households, N, that you have, the smaller the effect you can detect.
- The more *balanced* your sample is between treatment and control, the smaller the effect you can detect.
- The smaller the variance, σ^2, of the error term, the smaller the effect you can detect. Note: if we have control variables, or even fixed effects (for example, dummies for regions), then what matters is the residual error *after controlling for these observables*.

Question 2: I believe the effect is β_0. How big a survey do I need in order to have a κ% chance of finding this effect?

- We can answer this question, too, by rearranging the formula for the MDE to get:

$$\underline{N} = \frac{(t_{1-\kappa} + t_{\alpha/2})^2}{P(1-P)} \frac{\sigma^2}{\beta_0^2} \tag{1.6}$$

- The higher the confidence level we want to use in testing for an effect (the smaller is α)...
- Or, the more sure we want to be that we will actually find this effect (the bigger is κ)...
- Or, the farther from 50/50 is the balance of treated and untreated households in our experiment...
- Or, the more unexplained variation in Y there is...

...the bigger the sample size we need.

Question 3: I have a fixed budget for the study. What is the best design to use?

- We've already seen that, for a given N, the MDE will be smallest when treatment and control groups are balanced.
- But in practice you might pay for the intervention and survey out of the same budget, B:

$$N(1-P)c_{survey} + NP(c_{survey} + c_{intervention}) \leq B, \qquad (1.7)$$

where c_{survey} is the survey cost per household, and $c_{intervention}$ the cost of the intervention.

- Economists will recognize this as an optimization problem:

$$\min_{N,P} MDE, \qquad \textit{subject to}\,(1.7) \qquad (1.8)$$

with solution $\frac{P}{1-P} = \sqrt{\frac{c_{survey}}{c_{survey}+c_{intervention}}}$.

Power calculations with imperfect compliance

- Yesterday's microfinance example illustrated an important problem: sometimes not everyone to whom it is offered takes up an intervention.
- In 'encouragement designs' more generally, we may have two variants on this problem:
  - It may be that only a fraction c of the *treatment* group (the 'compliers') actually receive the intervention;
  - or that a fraction d of those supposed to be in the *control* group (the 'defiers') receive the treatment anyway.

  If we have prior information (an informed guess) about c and s, then just make a simple correction:

$$MDE = \frac{1}{c-s}(t_{1-\kappa} + t_{\alpha/2})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}} \qquad (1.9)$$

What is the unit of randomization?

- Sometimes the randomized allocation of a treatment occurs at the level of a group rather than an individual.
- For example, the treatment might occur at school, village, or district level. All individuals in this group uniformly either do or do not receive the treatment.
- This means we can't use group fixed effects at that level to control for unobservables: there is no within-group variation in the randomized treatment (or encouragement).

