ATAI-IITA IMPACT EVALUATION WORKSHOP

Managing and minimizing threats to analysis

Ruth Vargas Hill, IFPRI

Outline

- Randomized but not balanced?
- Attrition
- Spillovers
- Partial compliance and selection bias
- Choice and measurement of outcomes
- Protocol adherence
- External validity

Outline

- **Randomized but not balanced?**
- Attrition
- Spillovers
- Partial compliance and selection bias
- Choice and measurement of outcomes
- Protocol adherence
- External validity

Imbalance: risks

- For large N (where N is the unit of randomization) randomization ensures balance on observed and unobserved characteristics, however for smaller N this may not be the case
- Ensuring balance is important. Are respondents balanced on key variables (especially outcome variables) between treatment and control?
- Often they will not be balanced on all observables. For 10 variables, would expect one to be significant at 10 percent level.

Imbalance: solutions

- Re-randomize: keep reallocating to treatment and control until everything (or as much as possible) balances
  - not the best approach
  - not possible if lottery is being drawn in the field
- Stratification and Pairwise matching are preferred:
  - To increase balance on important characteristics (e.g. outcome variables)
  - Balance on unobservables is not improved in this method

Reference: Bruhn and McKenzie. 2009. "In pursuit of balance: randomization in practice in development field experiments" Amer. Econ. J.: Applied Econ.

Stratification

- Also called blocking
- Use baseline characteristics to split the sample up into strata or blocks. Eg:
  - Gender: women and men
  - Gender and above/below median age: old women, young women, old men, young men
- Randomize within each block:
  - Let δ be the proportion of total N to be treated, N_1 of which are women, N_2 of which are men.
  - Select δN_1 women for treatment, select δN_2 men
- Ensures balance for the characteristics which define strata
- Ideally want to choose strata that are highly correlated with the outcome variable

Stratification

- Quite simple, improves balance
- If simple, can be done in the field and without baseline data (e.g. region, gender of participant)
- Cannot be done for many variables at once, consider dichotomous variables:
  - One : 2 strata
  - Three: 2^3 =8 strata
  - Five: 2^5 =32 strata
- Control for this at analysis by including all *n* strata, S, as dummies in the analysis (otherwise over-estimate standard errors, and maybe bias point estimates too), i.e. for ANCOVA

$$y_{ti} = \beta_y y_{t-1,i} + \beta_T T_i + \beta_{S1} S_{1,i} + \ldots \beta_{Sn} S_{n,i} + \varepsilon_{ti}$$

Pairwise matching

- Choose a number of covariates (continuous also)
- Pairs are formed to minimize the Mahalanobis distance between the values of all the selected covariates within pairs
  - Greevy et al (2004): optimal multivariate matching
  - King et al (2007): optimal greedy algorithm. Stata code is online in "matching_algorithm.do" in the folder at: http://www.aeaweb.org/aej/app/data/2008-0182_data.zip
- One unit is assigned to treatment and one to control

Pairwise matching

- Improves balance
- Can be done for many variables at once
- Provides more options for managing attrition (see future slides)
- Takes time (algorithm can take many days to run)—requires baseline data sometime in advance of randomization
- Cannot be done in the field
- Control for this at analysis by including all *n* pair dummies, P, as dummies in the analysis (otherwise over-estimate standard errors), i.e. for ANCOVA

$$y_{ti} = \beta_y y_{t-1,i} + \beta_T T_i + \beta_{P1} P_{1,i} + \ldots \beta_{Pn} P_{n,i} + \varepsilon_{ti}$$

Outline

- Randomized but not balanced?
- **Attrition**
- Spillovers
- Partial compliance and selection bias
- Choice and measurement of outcomes
- Protocol adherence
- External validity

Attrition: risks

- Is it a problem if some of the people in the experiment vanish before you collect your data?
  - It is a problem if the type of people who disappear / appear is correlated with the treatment (and you want to do more than measure the rate of disappearance / appearance)

- Why is it a problem?
- Why should we expect this to happen?

Attrition bias: an example

- The problem you want to address:
  - Lack of price information causes farmers to sell to traders at the farmgate rather than in the nearby market.
- You start a price information program and randomize markets and their surrounding villages to treatment and control. You expect the following effects:
  - Increased quantities sold: farmers sell more when they can bargain for a higher price (Key, Sadoulet and de Janvry)
  - Increased sales at the market: particularly, farmers with smaller amounts of crop to sell start going to the market more when they know the price is higher (Key, Sadoulet and de Janvry, Hill and Fafchamps)
- You go to the markets (treatment and control) and record all farmer sales in a given month
- Will the treatment-control difference be accurately estimated for both outcomes?

Actual difference in quantity sold

	Before Treatment			After Treament	
	T	C		T	C
	20	20		**22**	20
	25	25		**27**	25
	30	30		**32**	30
Ave.					
	Difference			Difference	

Actual difference in quantity sold

	Before Treatment			After Treament	
	T	C		T	C
	20	20		**22**	20
	25	25		**27**	25
	30	30		**32**	30
Ave.	**25**	**25**		**27**	**25**
	Difference	0		Difference	2

Difference measured at the market

	Before Treatment			After Treament	
	T	**C**		**T**	**C**
	[absent]	[absent]		**22**	[absent]
	25	25		**27**	25
	30	30		**32**	30
Ave.					
	Difference			Difference	

Difference measured at the market

	Before Treatment			After Treament	
	T	**C**		**T**	**C**
	[absent]	[absent]		**22**	[absent]
	25	25		**27**	25
	30	30		**32**	30
Ave.	**27.5**	**27.5**		**27**	**27.5**
	Difference	0		Difference	-0.5

Attrition: solutions

- Don't rely on measures taken at a place where you think you will see attrition, measure where you think attrition will be the lowest:
  - E.g. survey farmers at home rather than at the market when you think the intervention will have an impact on where farmers sell.
- Devote resources to tracking participants
- If there is still attrition, check that it is not different in treatment and control. Is that enough?
- Also check that it is not correlated with observables.
- Try to bound the extent of the bias
  - suppose everyone who dropped out from the treatment hadthe lowest outcome that anyone got; suppose everyone who dropped out of control got the highest outcome that anyone got…
  - Why does this help?

Attrition and pairwise matching

	ATE for full sample	ATE on those who did not drop out	ITT
Attrition is random and treatment effect is constant across population	Drop a pair from analysis if one of the two partners drops out—increases power	Drop a pair from analysis if one of the two partners drops out	Use full sample
Attrition is random and treatment effect is heterogeneous	Use full sample	Drop a pair from analysis if one of the two partners drops out	Use full sample
Attrition is not-random	Use full sample	Drop a pair from analysis if one of the two partners drops out. But assumes attrition on observables	Use full sample

Outline

- Randomized but not balanced?
- Attrition
- **Spillovers**
- Partial compliance and selection bias
- Choice and measurement of outcomes
- Protocol adherence
- External validity

Spillovers: risks

- As discussed, for some interventions spillovers are likely.

- If this is not taken into account in the randomization design, the measured program impact will underestimate the direct program impact (positive externalities) overestimated (negative externalities)

Spillovers: solutions

- Randomize at a level which incorporates the spillover: at the village, at the group, or use distance between members.

