Causal Inference - Mini Course

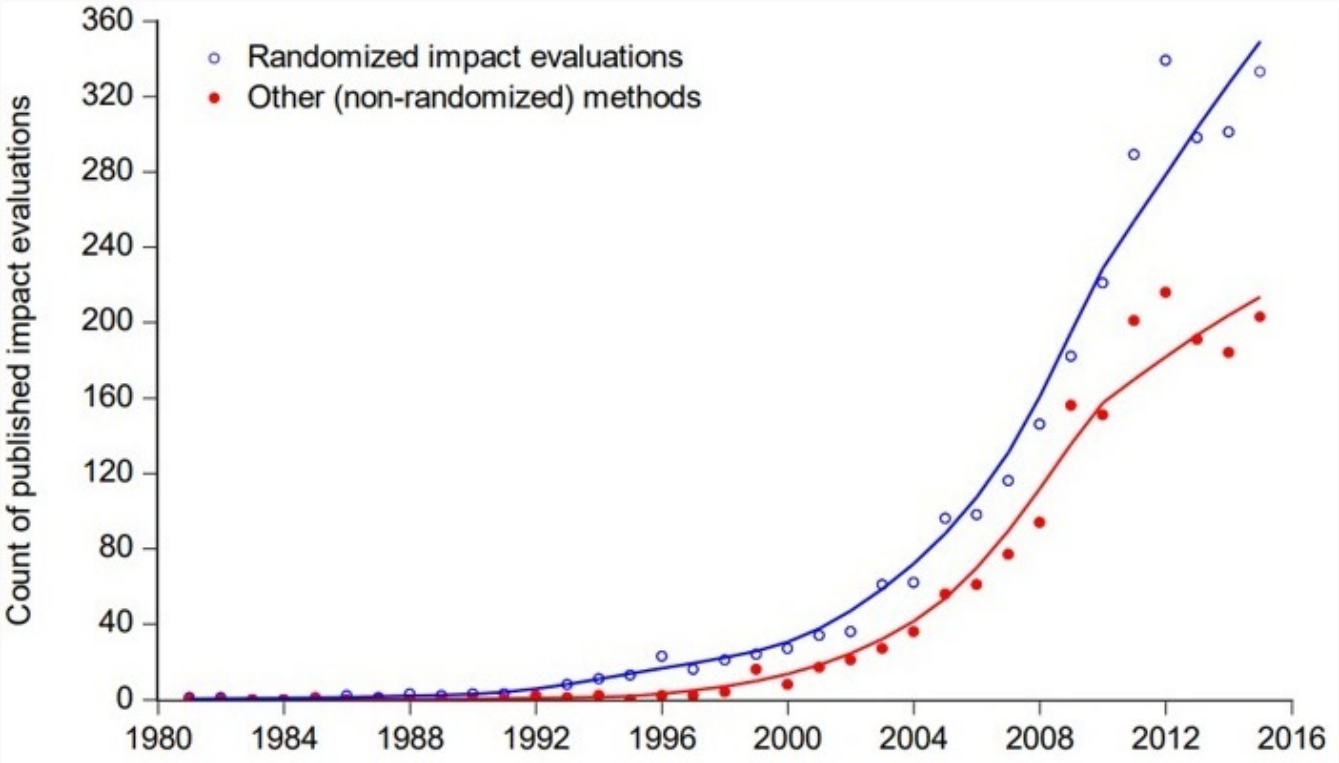
session 1 — RCTs: basics and an extended example

Simon Heß

August 23

RCT basics

RCTs' rise in development economics in recent years



(Source: Ravallion et al. 2018, based on IIIE)

- RCTs widely used in econ
- Nobel prize in Economics 2019 awarded to Abhijit Banerjee, Esther Duflo, and Michael Kremer "for their experimental approach to alleviating global poverty"

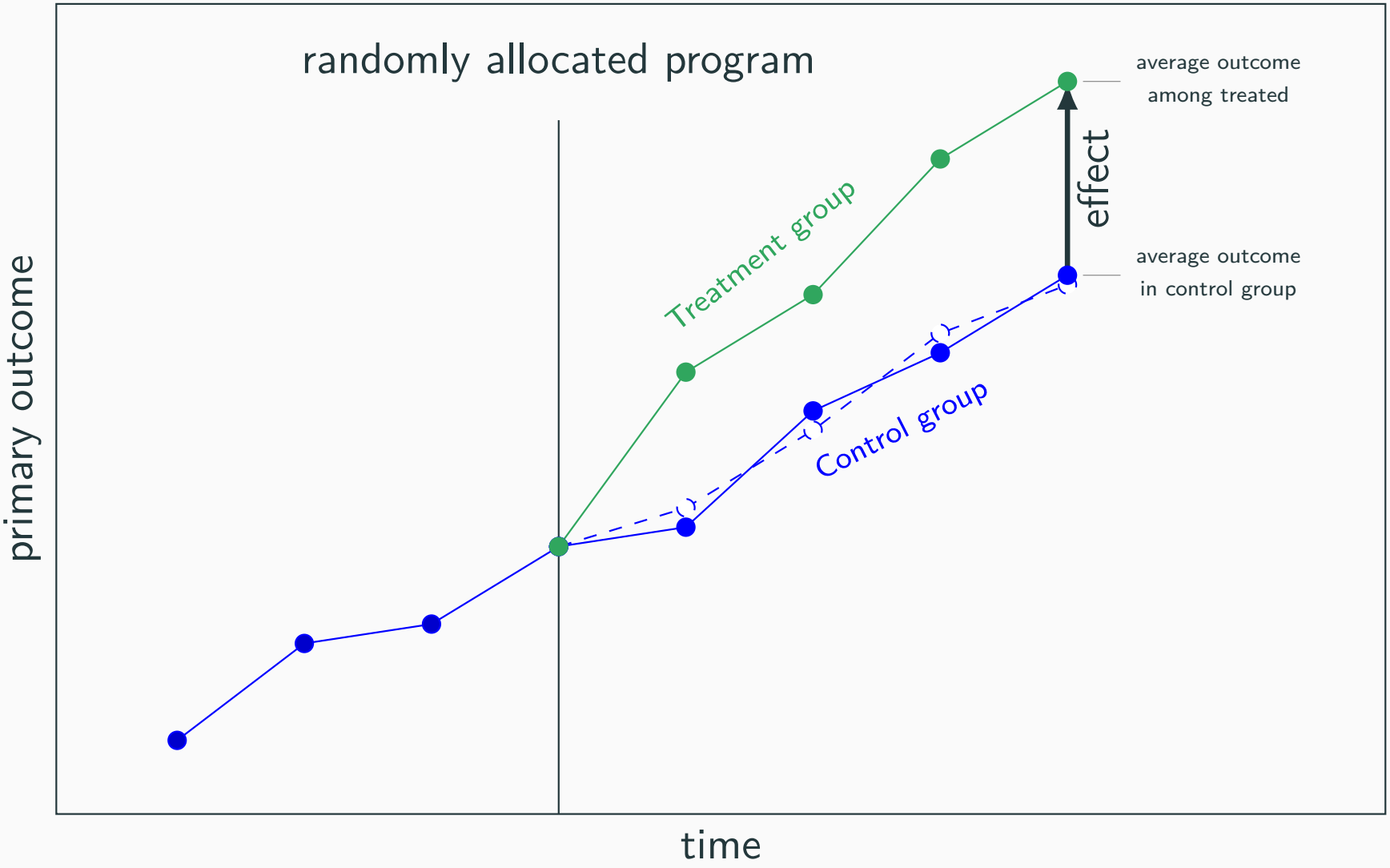
RCT topics for today

- 4 steps to conduct an RCT
- example: microfinance

Optional:

- balance
- attrition
- multiple hypotheses testing (and p-hacking)
- power calculations / minimal detectable effect

RCTs identify the counterfactual outcome



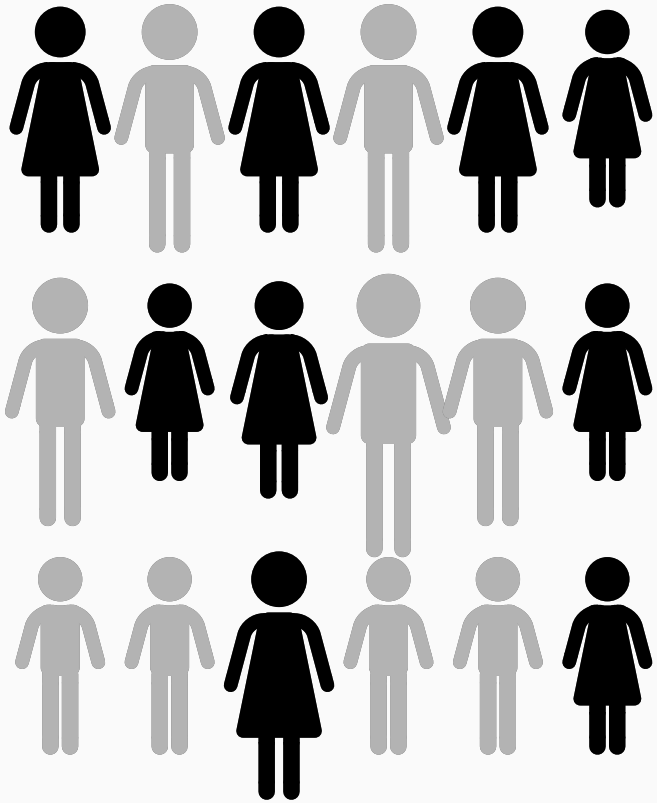
4 steps to an RCT

Steps to an RCT: 1 – Eligibility

1. Define who is eligible and draw sample
    - e.g. agriculture program for women.

Requires:

- a sampling framework
    - lists of households, firms, villages, ...