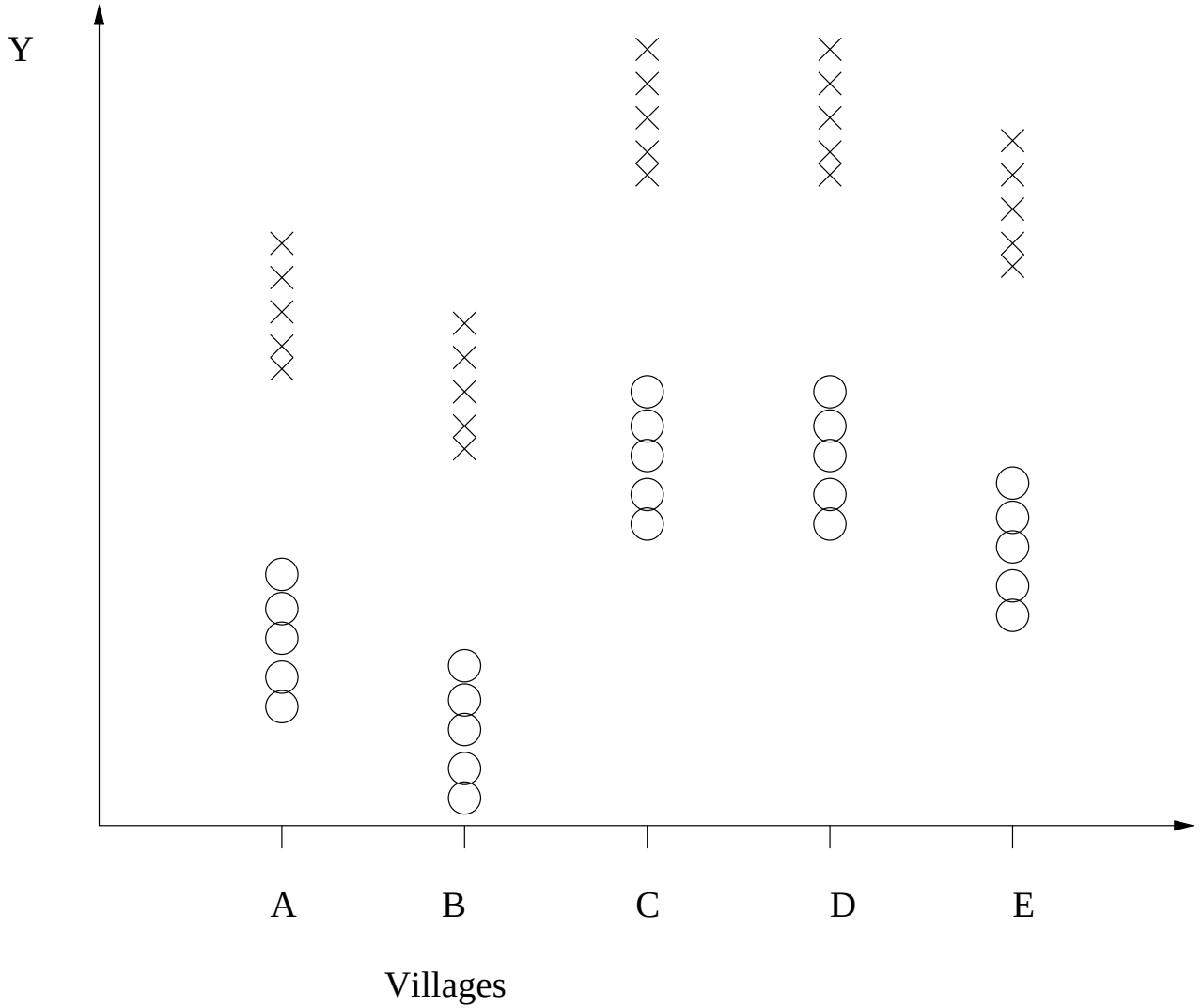
Power calculations under group randomization

- The power (or MDE) of the study in this case depends on how much of the variation in the outcome is explained by a common group-level shock, versus how much variation there is among individuals within the same group.
- Suppose the model we're estimating is

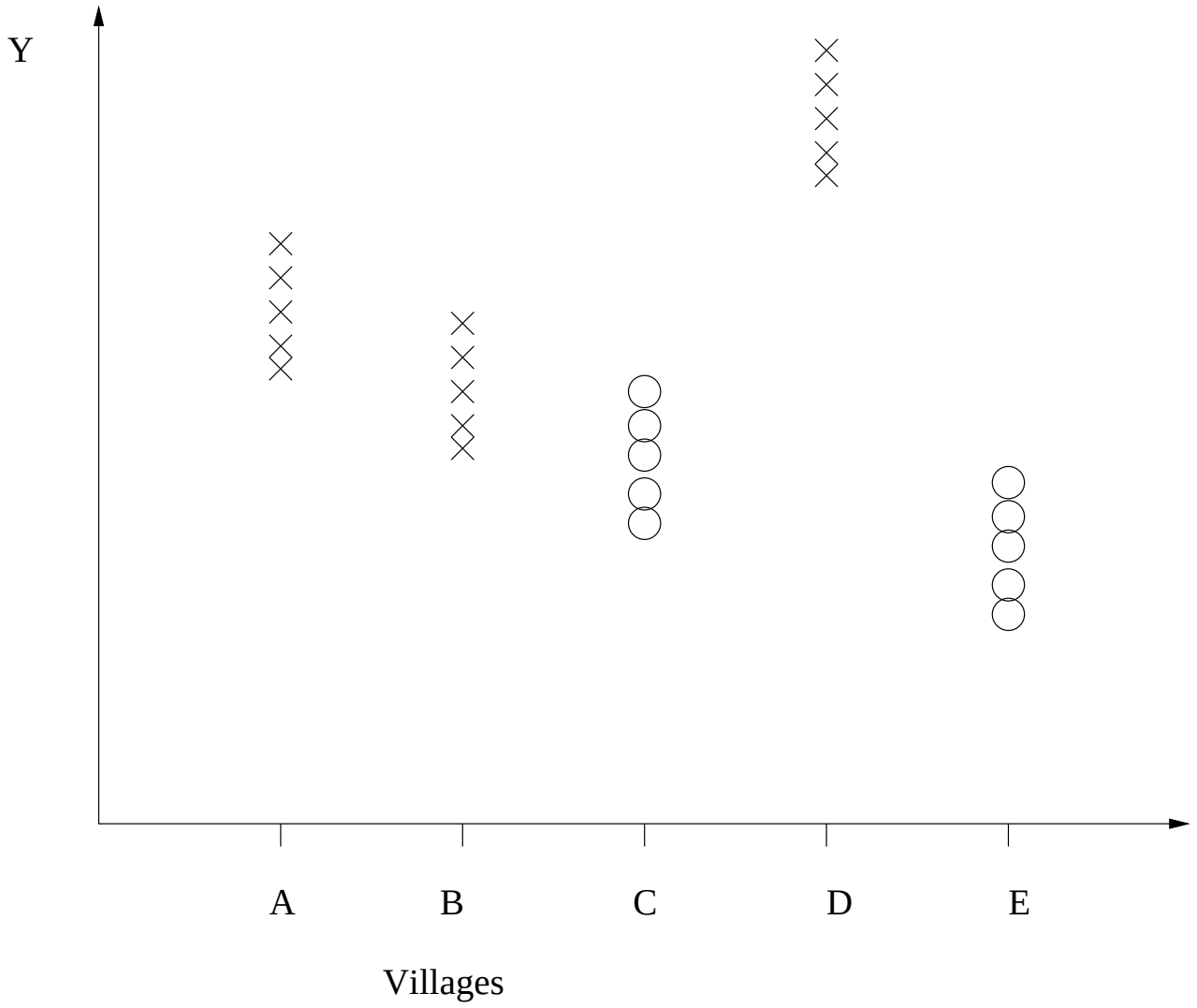
$$Y_{ij} = \alpha + \beta T + u_j + e_{ij} \tag{1.10}$$