- If interested in estimating direct program effect plus spillover: vary the intensity of treatment within the cluster
  - Control villages: no intervention
  - Treatment villages A: intervention to 10 people in village, randomly
  - Treatment villages B: intervention to 50 people in village, randomly
  - Treatment villages C: intervention to all in village, randomly

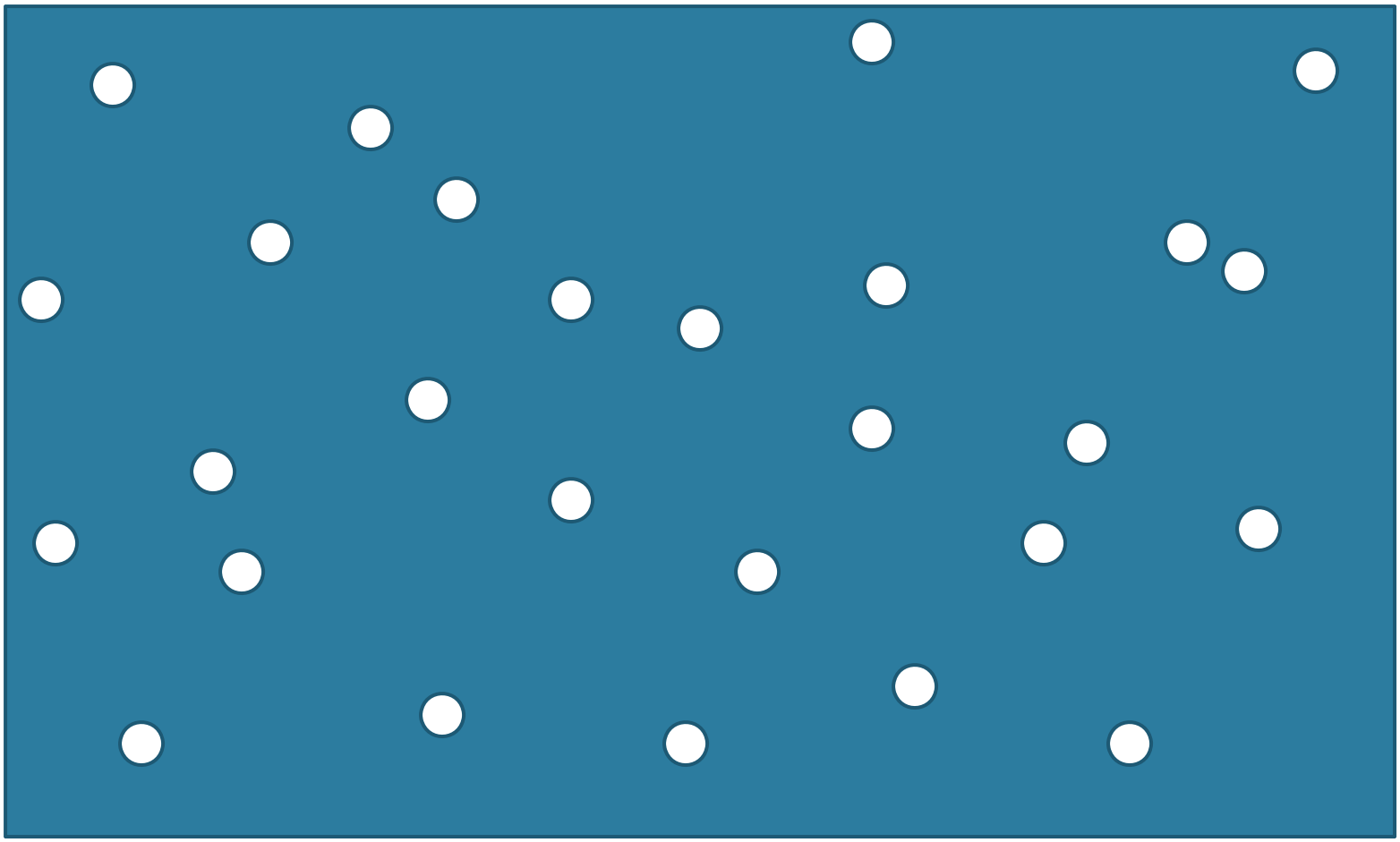
Spillovers: solutions

- May not have enough villages to randomize at the village level, but are concerned about spillovers within the village
- Conduct an initial analysis that tells you about the nature of the spillover:
  - Along what dimensions does the spillover occur?—physical distance, close relatives
  - Is there a "distance" beyond which spillover is unlikely?—individuals more than 2km apart, people outside your clan
- Analysis provides a "distance" that can be used to generate groups where spillovers are unlikely.
- Randomize on these groups