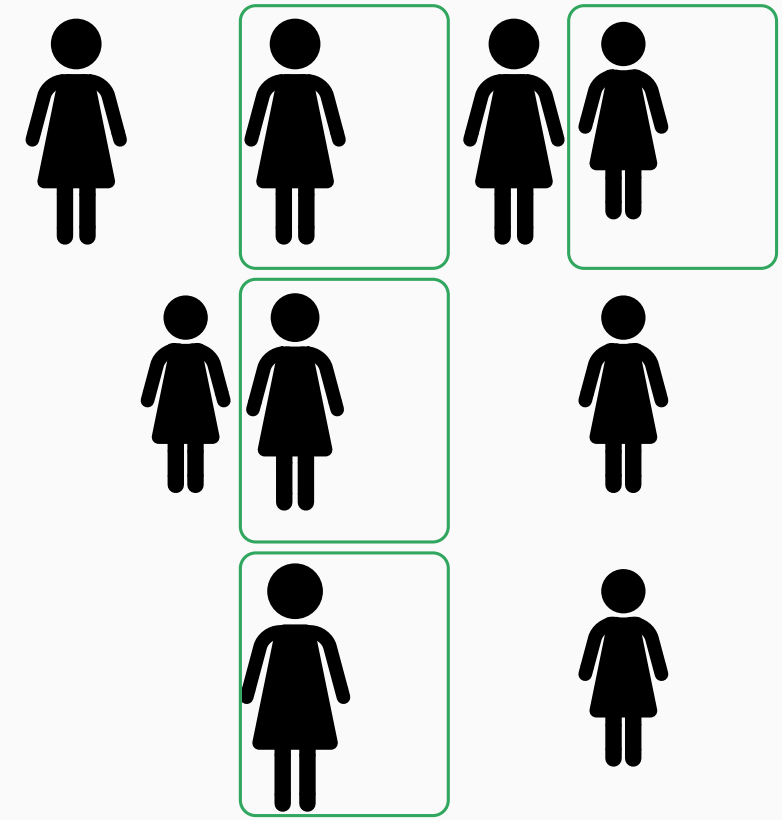


Steps to an RCT: 2 – Randomization

2. Randomize access to program:
   - among eligible participants
   - randomization ensures treatment and control are similar.
   - all characteristics are balanced in expectation.

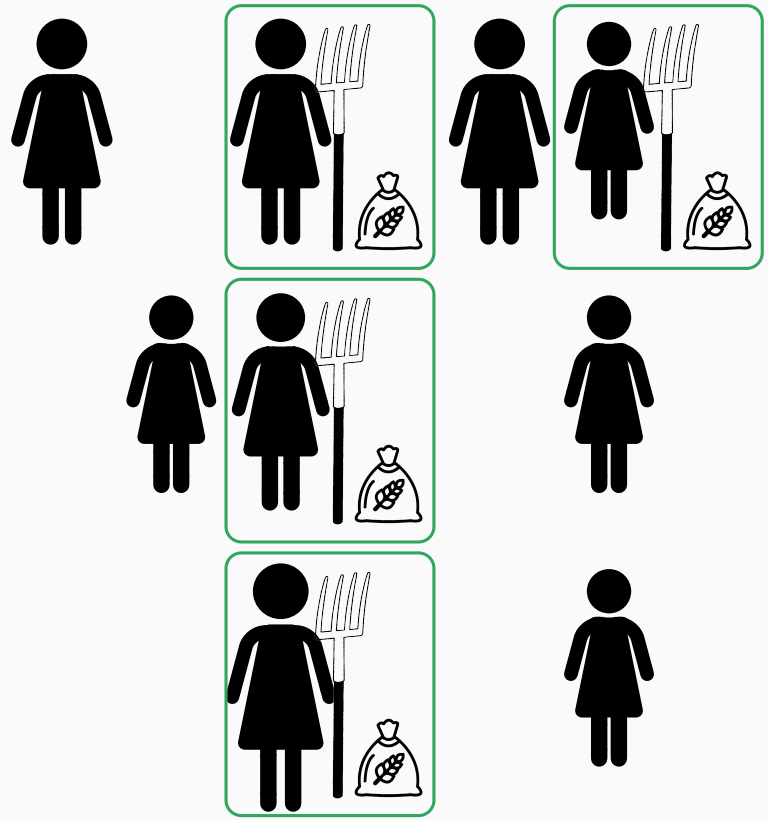
Requires:

- a randomization device
  - coin toss (bad)
  - code (good)
    - R: `treatment <- rbinom(n,1,0.5)`
    - Stata: `gen treatment = runiform()>0.5`



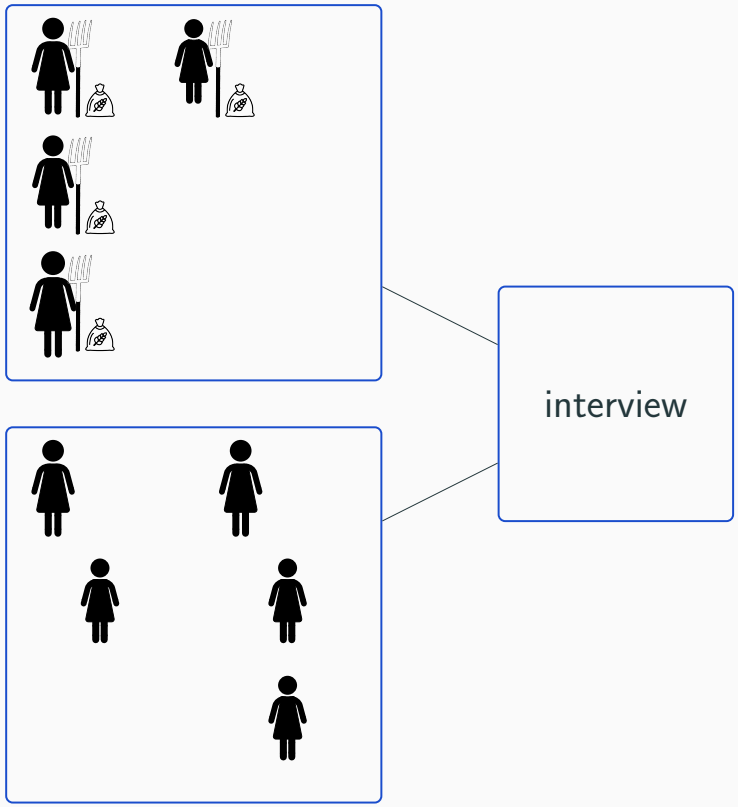
Steps to an RCT: 3 – Implementation

3. Implement intervention
    - only for treatment individuals



Steps to an RCT: 4 – Data collection

4. After the intervention, obtain data from (a sample of) people in the treatment and the control group.
   - Data is typically collected through surveys
   - Alterntives: official records (admin data), satellite images, ...
   - Baseline data is not required, but useful. E.g., to study balance of pre-treatment characteristics
     - "Balance tests" to verify if "randomization worked" (more later).