with individuals denoted by i and groups by j. Thus the unobserved characteristic/shock u_j is shared by all individuals in this group.

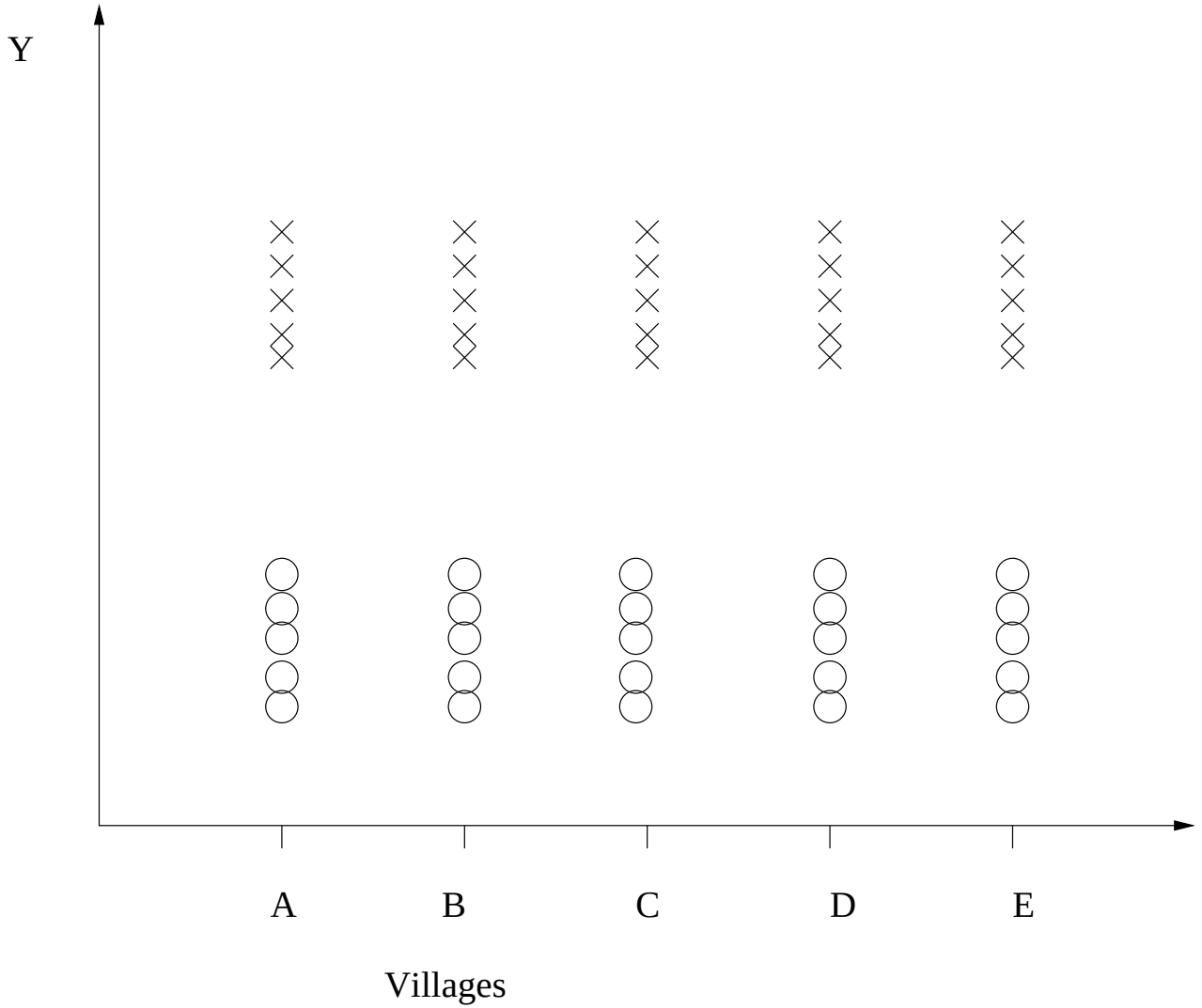
- What matters is the relative variance of these two error terms (σ_u^2 and σ_e^2).