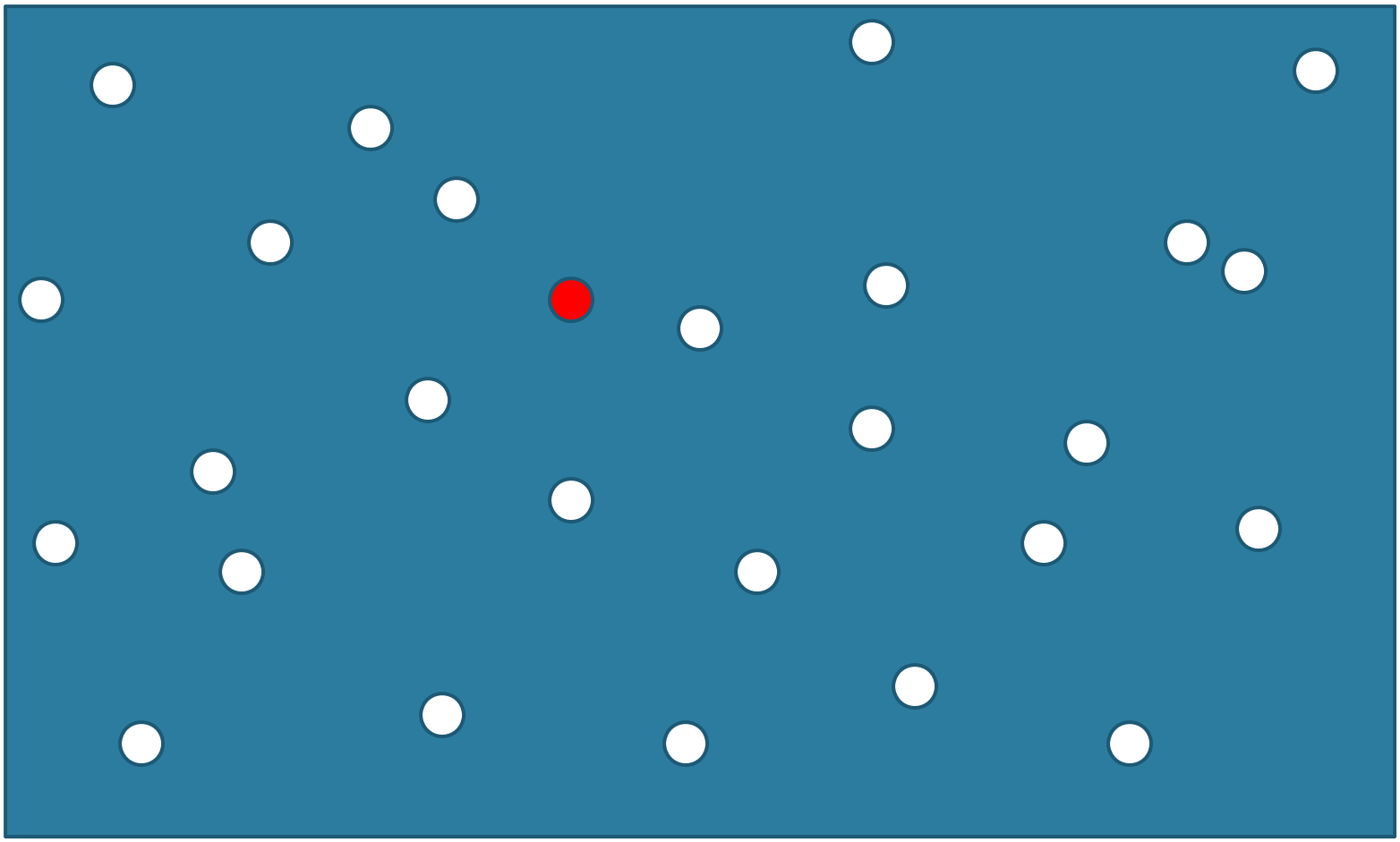
Distance: example 1

- Background:
  - Intervention to compare providing insurance to groups and individually
- Risk:
  - Groups do not stay within village boundaries, cannot randomize at the village
  - Randomizing at the next level up (kebele) implies working on a scale that is too large for the MFI
- Solution:
  - Conduct a baseline network map on membership of groups: which villages do farmers travel to be part of the group; questions on what determines choice of group; GPS coordinates of all villages
  - Probability of membership in a village more than 2km away is very unlikely, given regular attendance at meetings and funerals is required
  - Select villages for randomization that are more than 2km apart.