Treatment assignment in RCTs

in the previous example, treatment is assigned at the individual level

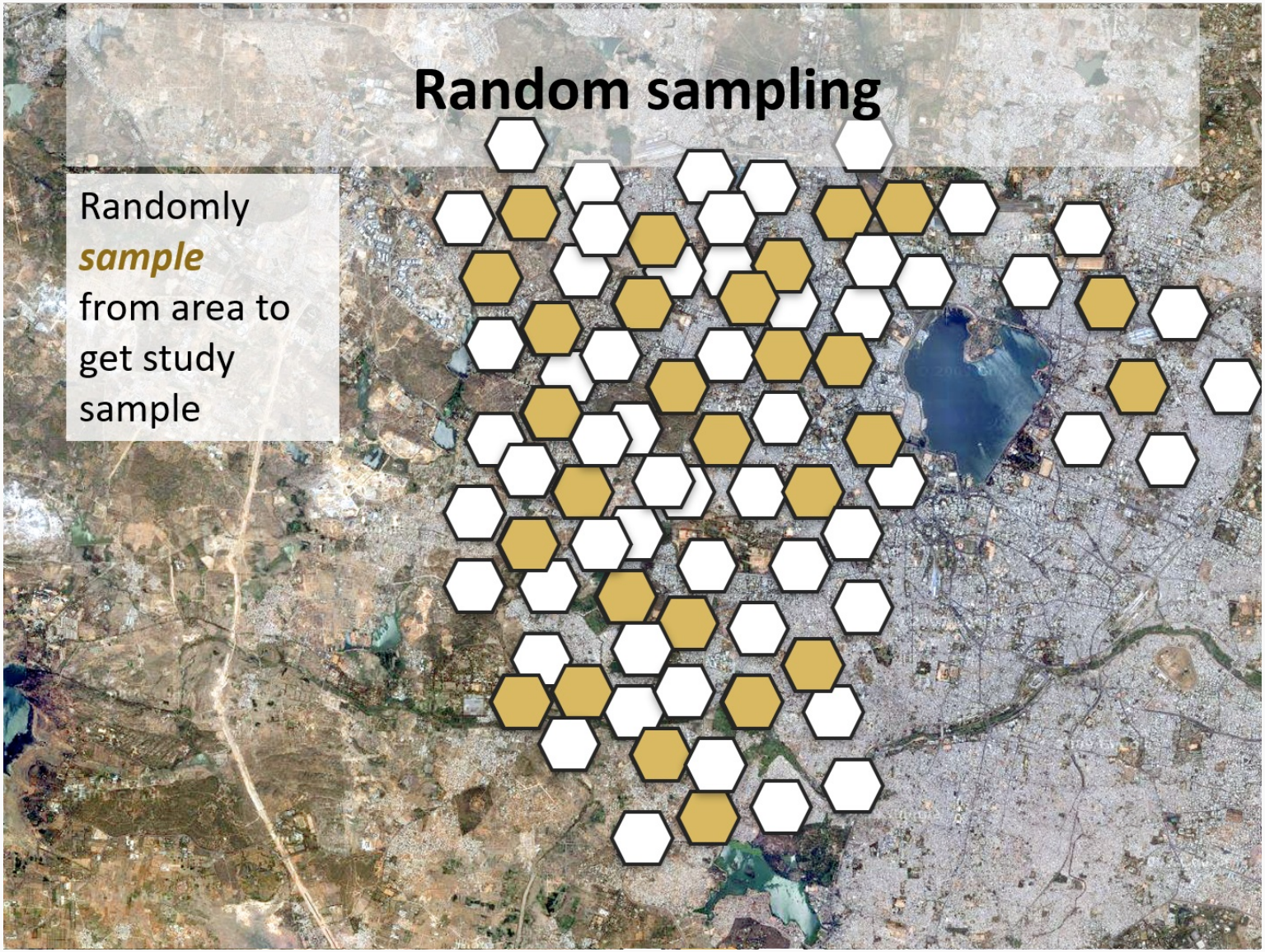
some RCTs assign treatment in groups

- "clustered treatment assignment"
- groups of individuals (neighborhoods, schools, . . .) jointly assigned to T or C
- rationales:
  - simplify treatment/data collection (e.g., minimizing travel costs)
  - treatments cannot be administered to individuals (e.g., infrastructure)
  - to avoid (or study) interaction effects (e.g., externalities, spillovers)

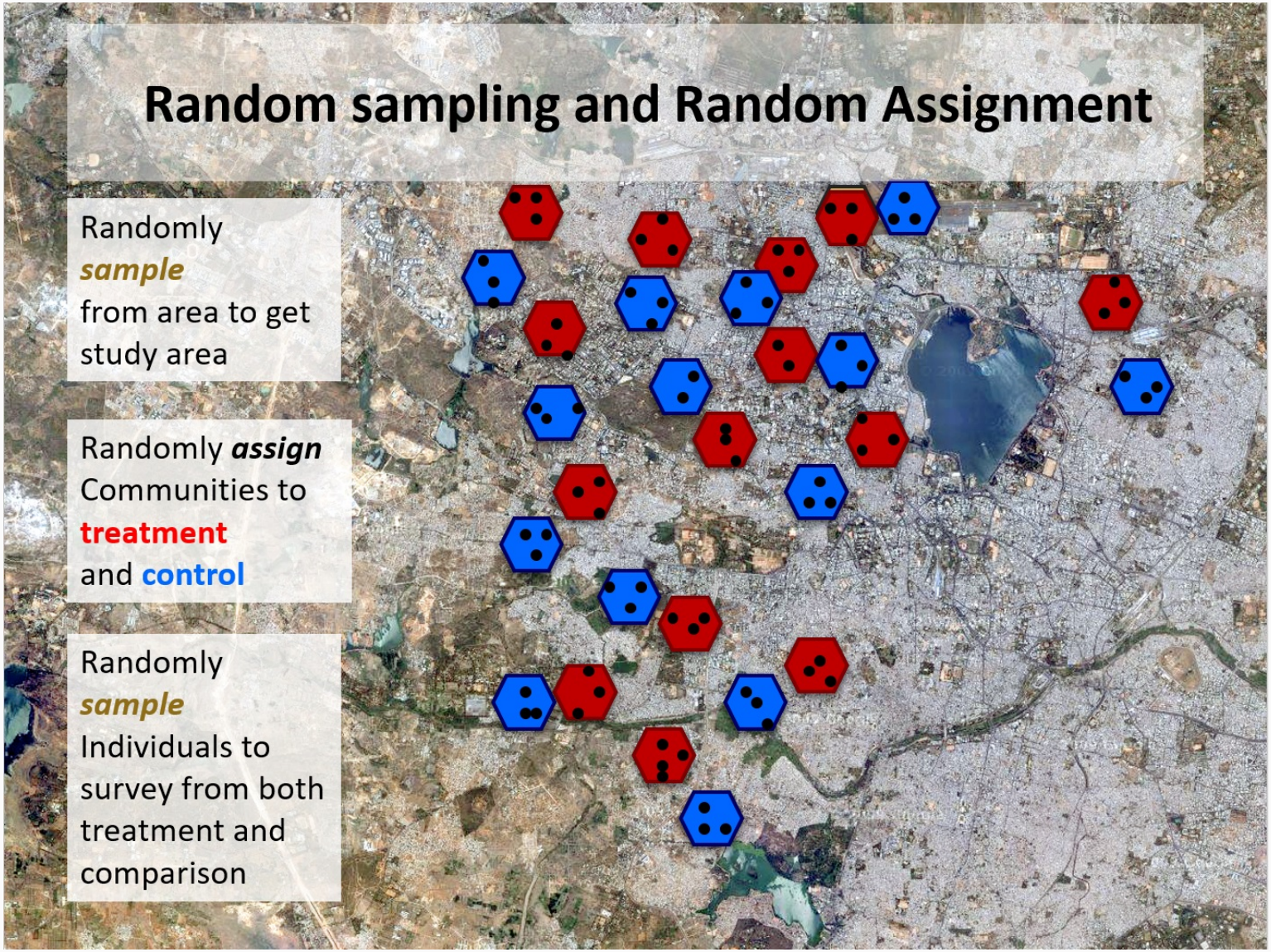
often treatment assignment is stratified (aka blocked)

- form groups (or pairs) of units with similar characteristics and assign a fraction of each group to treatment
- rationales:
  - ensure characteristics used to form groups are balanced.
    (In expectation they are always balanced.)

Example with neighborhood-level treatment



Example with neighborhood-level treatment



How to estimate the impact of a program with an RCT?

compare means, or run a regression:

$$y = \beta_0 + \beta_1 \text{treatment} + ... + \varepsilon$$

- "treatment" is a dummy variable indicating the treatment group
- β_0 captures the mean in the control group
- β_1 captures the treatment effect
- ε might be correlated across obs in the same treatment cluster:
  - needs to be addressed (e.g., compute cluster standard errors).
- control variables
  - are not required for identification in RCTs
  - can improve estimation precision
  - can use machine learning to choose optimal controls (Belloni, Chernozhukov, and Hansen 2014)
- controlling for randomization stratification fixed effects?
  - is recommended to increase power (Bruhn and McKenzie 2009)

How is the impact of a program estimated with an RCT in practice?