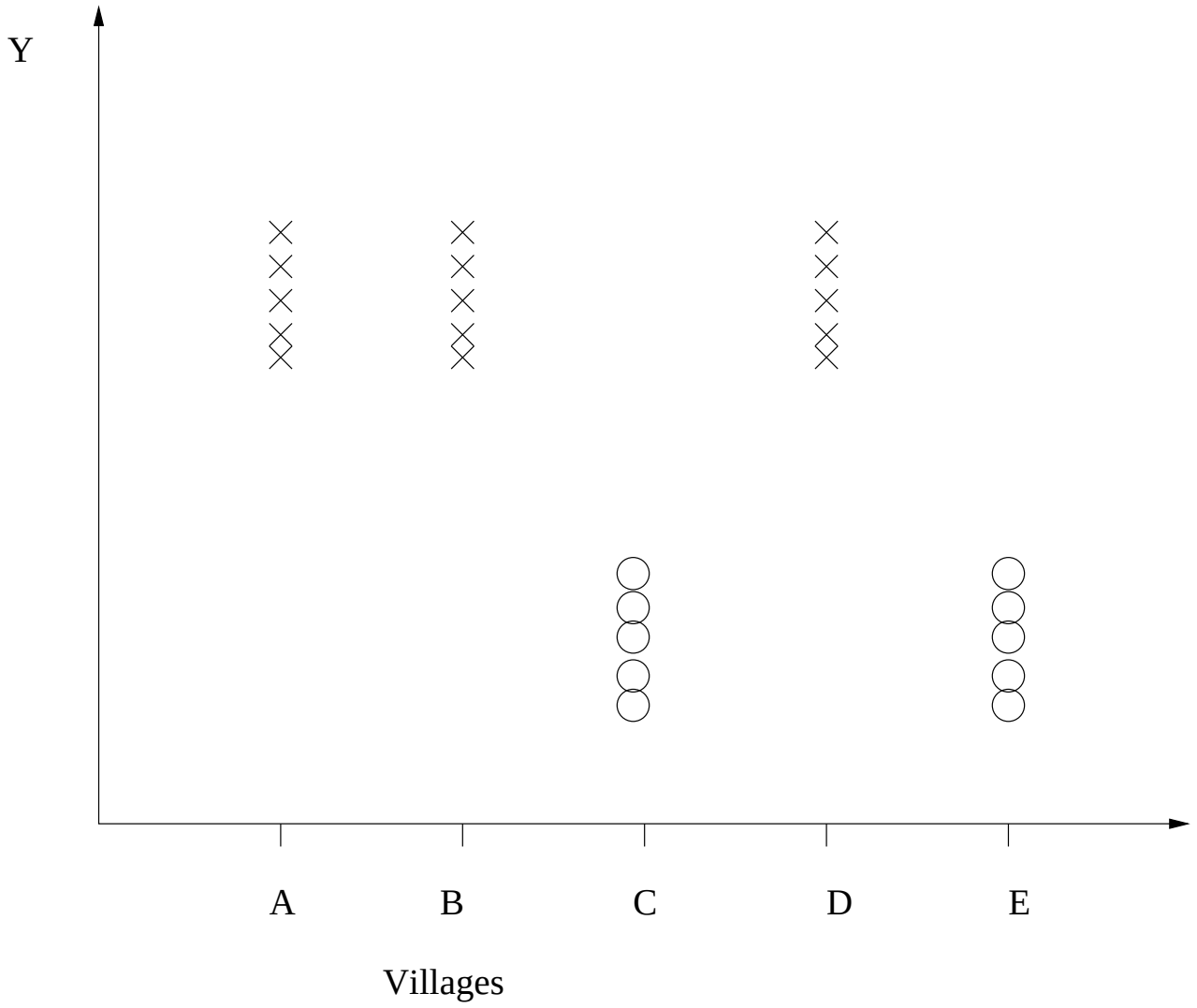
Hypothetical outcomes with/without treatment