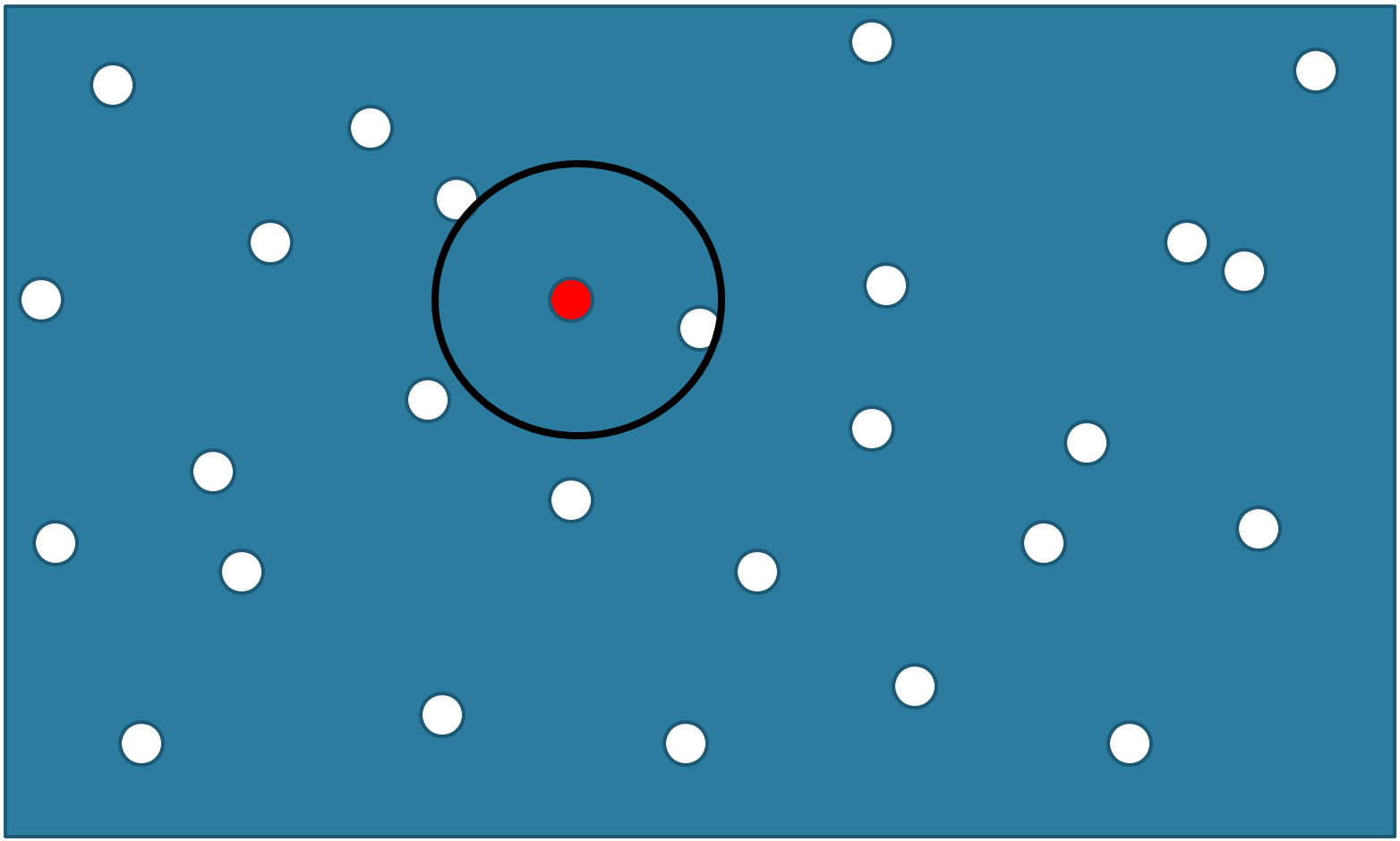
All villages



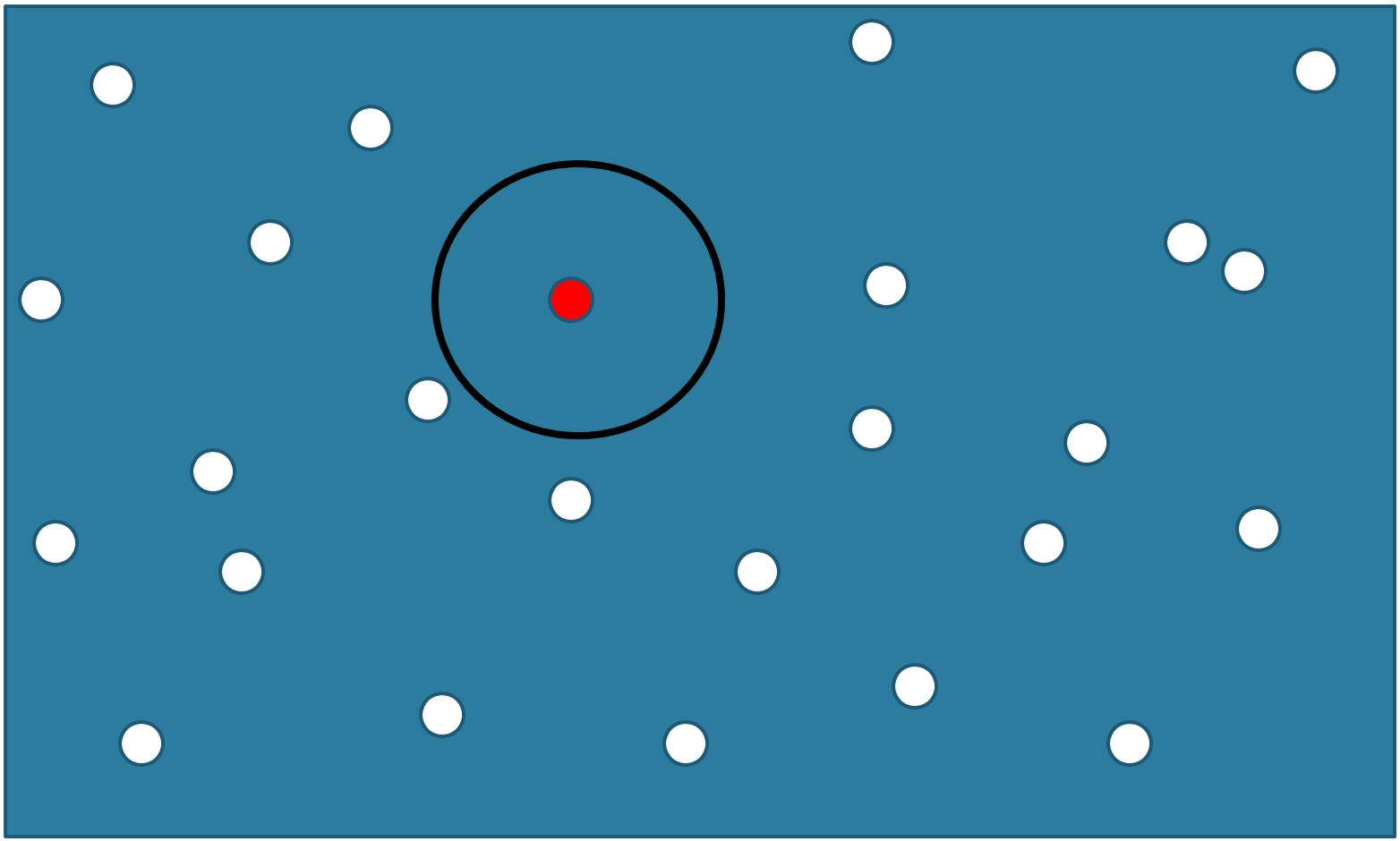
Select first village



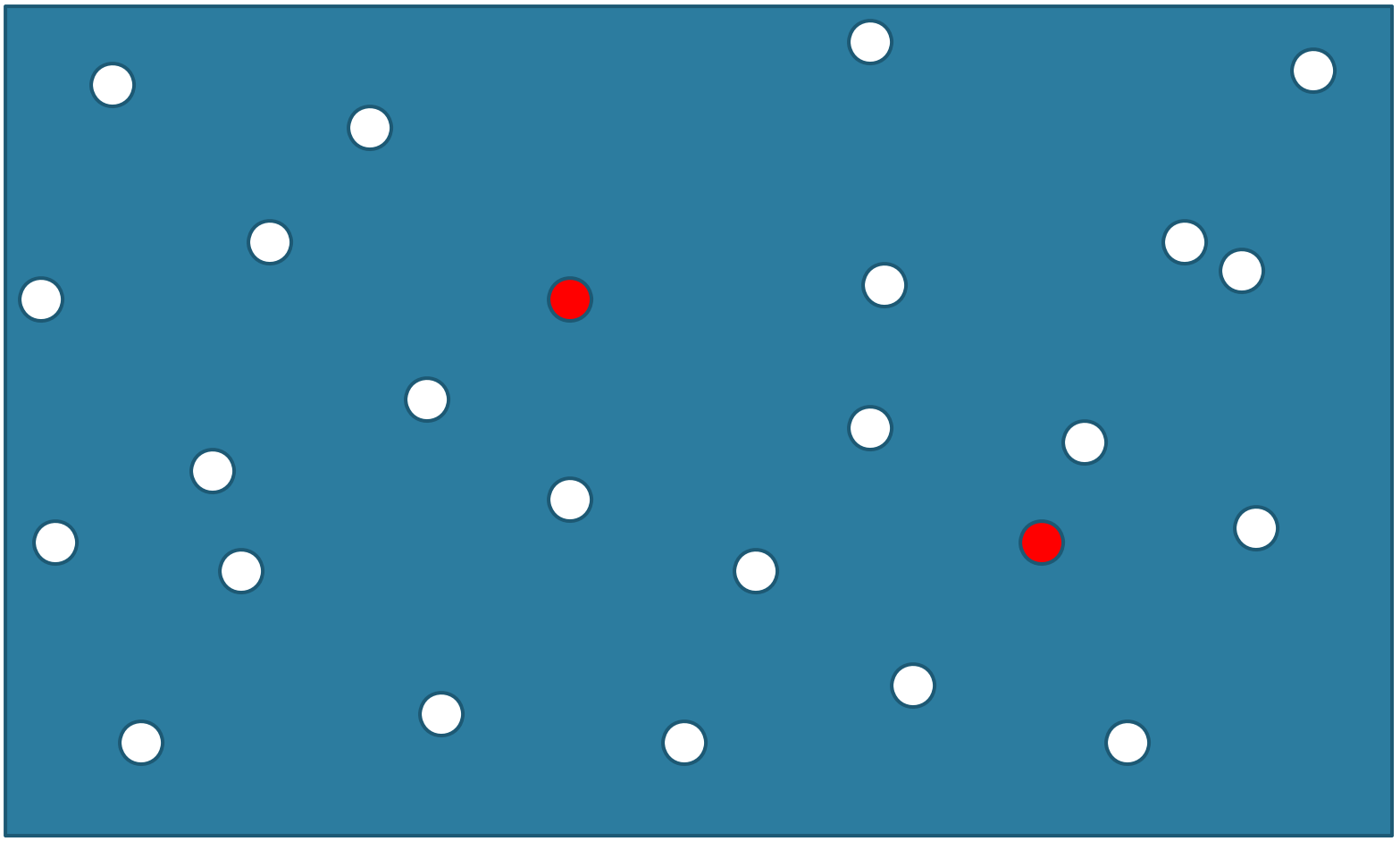
Drop villages within a 2km radius



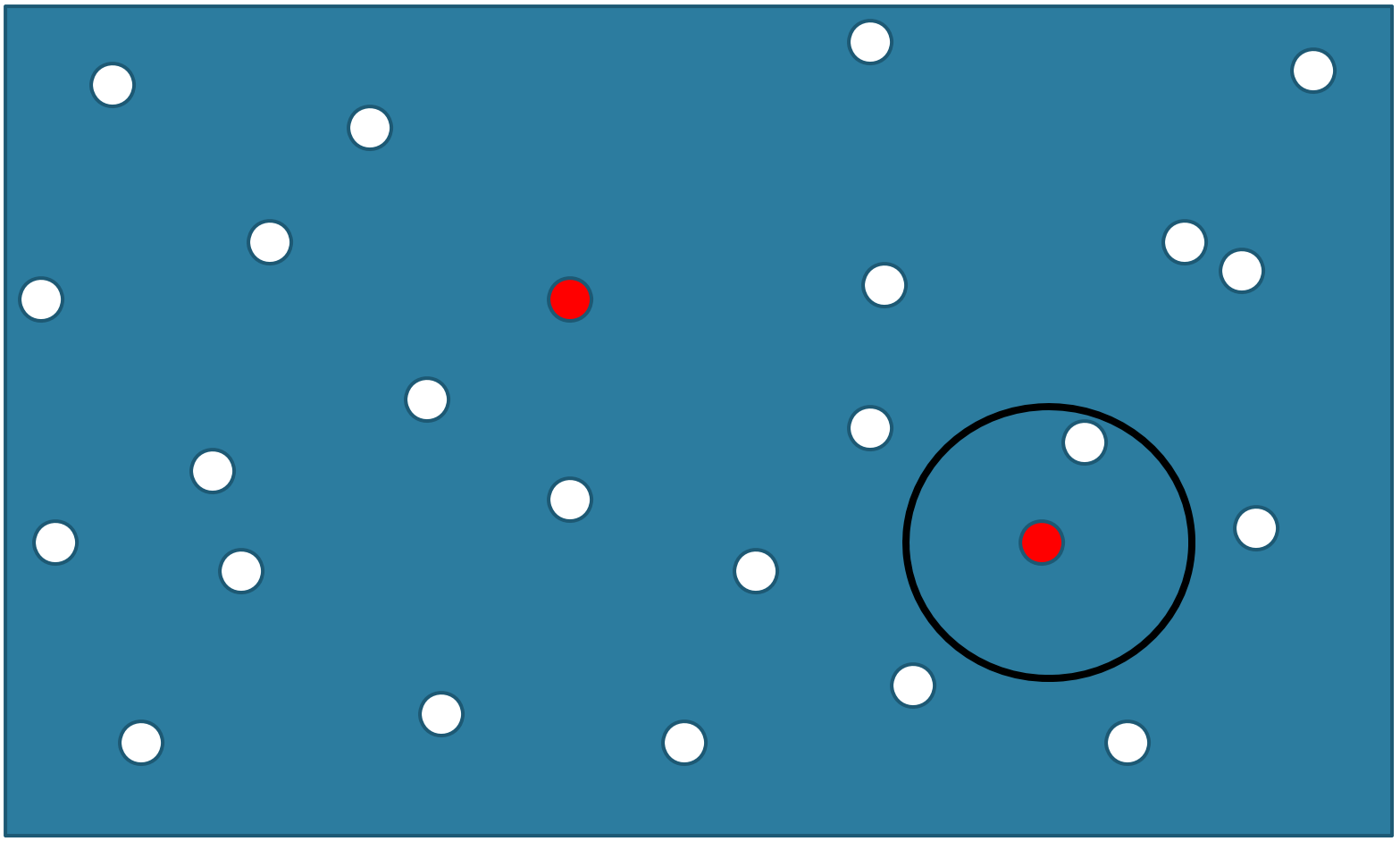
Drop villages within a 2km radius



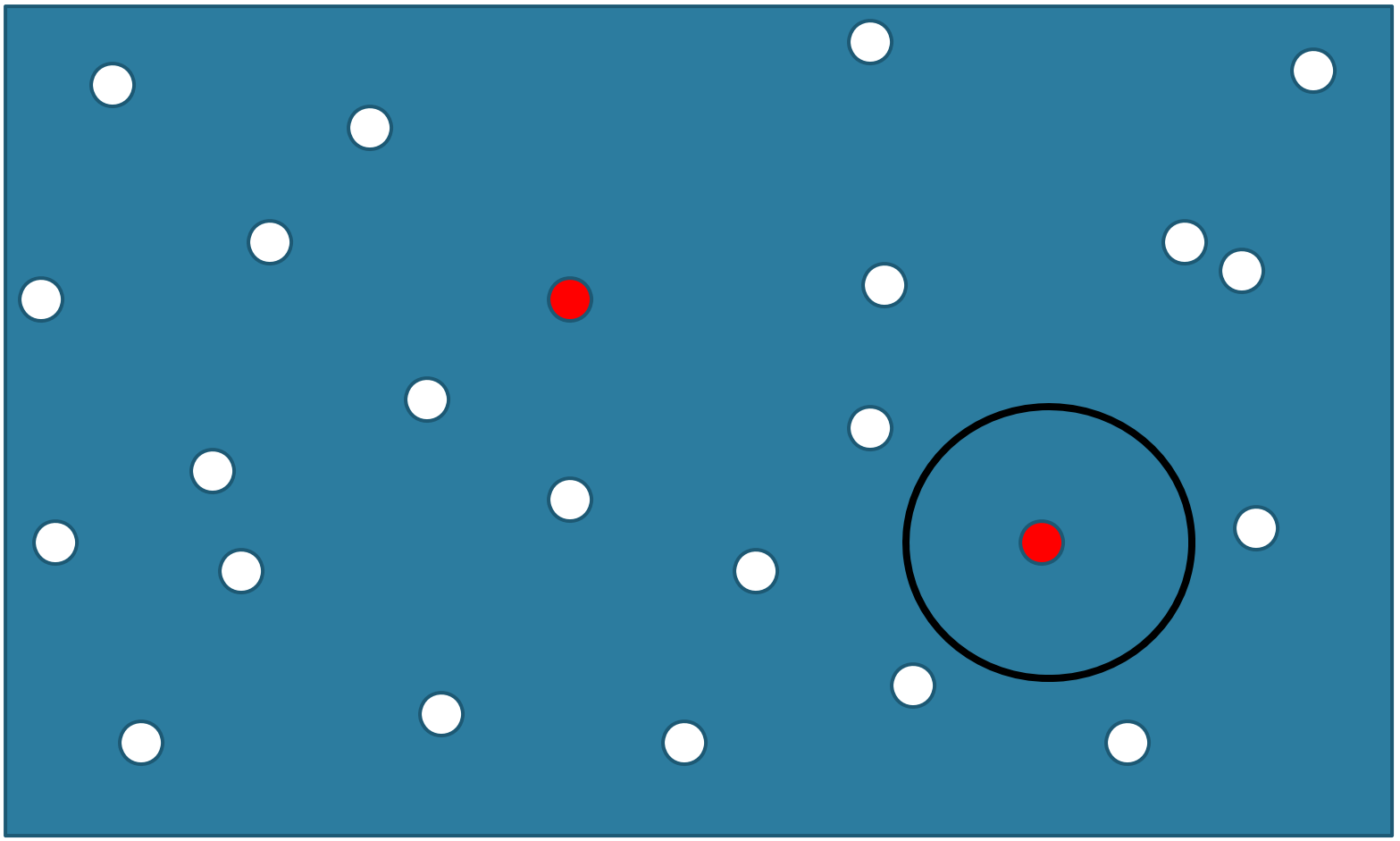
Select second village



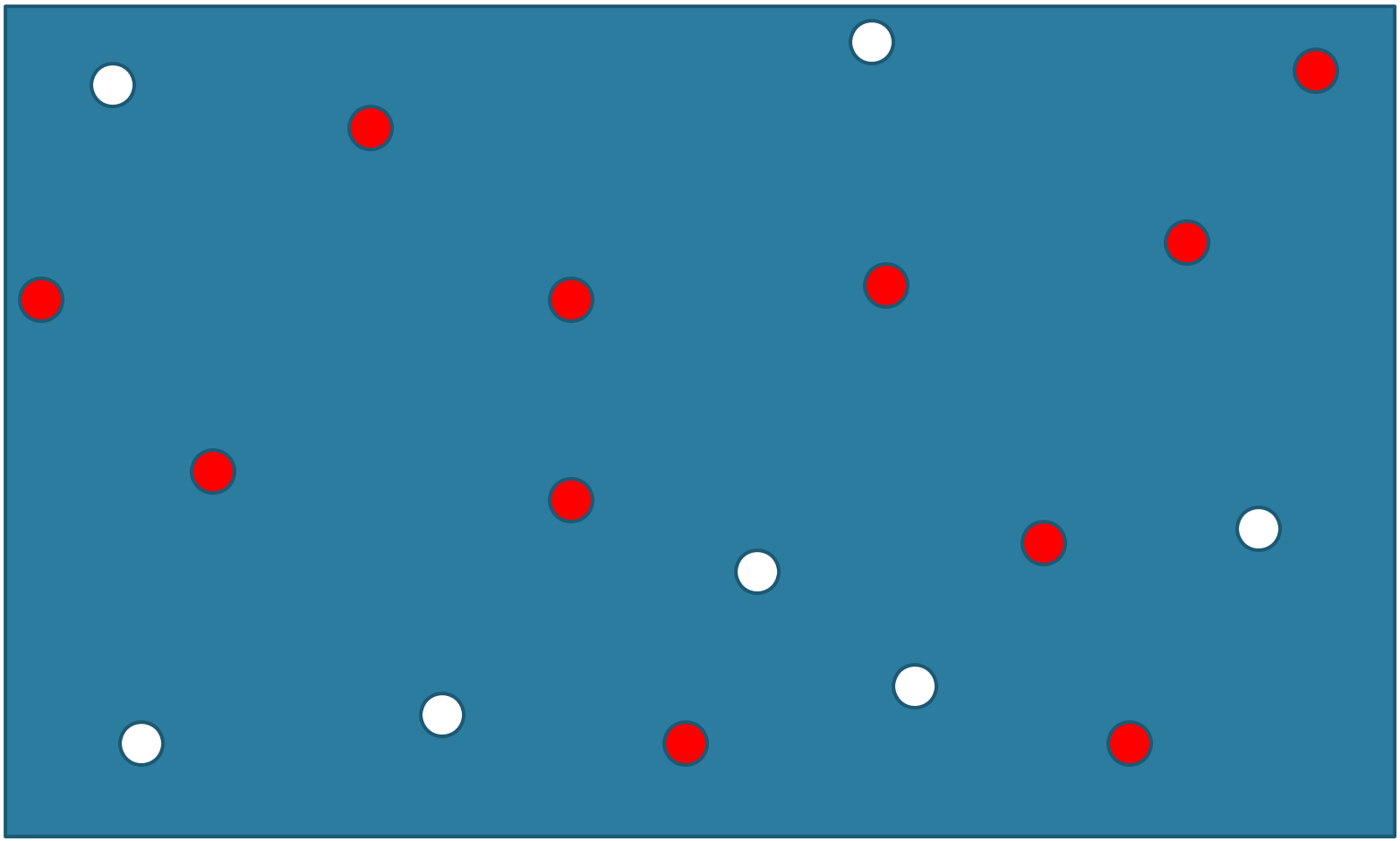
Drop villages within a 2km radius