Compare means: $\text{impact} = \overline{Y(\text{treatment})} - \overline{Y(\text{control})}$

Through regression analysis:

$$y = \beta_0 + \beta_1 \text{treatment} + \varepsilon$$

- Stata:

  `regress y treatment`

  - With stratification and/or clustering, we will have to take that into account:
    - With clusters indicated by a variable `cluster_ID`:

      `regress Y treatment, cluster(cluster_id)`
    - With strata indicated by a variable `stratum_ID`:

      `regress Y treatment i.stratum_id`
- R:

  `lm(y ~ treatment)`

RCT regression output example

```
. reg wage treatment, cluster(cluster_id)
Linear regression                               Number of obs     =        531
                                                F(12, 55)         =       5.26
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0760
                                                Root MSE          =     .93197
                                  (Std. Err. adjusted for 56 clusters in village)
------------------------------------------------------------------------------
wage         |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   treatment |   .2059332   .1022019     2.01   0.049     .0011159    .4107504
------------------------------------------------------------------------------
```

Critical assessment: Implementation issues (internal validity)

- **Compliance**
  - some subjects do not follow assigned treatment
- **Attrition**
  - some subjects drop out of the sample
- **Externalities and spillovers**
  - treated units and control units interact or affect each other
- **Hawthorne effects**
  - subjects change behavior because under observation, e.g., extra effort
- **John Henry effect**
  - subjects change behavior because in control group, e.g., to overcome the –presumed– disadvantage of being in control group

Also:

- **Ethical concerns**?

Critical assessment: Limits of findings (external validity)

- **Scaling up**
  - equilibrium effects: scaling up programs may lead to changes in prices or behavior of people, which may cause outcomes to differ
    - E.g., if providing a training program for all people substantially increases skilled labor supply, wages for skilled labor may fall, and the benefits per individual may be smaller than in a small-scale RCT
    - Egger et al. (2022) find aid in Kenya had (small) effects on local price inflation
  - political economy effects, difficulties with implementation at scale
    - researchers/NGOs in small-scale projects might be more devoted
  - External validity: How general are results?
    - Effect at another time might be different
    - Effect at another location might be different
- **Treatment intensity**: How to extrapolate out of sample?

These limits are not specific to RCTs (also apply to other methods)

Recap: RCT basics

- randomization to ensure a counterfactual outcome can be estimated
- 4 steps to an RCT
- modes of treatment assignment (e.g., clustered or stratified)
- estimating treatment effects with regression
- common threats to internal and external validity

Next:

- an RCT on microfinance

Miracle of microfinance?

Miracle of microfinance? A seminal study on microfinance

- poor people often have limited access to banking
- microfinance was thought by some to be a simple universal solution

Does microfinance work?

- large growth in microfinance over recent decades
- many *stories* of successes, but also of failure
- until 2015, there were few robust impact evaluations of microfinance programs
- consider in detail one impact evaluation from India (Banerjee et al. 2015)
  - randomized microfinance, thus eliminating selection bias to then compare welfare in neighborhoods with and without microfinance

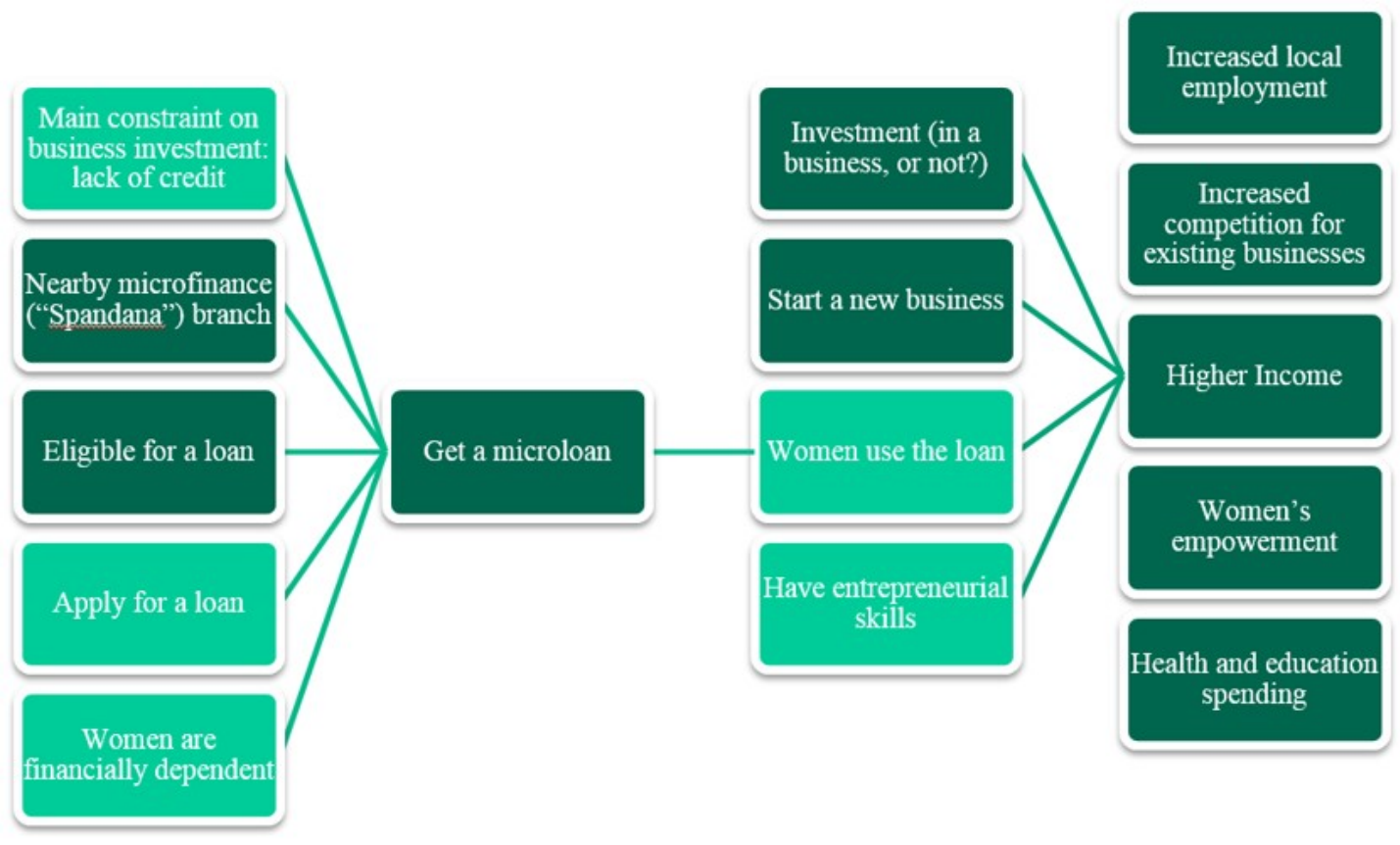
An impact evaluation: Theory of change and RCT

Many development impact evaluations consist of two basic components

1. *A theory of change*
   - spells out all the outcomes we think will change with microfinance based on economic theory (consumption, investement, business operations, . . .)
2. An RCT where some get treated and other do not
   - outcomes from the *theory of changed* are measured in surveys

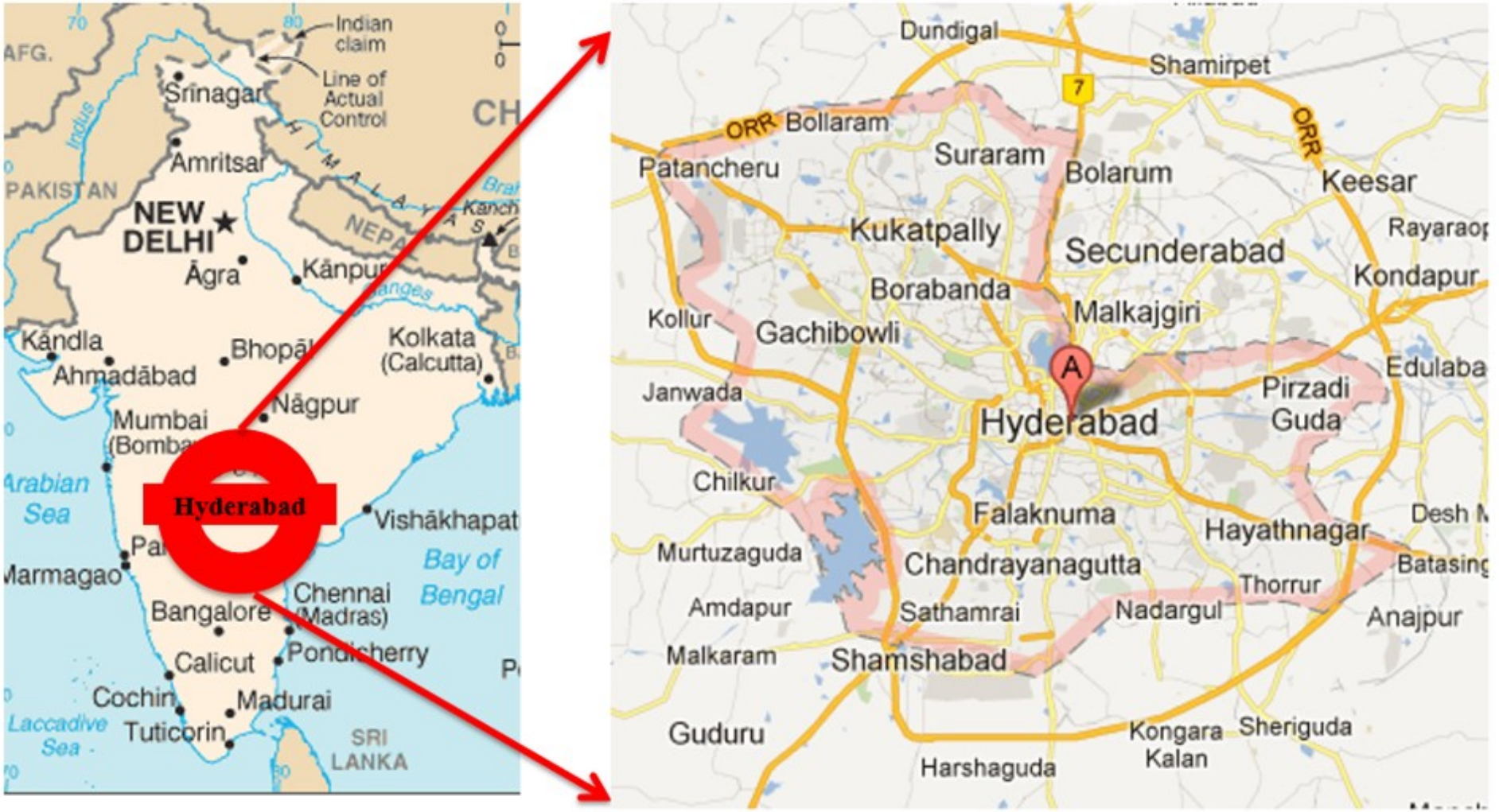
Theory of change: Example of entrepreneurship