Observed outcomes after randomization



Hypothetical outcomes with/without treatment



Observed outcomes after randomization

Power calculations for group randomization, cont'd

- Let $\rho = \frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2}$, the fraction of the residual variation explained by the group-level effect.
- Then if we let J be the number of groups and n be the number of individuals per group, the MDE becomes

$$\beta_{MDE} = \frac{t_{1-\kappa} + t_{\alpha/2}}{\sqrt{P(1-P)J}} \sqrt{\rho + \frac{1-\rho}{n}} \sqrt{\sigma_u^2 + \sigma_e^2} \qquad (1.11)$$

- Intuitively, as the variation in the u_j goes down, then group membership doesn't matter for individuals' outcomes. The MDE becomes equivalent to our original version for $\rho = 0$.

- Further design improvements are possible if we're particularly interested in the effect of the treatment on a particular subset of the population, say, women in rural villages.
- In this case, instead of doing a pure random sample, we can dedicate a fixed proportion of the sample
- In the extreme, you could think of your blocks as pairs of matched individuals—e.g., twins—and randomly allocate one of each twin to control and one to treatment.
  - A dummy variable for each sibling pair would then control for all unobserved characteristics
  - In the extreme, the *only* difference between the two would be attributable to the treatment...
- This approach is particularly valuable if you expect the effect of the treatment to be different across these groups.

Oversubscription designs

- **Oversubscription** designs use randomization as a rationing rule, to choose among a pool of eligible participants in a program.
- Strong case can be made on grounds of fairness in such designs.
- Ex: Angrist et al. assess the impact of a credit program by randomizing acceptances of applicants among a marginal group of prospective borrowers.

Phase-in as a route to randomization

- **Randomized phase-in** can be used when practical constraints mean a program will be introduced gradually over time.
- Randomly selected *late recipients* provide a control group for randomly selected *early recipients*. For example, 'Worms'.
- A few concerns:
  - Do people change their behavior in *anticipation* of a foreseen treatment? [How might this bias results?]
  - Are the effects of the treatment felt sufficiently quickly that they will occur before the late recipients are treated?

Not all phase-ins are created equal

- But note that not all phase-in is random. In spite of this, authors often employ such a 'pipeline comparison' method
- For example, Field (2005) looks at the issuance of property titles to slum residents in Lima. A snapshot halfway through implementation allows comparison of early with late recipients.
- But are late recipients a valid comparison group? Mitchell (2005); see also Conning and Deb (2007)
  - Early titled neighborhoods were done as a demonstration project by de Soto's ILD;
  - Early neighborhoods were more central and better off;
  - Late neighborhoods composed of refugees.

  Might these effect credit, investment, and labor supply?

Within-group randomization

- Key idea in **within-group randomization** is to randomize within *subgroups*, while ensuring that each group gets something.
- For example, Banerjee et al. (2007) allocate a teaching assistant to all schools, but randomly choose whether to give this to the 2nd or 3rd grade.

Encouragement designs

- **Encouragement designs** can be applied when everyone has access to a program.
- Think of this as generating an *instrument* for participation in a program. Key is that participation should be higher for those randomly *encouraged* to participate (by financial or other means) from those without such added incentives.
- Examples
  - Duflo and Saez (2003) provide (via letter) a financial incentive for individuals to attend meetings about a particular pension plan; those receiving this incentive are more likely to attend a meeting and, ultimately, more likely to sign up.
- Encouragement designs are very flexible, and a way of introducing an exogenous source of variation in treatment when *de jure* access is universal.
- Key is to find an incentive that works. . .

References I

Duflo, E., R. Glennerster, and M. Kremer (2007): "Using Randomization in Development Economics Research: A Toolkit," Centre for Economic Policy Research, Discussion Paper No. 6059.

Duflo, E., and E. Saez (2003): "The role of information and social interactions in retirement plan decisions: evidence from a randomzied experiment," *Quarterly Journal of Economics*, 118(3), 815–842.

Field, E. (2005): "Property Rights and Investment in Urban Slums," *Journal of the European Economic Association*, 3(2/3), 279–290.

Mitchell, T. (2005): "The work of economics: how a discipline makes its world," *European Journal of Sociology*, 6(2), 297–320.