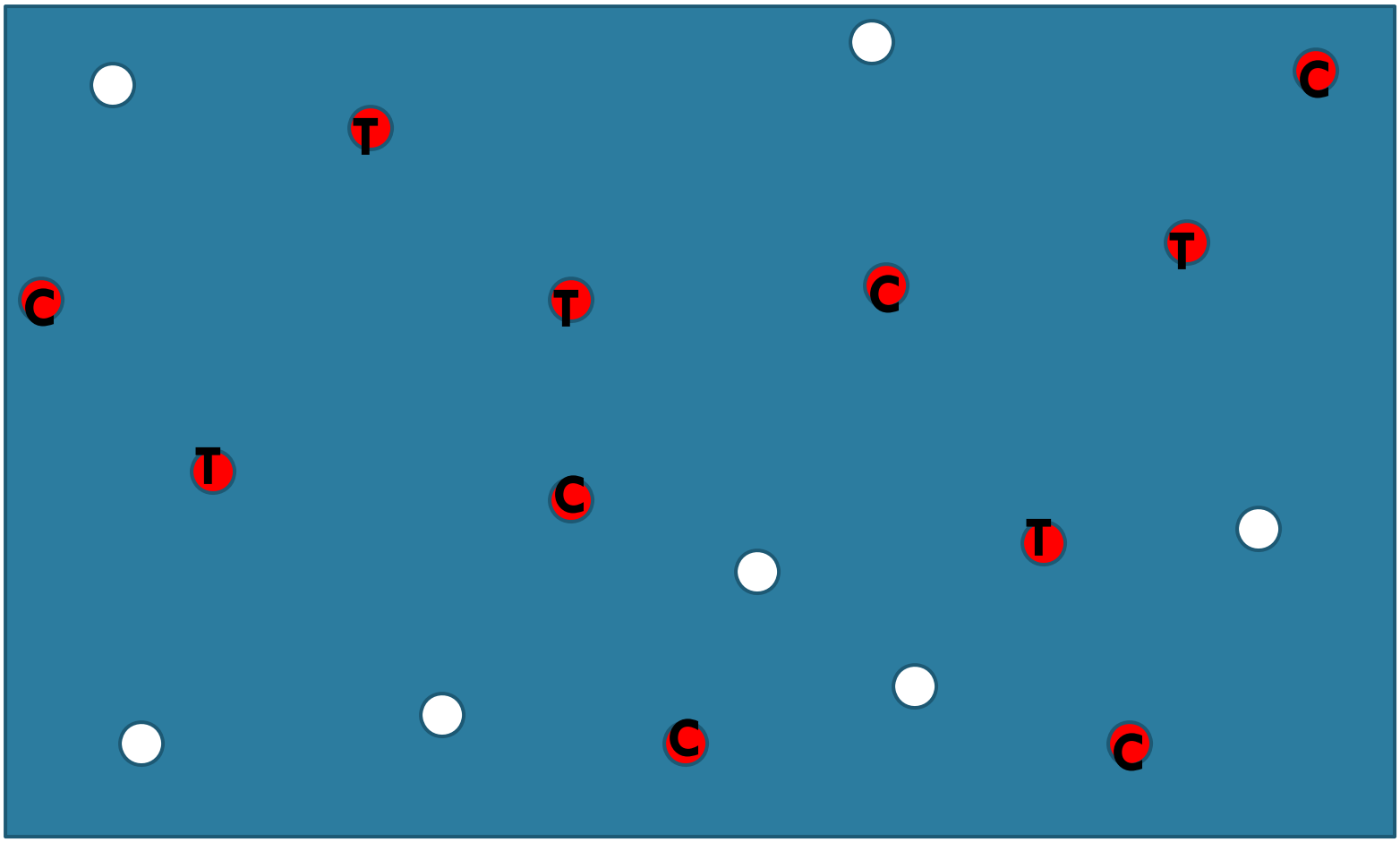
Drop villages within a 2km radius



Repeat until enough study villages are selected



Randomize treatment across study villages



Distance: example 2

- Background:
  - Intervention in Tanzania to look at impact of incentive schemes between a milk company and farmers.
  - Randomly selected farmers given a new incentive scheme that rewards loyalty to test the scheme.
- Risk:
  - Spillovers are possible: those with an incentive can deliver milk on behalf of others.
  - The firm is only operational in 10-12 villages, randomization at the village level is not possible.
- Solution:
  - Analysis of baseline data suggests that delivering on behalf of others occurs within a clan, but rarely outside of a clan.
  - Randomize at the clan level.
  - Collect data on GPS coordinates of clans.