- Savings: Microcredit helps women save for durable goods
- Consumption: Microcredit allows women to increase consumption temporarily
- Entrepreneurship: Microcredit helps women create or expand businesses



(Glennerster and Takavarasha 2013)

Setting: Hyderabad, India



(Glennerster and Takavarasha 2013)

Setting (2)

The loans

- group-loans; groups of 6-10 women; 25-45 groups form a "center"
- women are jointly responsible for the loans of their group
- the first loan is Rs. 10,000, about $200
- repayment over 50 weeks; interest rate 12 percent
- if all repay, loan amounts increase up to Rs. 20,000

Eligibility

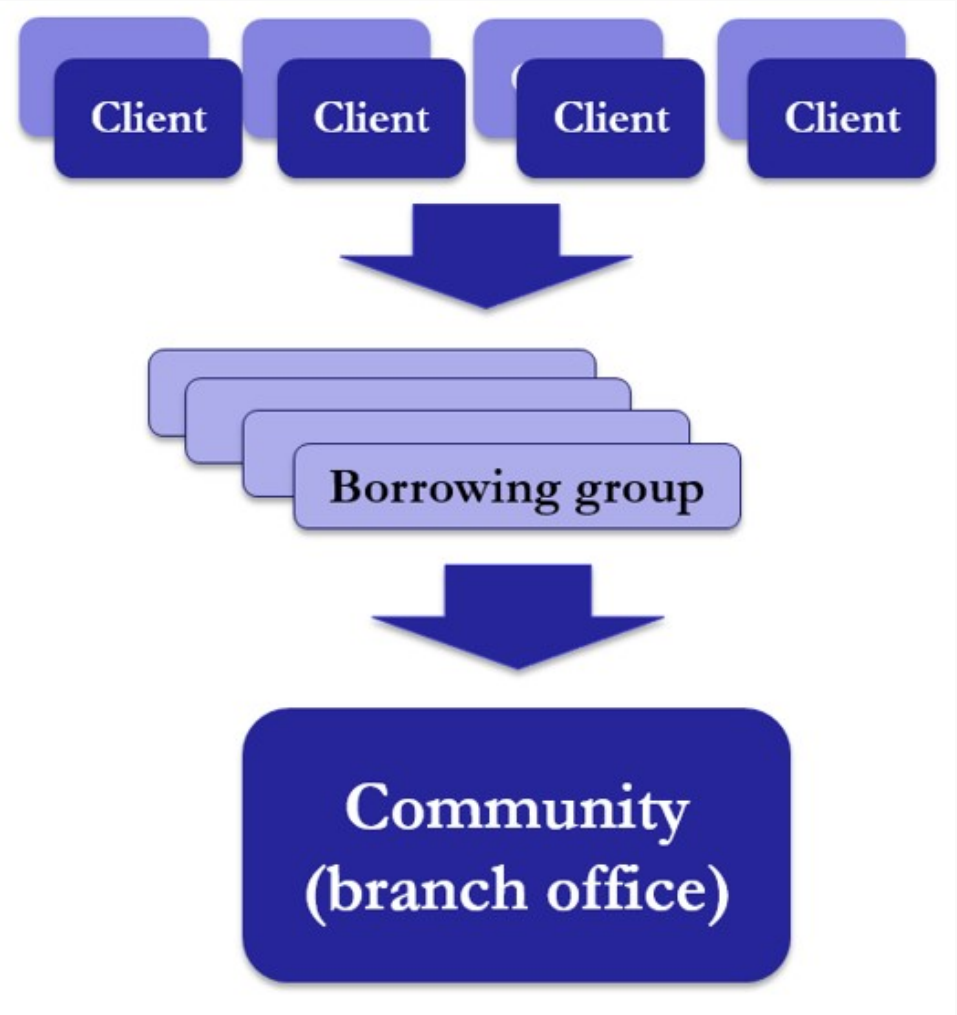
- women, aged 18 to 59
- must have resided in the same area for at least one year
- must have valid identification and residential proof
- at least 80 percent of women in a group must own their home
- groups are formed by women themselves, not by Spandana

Setting (3)

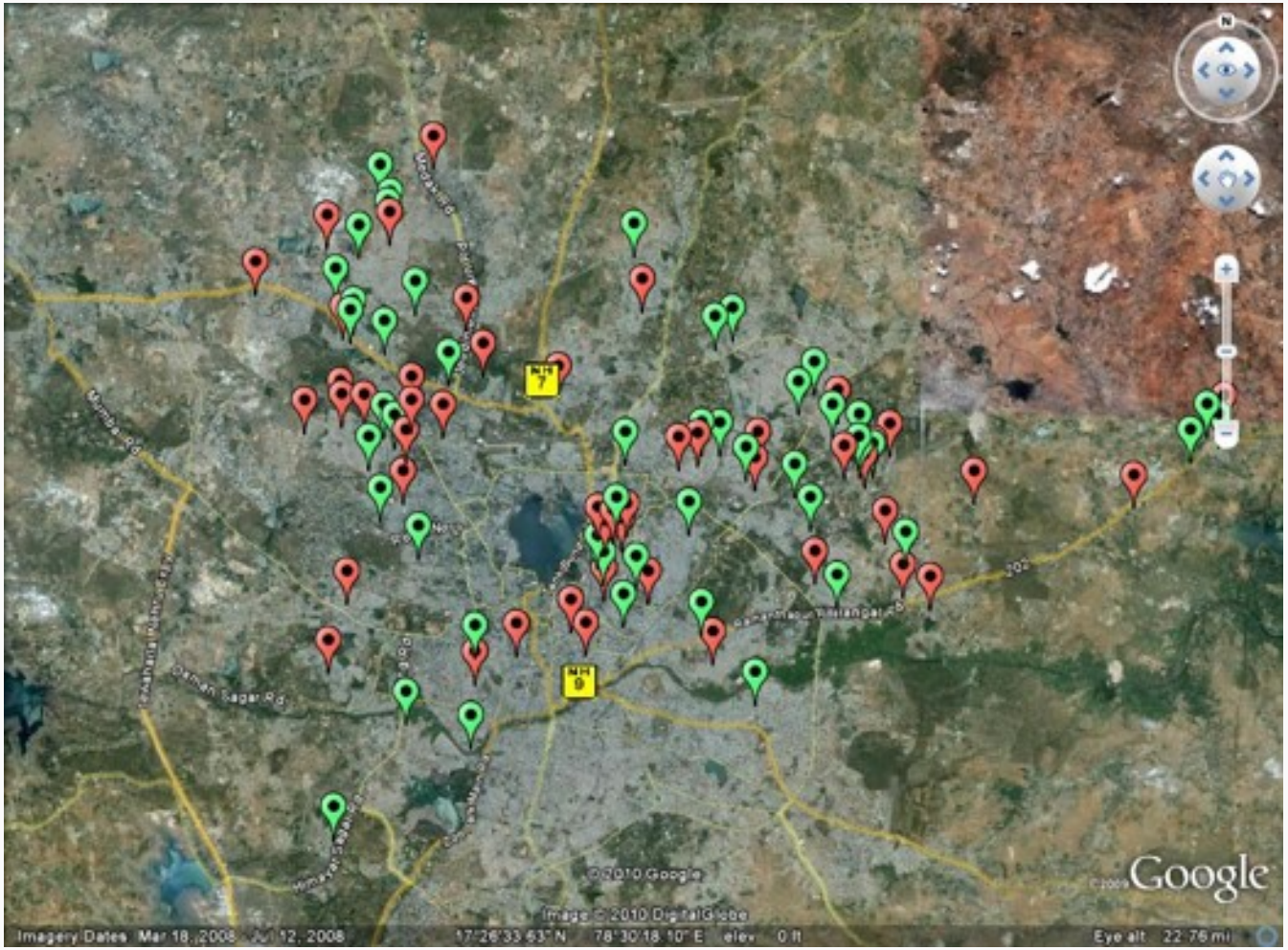
- $1/3$ of Hyderabad's population lives in slums
- in 2004, no MFIs were working in targeted neighborhoods
- yet 69% of households had an informal loan
- average expenditure per person per month: $18
- average household debt $670
- literacy rate: 68%

Unit of randomization?

- practical considerations for implementation
- econometric considerations, e.g., spillovers

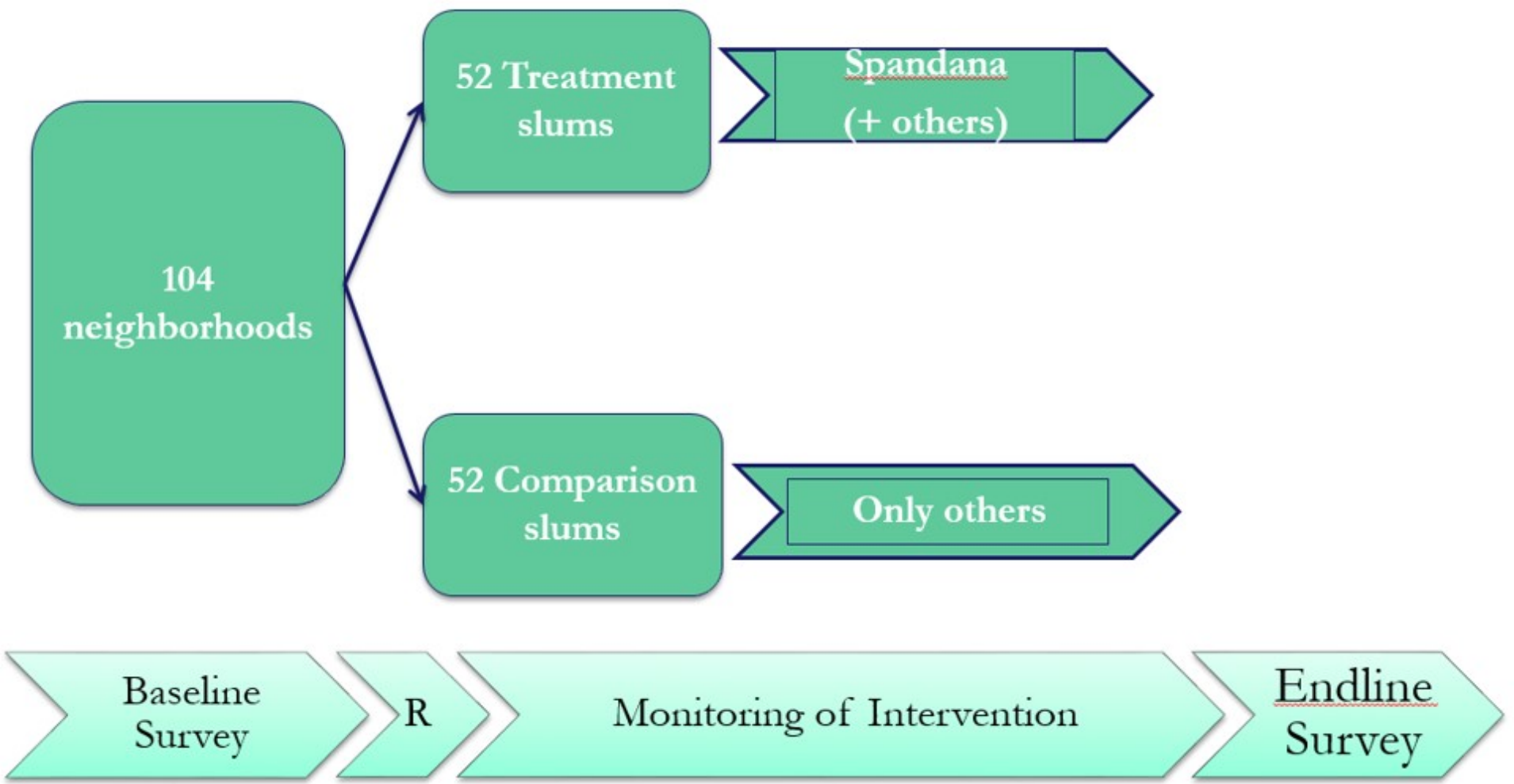


Treatment and control areas



(Glennerster and Takavarasha 2013)

Research design



(Glennerster and Takavarasha 2013)

Measurement

- each endpoint in the *theory of change* is one indicator for data collection

Indicator	Instrument
Investment	Number of businesses per household; business size; duration; costs and revenue; sales
Consumption	Monthly expenditures of the household, itemized; "Special" spending (e.g. weddings)
Women's empowerment	Decision-making by household members
Health and education	Number of health events; tuition spending; education completed of all household members

Analysis

- collected data in 2007-2008
- complication: other microfinance organizations entered control communities (as well as treatment communities)
  - take-up in the control group was 18.7%.
  - take-up in the treatment group was 27%.
  - $\Rightarrow$ comparing treatment and control does not imply a comparison of "access to microfinance" with "no access to microfinance"
    - but variation in the "degree of access to finance"
    - i.e., intention to treat (ITT) analysis

First results: Borrowing

	Spandana (1)	Other MFI (2)	Any MFI (3)	Other bank (4)	Informal (5)	Total (6)	Ever late on payment? (7)	Number of cycles borrowed from an MFI (8)	Index of dependent variables (9)
Panel A. Endline 1									
Credit access									
Treated area	0.127***	−0.012	0.084***	0.003	−0.052**	−0.023	−0.060**	0.084**	0.106***
	(0.020)	(0.024)	(0.027)	(0.012)	(0.021)	(0.014)	−0.026	(0.041)	(0.0291)
Observations	6,811	6,657	6,811	6,811	6,811	6,862	6,475	6,811	6,862
Control mean	0.051	0.149	0.183	0.079	0.761	0.867	0.616	0.330	0.000
Hochberg-corrected *p*-value									0.000
Loan amounts (*in Rupees*)									
Treated area	1,334***	−94	1,286***	75	−1,069	2,856			
	(230)	(336)	(439)	(2,163)	(2,520)	(4,548)			
Observations	6,811	6,708	6,811	6,811	6,811	6,862			
Control mean	597	1,806	2374	8,422	41,045	59,836			

Source: Banerjee et al. (2015), Table 2 (excerpt)

Results: Self-employment/businesses (all households)

	Assets (stock) (1)	Investment in last 12 months (2)	Expenses (3)	Profit (4)	Has a self-employment activity (5)	Number of self-employment activities (6)	Has started a business in the last 12 months (7)	Has closed a business in the last 12 months (8)	Index of dependent variables (9)
Panel A. Endline 1									
Treated area	598	391*	255	354	0.0083	0.018	0.009	0.002	0.0357
	(384)	(213)	(1,056)	(314)	(0.0215)	(0.0380)	(0.006)	(0.008)	(0.0188)
Observations	6,800	6,800	6,685	6,239	6,810	6,810	6,757	2,352	6,810
Control mean	2,498	280	4,055	745	0.349	0.503	0.047	0.037	0.000
Hochberg-corrected *p*-value									0.175
Panel B. Endline 2									
Treated area	1,261**	−134	−530	542	0.023	0.045	−0.000	−0.000	0.0151
	(530)	(207)	(547)	(372)	(0.023)	(0.040)	(0.010)	(0.006)	(0.0186)
Observations	6,142	6,142	6,116	6,090	6,142	6,142	6,142	6,142	6,142
Control mean	5,003	1,007	5,225	953	0.418	0.561	0.083	0.053	0.000
Hochberg-corrected *p*-value									>0.999