Outline

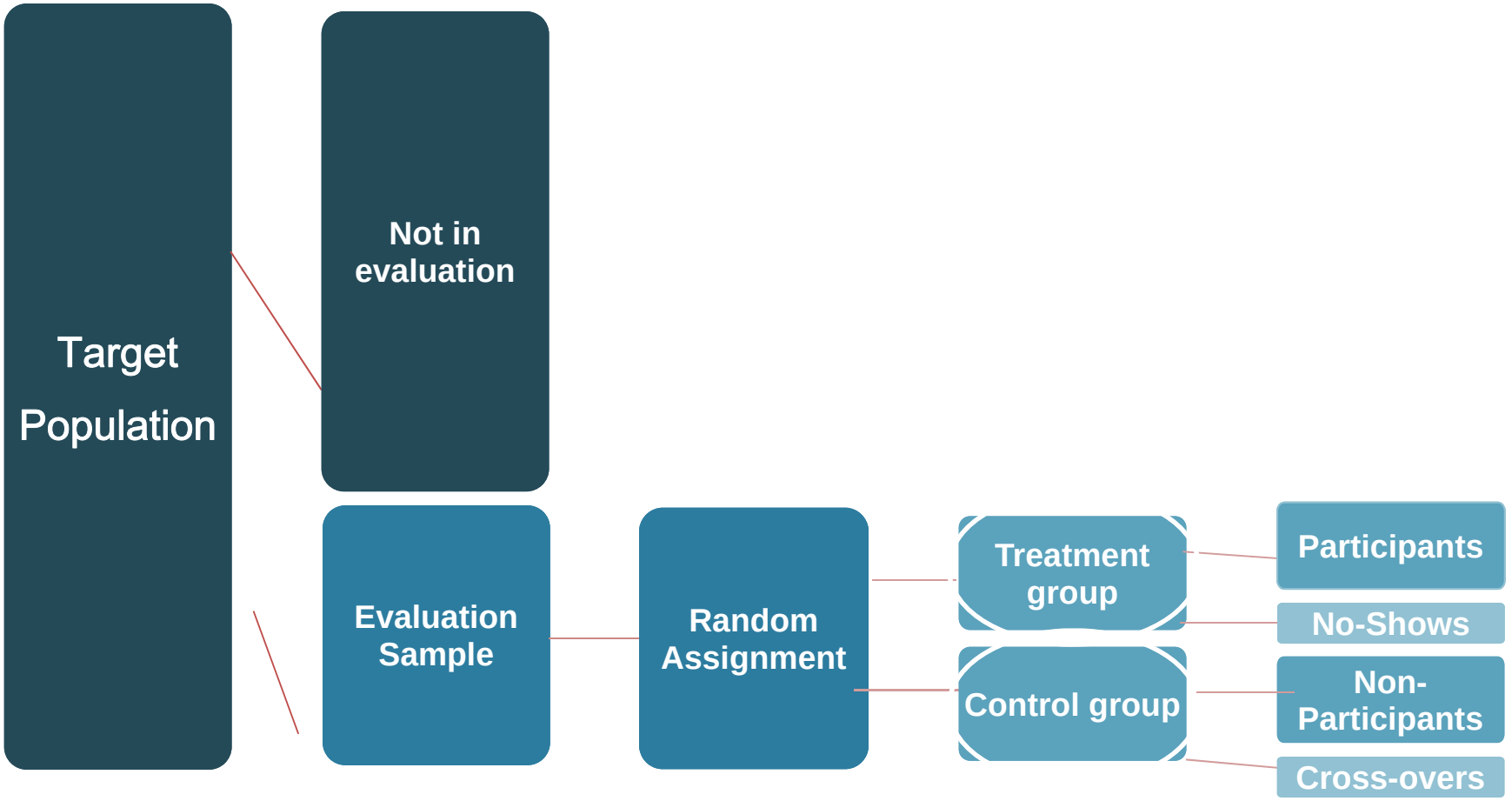
- Randomized but not balanced?
- Attrition
- Spillovers
- **Partial compliance and selection bias**
- Choice and measurement of outcomes
- Protocol adherence
- External validity

Selection bias: risks

- Sample selection bias could arise if factors other than random assignment influence program allocation
  - Even if intended allocation of program was random, the actual allocation may not be

- Individuals assigned to comparison group could attempt to move into treatment group
- Alternatively, individuals allocated to treatment group may not receive treatment (partial compliance)
- Can you just compare beneficiaries and non-beneficiaries anyway? Why not?

Selection bias: risk



Selection bias: solutions

- Intent to treat (ITT):
  - Average impact of program in practice: treats all noncompliars as treated, and treats all crossovers as remaining in the control
  - Problem: power is reduced by noncompliance and does not provide an idea of what the average impact of the program on the treated is.
- Treatment on the treated (ToT):
  - Instruments for take-up with assignment: gives an idea of the average impact of the program for a specific group
- Encouragement design:
  - To encourage compliance

Intent to treat

- Instead of randomizing treatment (T), we randomized assignment/eligibility (Z)
  - e.g. we randomized access to improved seeds rather than whether or not they used improved seeds.
- We can estimate the impact of assignment to treatment, this is called the intent to treat.
- Z replaces T in the previous regression models.
- For ANCOVA:

$$y_{ti} = \beta_y y_{t-1,i} + \beta_z Z_i + \varepsilon_{ti}$$

- This may often be what we want to estimate, as the effect of a policy is often the impact of Z.

Treatment on the treated

- Sometimes we want to know the impact of the treatment, e.g. what actually is the impact of improved seeds on farmer yields and crop revenue.
- Use Z as an instrument for T to estimate the LATE (local average treatment effect) – the average effect of the treatment for those who were induced into treatment by being assigned in the program
- Estimate

$$y_{ti} = \beta_y y_{t-1,i} + \beta_T T_i^* + \varepsilon_{ti}$$

Where T^*_i is the fitted value from the regression:

$$T_i = \alpha_y y_{t-1,i} + \alpha_Z Z_i + \mu_{ti}$$

Treatment on the treated: assumptions

- Z does not have a direct effect on y other than through T
  - Not always the case: spillovers? Can also be a program effect even if full treatment is not selected into (information about improved farming practices even if did not use improved seeds)

- Z has a monotonic effect on T: the probability of being treated increases for everyone when Z=1 or the probability of being treated decreases for everyone when Z=1
  - Usually the case, but worth thinking through
  - It is always the case when no-one in the control group was treated. In this case the LATE gives the average effect of the treatment on the treated.

Outline

- Randomized but not balanced?
- Attrition
- Spillovers
- Partial compliance and selection bias
- **Choice and measurement of outcomes**
- Protocol adherence
- External validity

Measurement of outcomes: risks

- Choosing too many outcome variables will inevitably result in one of them being positive.
- Variables of ultimate interest (e.g. consumption per capita) have many determinants, so it is unlikely that the intervention will have a large detectable effect.
- Respondents may be tempted to report changes in stated outcomes (did you change your behavior as a result of….) that do not reflect change in underlying behavior
- Highly variable outcomes or outcomes measured with a lot of noise, have a very large MDE for a given randomization design.