Source: Banerjee et al. (2015), Table 3

Results: Consumption

	Total (1)	Durables (2)	Nondurable (3)	Food (4)	Health (5)	Education (6)	Temptation goods (7)	Festivals and celebrations (8)	Home durable good index (9)
Panel A. Endline 1									
Treated area	10.24	19.73*	−6.50	−12.11	−3.7	−2.061	−8.785*	−14.16*	−0.051
	(37.22)	(11.35)	(31.81)	(12.06)	(11.51)	(9.865)	(4.92)	(8.09)	(0.057)
Observations	6,827	6,781	6,781	6,827	6,827	5,415	6,827	6,827	6,841
Control mean	1,419	116	1,305	525	140	168	84	69	2.37
Hochberg-corrected *p*-value	>0.999								
Panel B. Endline 2									
Treated area	−48.83	0.42	−45.45	−11.20	−22.54	12.16	−10.07	6.17	−0.0127
	(51.53)	(9.88)	(46.92)	(17.88)	(17.50)	(15.19)	(6.61)	(4.12)	(0.0426)
Observations	6,142	6,140	6,142	6,142	6,141	4,910	6,142	6,103	6,142
Control mean	1,914	131	1,755	687	187	206	118	90	2.66
Hochberg-corrected *p*-value	0.691								

Notes: Columns 1–8: Monthly per capita household expenditures. Temptation goods include alcohol, tobacco, betel leaves, gambling, and food consumed outside the home. Column 9 calculated on a list of 40 home durable goods (stock, not flow). Each asset is given a weight using the coefficients of the first factor of a principal component analysis. The index, for a household *i*, is calculated as the weighted sum of standardized dummies equal to 1 if the household owns the durable good, 0 otherwise. See online Appendix 1 for description of the construction of the consumption variables. *p*-values for the regression in column 1 (total consumption) reported using Hochberg's step-up method to control the FWER across all outcomes. See text for details.

Results: Summary

People in neighbourhoods where microfinance was available:

- Had more MF loans
- Were not more indebted
- Did not have more successful businesses
- Consumed less "Temptation goods" and more Durables in the short run
- Did not consume more in the long run

MF had some limited effects, but no positive transformative "shock"

- example of a long controversial debate involving many RCTs, academics, policy makers, and the public

External validity?

- Urban vs. rural?
- Selection of neighborhoods without MFI in 2005?
- India vs. other countries?
  - Largely comparable results in Ethiopia, Morocco, Bosnia-Herzegovina, Mexico, and Mongolia
  - see 6 papers in **AEJ: Applied**, January 2015

▸ Do six studies show external validity? ▸ Other empirical concerns? ▸ Further (optional) materials

Optional

Optional

- Balance
- Attrition
- Multiple testing
- Power calculations

Balance

Balance - Are the two groups really the same?

- During treatment randomization, stratification ("blocking") can ensure balance on baseline observables
  - Can increase statistical power
  - We cannot stratify on *all* variables or on unobservables
- After randomization, we can test if T&C are different
  - In expectation, the groups are identical
  - In a given experiment, they will be ~different
- Many papers report "balance tests" for pre-treatment variables
  - e.g., showing that age, gender, income, ... are similar in T&C
  - If we test many variables, some differences will be significant
    - 1 in 100 will be significant at the 1%-level
    - 1 in 20 will be significant at the 5%-level
  - Is that a problem? What do we learn from those tests?

Balance - So what if the two groups *are* different?

- If differences are large and significant, worry whether randomization was correctly carried out
- If we think it was implemented correctly:
  - unbiasedness of RCT estimates is not conditional on insignificant differences in controls or pre-treatment variables
  - p-values, std. errors, ... are designed for "balance in expectation" (holds by design in RCTs) and account for the possibility of chance imbalances
- Still worried that observed differences in outcomes are not causal effects but pre-existing imbalances in expected outcomes?
  - Methods exist to adjust inference for chance imbalances (e.g., Hennessy et al. 2016) ...
  - ... and debates exist whether to avoid these (Mutz, Pemantle, and Pham 2019)
- In practice, balance tables are often used to convince readers that randomization was carried out well and to provide summary statistics

Attrition

Attrition - When subjects quit the sample

- During a study, respondents might drop out before data collection (migrate, pass away, not want to participate anymore, . . .)
- Independence assumption might not hold in remaining sample
  - $\Rightarrow$ selection bias

Ex.: Attrition based on negative treatment effects

Treatment has positive effects on some and negative effects on others. The negatively affected stop participating.

Comparing the remaining (positively affected) to the control group, while ignoring the unobserved (negatively affected) people who left yields biased results.

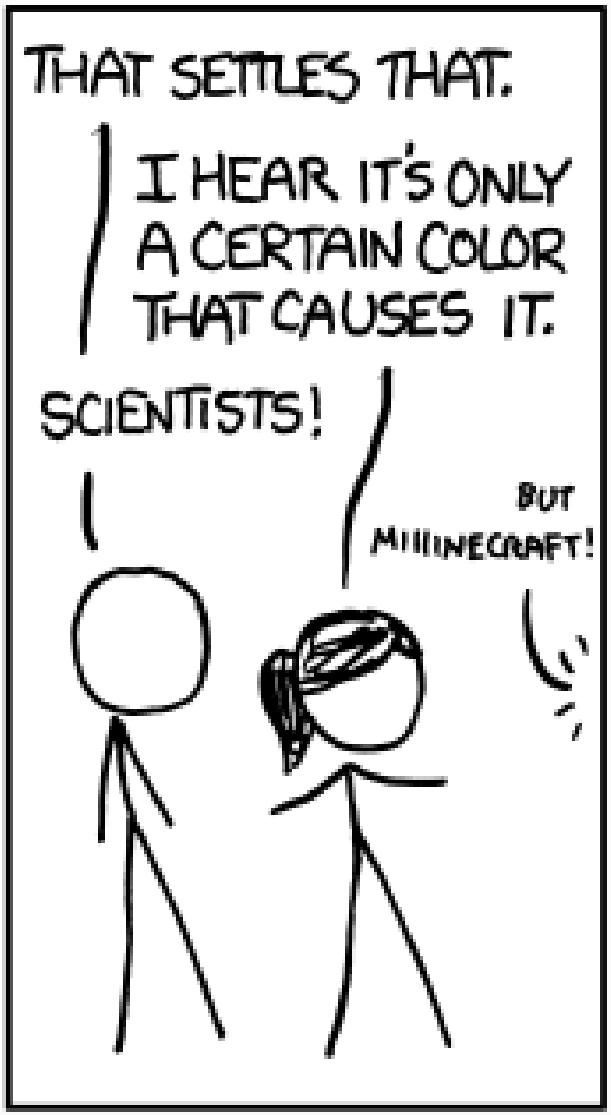
Attrition - How to investigate if this causes a bias?

Two (complementary) approaches are common

1. Testing differential attrition: Is attrition as good as random?
   (i.e., independent from treatment or potential outcomes)
   - Are the rates of attrition the same in T&C?
     - e.g., regress an attrition dummy on treatment
   - Are the patterns of attrition the same in T&C?
     - e.g., regress an attrition dummy on interactions of treatment and age
2. Estimate bounds: Is attrition large enough to change conclusions?
   - Impute extreme outcomes for missing observations ("Manski bounds")
     - High values for missing T and low values for missing C
       $\Rightarrow$ Upper bound estimate
     - High values for missing C and low values for missing T
       $\Rightarrow$ Lower bound estimate
   - Alternative: "Lee bounds"
     - remove high/low observations from the group with the least attrition to estimate lower upper/bounds

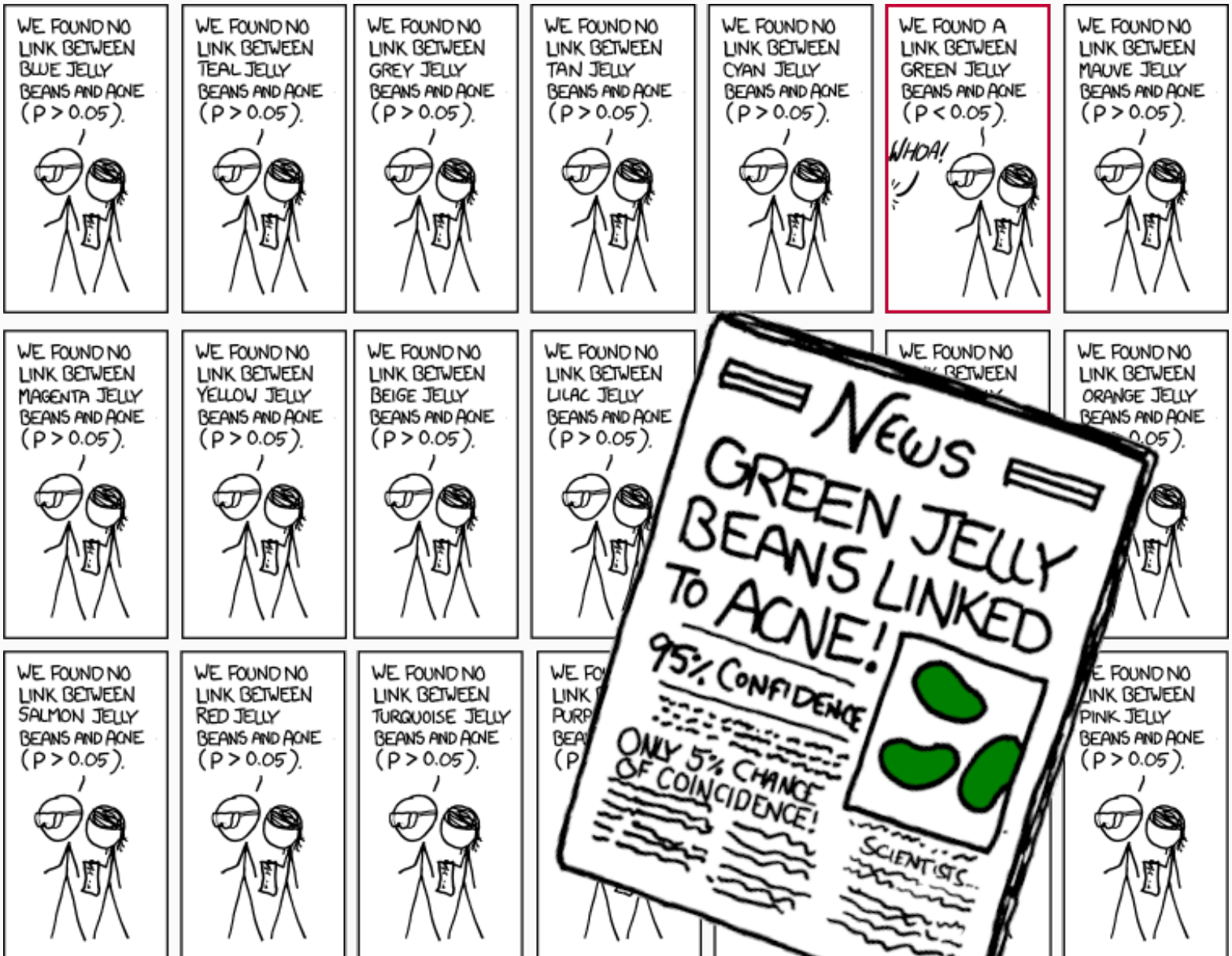
Multiple testing

Running many tests



Source: xkcd.com

Running many tests



Running many tests . . .

. . . increase the likelihood of drawing least one "false positive" conclusion.

The problem exists irrespective of whether we . . .

- test different outcomes against a single treatment, or
- test the same outcome regression with different controls.

What can we do to make our results more credible?

Solutions to the multiple testing problem

1) When many related outcome variables are tested: Use an index
2) When many separate outcomes are used: Use p-value corrections
3) When you need to convince an audience that you did not run all separate tests and only show the one that popped up significant: Use a pre-analysis plan.

Pre-analysis plans

- ... are a step-by-step plan setting out how a researcher will analyze data written in advance of them seeing/collecting this data
  - specify how data will be prepared, regression specifications, how variables will be constructed, etc.
  - formulated before data analysis/collection
- some debate in the literature as to when/whether to do one and how extensive it should be:
  - early pre-analysis plans often were 30+ pages, tried to pre-specify all eventualities
  - Olken (2015) suggests simpler plans may have most advantages
  - Coffman and Niederle (2015) – may not be needed if easy to replicate (e.g., lab experiments)

Pros of pre-analysis plans

- most useful for field experiments that may be expensive and difficult to replicate
  - (other false positives in academic papers may get corrected through replications, Coffman and Niederle 2015)
- credibility
- several other uses:
  - helps focus on what key outcomes one most cares about and get agreement on this in advance rather than after seeing results
  - forces you to think through a design before implementation (and design a questionnaire that has (only) what is required).

Recap: Multiple inference

statistical testing is a "decision problem":

- need to make a decision based on "noisy" information
- willing to accept *some* error in that.
- if we increase the number of decisions we make, errors become more likely
- we discussed several methods to
  - keep the number of test decisions low (indexing)
  - adjust the decision rule to keep the number/rate of overall errors low (p-value adjustments)
  - be transparent about the number of test decisions (pre-analysis plans)

Power calculations

Power calculations

Important concepts:

- **significance** level: α - probability of a **false positive**
- **power** level: $1 - \beta$, one minus the probability of a **false negative**
- minimum detectable effect size (MDES): smallest effect a study has sufficient power to detect at given significance level and power level

Power calculations cen be used to either

1. compute required sample size (given power and MDES)
2. compute power (given sample size and MDES)
3. compute MDES (given power and sample size)

Power calculations help design the study in a way that will allow for meaningful conclusions.

To common ways to do power calculations

1. by simulating data with properties that we expect to find in the "field" (samples size, effect size, variance) and testing if/when we can detect an effect with the methods we intend to use (e.g., regression)
2. based on theory/properties of the estimators to be used

Ex.: Power calculations for t-tests

$$n = \frac{4\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{D^2},$$

- n - sample size
- σ^2 - variance of the outcome
- z - critical values (e.g. from a Normal)
- D - MDES

What else about power calculations

Aside from telling us how big a data set to collect in our own study, power calculations help us answer an important question about existing studies:

What to make of an insignificant effect estimate?

If there's no significant effect, is this because there is no effect, or because the study designed in a way that makes us unable to detect an effect?

I.e., was the study "under-powered" to find an effect?

Recap

Recap

- 4 steps to an RCT
- Microfinance example

Optional:

- Baseline balance
- Attrition
- Multiple testing
- Power calculations

Next: Regression discontinuity

- I will email a problem set with data on RCTs
- Please read Suri, Bharadwaj, and Jack (2021)

Bibliography

Banerjee, Abhijit, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan. 2015. "The Miracle of Microfinance? Evidence from a Randomized Evaluation." *American Economic Journal: Applied Economics* 7 (1): 22–53.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on Treatment Effects After Selection Among High-Dimensional Controls." *The Review of Economic Studies* 81 (2): 608–50.

Bruhn, Miriam, and David McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics* 1 (4): 200–232.

Dahal, Mahesh, and Nathan Fiala. 2020. "What Do We Know about the Impact of Microfinance? The Problems of Statistical Power and Precision." *World Development* 128: 104773.

Egger, Dennis, Johannes Haushofer, Edward Miguel, Paul Niehaus, and Michael Walker. 2022. "General Equilibrium Effects of Cash Transfers: Experimental Evidence from Kenya." *Econometrica* 90 (6): 2603–43.

Glennerster, Rachel, and Kudzai Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide*. Princeton University Press, Slides available from runningres.com.

Hennessy, Jonathan, Tirthankar Dasgupta, Luke Miratrix, Cassandra Pattanayak, and Pradipta Sarkar. 2016. "A Conditional Randomization Test to Account for Covariate Imbalance in Randomized Experiments." *Journal of Causal Inference* 4 (1): 61–80.

Meager, Rachael. 2019. "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments." *American Economic Journal: Applied Economics* 11 (1): 57–91.

Mutz, Diana C, Robin Pemantle, and Philip Pham. 2019. "The Perils of Balance Testing in Experimental Design: Messy Analyses of Clean Data." *The American Statistician* 73 (1): 32–42.

Pritchett, Lant, and Justin Sandefur. 2015. "Learning from Experiments When Context Matters." *American Economic Review* 105 (5): 471–75.

Ravallion, Martin et al. 2018. "Should the Randomistas (Continue to) Rule." *Center for Global Development Working Paper*. Vol. 492.

Suri, Tavneet, Prashant Bharadwaj, and William Jack. 2021. "Fintech and Household Resilience to Shocks: Evidence from Digital Loans in Kenya." *Journal of Development Economics* 153: 102697.

Wydick, Bruce. 2016. "Microfinance on the Margin: Why Recent Impact Studies May Understate Average Treatment Effects." *Journal of Development Effectiveness* 8 (2): 257–65.

Appendix

Do six studies show external validity?

- no clear way to extrapolate to microfinance in other regions (Pritchett and Sandefur 2015)
- the six studies reflect very different microfinance markets
  - early stages: Ethiopia, urban India, and Morocco
  - near-saturated market: Bosnia-Herzegovina, Mexico, and Mongolia
  - in saturated market, experiments can only be run in marginal contexts, where effects are weakest (Wydick 2016)

Other empirical concerns?

- Entry of other MFIs non-random?
- Statistical power (sample sizes)
  - Dahal and Fiala (2020):
    - "Every one of the studies is significantly underpowered to detect reasonable effect sizes"?
    - Find significant effect on a pooled sample
  - Meager (2019):
    - "I jointly estimate the average effect and the heterogeneity in effects across seven studies using Bayesian hierarchical models. I find the impact on household business and consumption variables is unlikely to be transformative and may be negligible."

Back Further (optional) materials

- 10 minute video by Josh Angrist on RCTs

  mru.org/courses/mastering-econometrics/introduction-randomized-trials
- Glennerster, Rachel, and Kudzai Takavarasha. Running randomized evaluations: A practical guide. Princeton University Press, 2013.