Measurement of outcomes: solutions

- Too many outcome variables:
  - Pre-specify outcomes of interest
  - Report results on all measured outcomes, even null results
  - Correct statistical tests (but don't overcorrect a la Bonferroni)
- Variables of ultimate interest have many determinants:
  - Look at intermediate outcomes, have a model of change
- Stated changes:
  - Back these up with measurement of the underlying change in behavior

Outcomes: solutions

- Highly variable outcomes or outcomes measured with a lot of noise:
  - Take repeated measures. Baseline data is key. More than one follow-up can help.
  - Improve accuracy of measurement with shorter recall or other means of collection (diaries, regular visits, records at marketing place of extension agent)
  - Must do the same for both treatment and control.

Improving measurement and repeated measures

Improving measurement:

- Careful supervision of surveys, use PDAs where possible, multiple questions key outcome variables
- Visiting a household at the right time reduces recall error, conduct surveys after the main agricultural events to be assessed – planting, fertilizer application, harvest, sales of harvest
- Visiting a household more often reduces recall error: number of loans taken in a year, number of gifts given or received.
- Rely on more than just survey responses: field visits, extension officer reports, MFI loan data, sales data, data collected by traders or in markets

Repeated measures:

- If the outcome of interest is highly variable with little autocorrelation across time (e.g. trader sales) then repeated surveys increases power

McKenzie. 2011. Beyond baseline and follow-up. The case for more T in experiments. World Bank Policy Research Working Paper 5639

Repeated measures

DIF-DIF for multiple rounds

$$Y_{ti} = \beta\, EVERT_i + \gamma\, T_{ti} + \delta_{m-1} + \ldots + \delta_r + \varepsilon_{ti}$$

Y_{ti}: outcome of interest for individual i at time t

$EVERT_i$: i is in the treatment group

T_{ti}: i is treated at time, t

δ_t: time dummies for each survey round: m pre-treatment survey rounds labeled m-1 to 0 and r post-treatment survey rounds labeled from 1 to r.

ε has a mean of 0 and a cross-section variance of σ^2

ρ is the autocorrelation of ε across time

Repeated measures

ANCOVA for multiple rounds

$$Y_{ti} = \theta Y_{PRE,i} + \gamma T_{ti} + \delta_1 + \ldots + \delta_r + \varepsilon_{ti}$$

Y_{ti}: outcome of interest for individual i at time t

$Y_{PRE,i}$: mean of Y for individual i over m pre-treatment rounds

T_{ti}: i is treated at time, t

δ_t: time dummies for each of r post-treatment survey rounds labeled from 1 to r.

ε has a mean of 0 and a cross-section variance of σ^2

ρ is the autocorrelation of ε across time

Repeated measures

Estimation method	m	r	Estimator	Variance of estimator
Dif-dif	1	1	$(Y(T)_1 - Y(C)_1) - (Y(T)_0 - Y(C)_0)$	$4\sigma^2(1-\rho)/n$
Ancova	1	1	$(Y(T)_1 - Y(C)_1) - \theta^*(Y(T)_0 - Y(C)_0)$	$4\sigma^2(1-\rho^2)/n$
Dif-dif	m	r	$(Y(T)_{POST} - Y(C)_{POST}) - (Y(T)_{PRE} - Y(C)_{PRE})$	$2\sigma^2/n\ [(1 + (r-1)\rho)/r - ((m+1)\rho - 1)/m]$
Ancova	m	r	$(Y(T)_{POST} - Y(C)_{POST}) - \theta^*(Y(T)_{PRE} - Y(C)_{PRE})$	$2\sigma^2/n\ [(1 + (r-1)\rho)/r - m\rho^2 / (1 + (m-1)\rho)]$

n is the number of people in the treatment group and the control group

Implications

- ANCOVA provides more power than dif-dif
- How to split a survey budget between pre and post-treatment rounds?
  - the lower the autocorrelation, the more post-treatment survey rounds should be conducted
  - If ρ=0.25 and there is only a budget for 3 rounds, it is best to have three follow-up waves and no baseline.
- What is the gain from an additional follow-up round?
  - Going from r to r+1 rounds increases power by $(1-\rho)/r(r+1)$
  - Greatest gain going from r=1 to r=2
  - Gains are smaller the higher the autocorrelation
- How to choose n and T with a fixed budget for nT surveys?
  - high n when ρ is high
  - high T when ρ is low

Repeated measures: example

Tanzanian milk firm, 131 households in 77 clans

- Surveys (wide variety of variables)
- Daily delivery data collected at firm (key variable of interest)

Measure	Source	No. of obs per individual	Treatment effect
Prob of daily delivery	Delivery data	About 150	0.060*** (0.011)
Number of monthly deliveries	Delivery data	5	2.275** (1.026)
Number of monthly deliveries in season	Delivery data	2	2.274 (1.532)
	Survey data	2	4.125 (2.698)

Outline

- Randomized but not balanced?
- Attrition
- Spillovers
- Partial compliance and selection bias
- Choice and measurement of outcomes
- **Protocol adherence**
- External validity

Protocol adherence: risks

- Perfect designs can be ruined with bad implementation
  - Low levels of compliance jeopardize power calculations
  - Endogenous switching of treatment groups
  - Differences in carefully designed interventions can change the hypothesis that can be tested

Protocol adherence: solutions

- Implementing agency sees the benefit of randomization
- Oversight of implementation as it progresses
- Ensuring randomization is manageable:
  - As Karen discussed, this is one reason for randomizing at the cluster, not the individual level
  - Simple designs
- Ensuring randomization is perceived as fair by the implementing agency and the recipients (farmers):
  - Transparent lotteries, gradual rollout of program, alternative intervention for control (if appropriate)
- Implementation at a manageable scale for implementing agency and oversight

Oversight

- It is important for the evaluator to be involved with the implementation of the project in order to ensure that:
  - Random assignment was adhered to
  - Spillovers and non-compliance weren't major issues
  - The design of the program wasn't changing in ways that could undermine the validity of the experiment
  - Particularly at beginning, but also throughout

Implementing at a manageable scale

- Power tests suggest more N
- Ability to implement protocol correctly is important. How many villages and households can field staff reach? How well can supervisors oversee their implementation.
- Operating at too large a scale for proper implementation can result in imperfect compliance and cross-overs which weaken power considerably.
- Thinking of other ways to increase power is important:
  - Encouragement design (to increase compliance)
  - Baseline measures of outcome variables (ANCOVA estimation)
  - Repeated measures of outcome variables
  - Investing in more accurate collection of data on outcome variables

Outline

- Randomized but not balanced?
- Attrition
- Spillovers
- Partial compliance and selection bias
- Choice and measurement of outcomes
- Protocol adherence
- **External validity**

External validity: risks

- Often working in a few selected sites is easiest, but hard to extrapolate findings in this case—will the measured impact be the same somewhere else
- Being observed carries its own effect, which may affect estimates of program impact
- We often don't have good placebo treatments
- Encouragement designs may bring in people that are unlikely to participate in reality

External validity: solutions

- In the short run: work in varied sites
- In the long run: conduct repeated experiments in different contexts and do a Cochrane review
- Be careful about how repeated measurements are taken
- Think of good placebo treatments where possible: a different seed? Different type of information? Cash equivalent?
- Randomize encouragement designs so we have some idea of the impact of the encouragement on participation:
  - Randomize the value of the subsidy
  - Randomize the intensity of training / house visits