Implementation Issues for Rigorous Evaluations

Clair Null

November 24, 2008

Ethics

- Which of the 4 methods require consideration of ethical issues?

Ethics

- Which of the 4 methods require consideration of ethical issues?
  - Randomized Evaluations
  - Regression Discontinuity
  - Matching
  - Differences-in-Differences

Ethics

- Which of the 4 methods require consideration of ethical issues?
  - Randomized Evaluations
  - Regression Discontinuity
  - Matching
  - Differences-in-Differences
- Biggest worry with randomized evaluations
  - b/c researcher has control over how people are treated
  - "quasi-experimental" methods usually rely on pre-existing data

Ethical Basis for Randomization

- Un-ethical NOT to try to learn what works best
- If there's not enough to go around, randomizing who gets it & who doesn't is pretty fair

Ethical Basis for Randomization

- Un-ethical NOT to try to learn what works best
- If there's not enough to go around, randomizing who gets it & who doesn't is pretty fair
  - lottery (what it sounds like)
    - housing vouchers
  - phase-in (eventually everybody gets "treated")
    - de-worming
  - encouragement (vary the intensity of encouragement and use that to estimate effect of treatment)
    - invite random subset of friends to witness fertilizer application
  - cluster (every large unit gets treated in some way, but individuals within unit are randomized)
    - tutors (some schools get 4th grade, others get 3rd grade)

Ethical Basis for Randomization

- Un-ethical NOT to try to learn what works best
- If there's not enough to go around, randomizing who gets it & who doesn't is pretty fair
  - lottery (what it sounds like)
    - housing vouchers
  - phase-in (eventually everybody gets "treated")
    - de-worming
  - encouragement (vary the intensity of encouragement and use that to estimate effect of treatment)
    - invite random subset of friends to witness fertilizer application
  - cluster (every large unit gets treated in some way, but individuals within unit are randomized)
    - tutors (some schools get 4th grade, others get 3rd grade)
- Committee for the Protection of Human Subjects (CPHS) / Internal Review Board (IRB)

Externalities

- Social benefits (costs) not the same as private benefits (costs)
- For impact evaluation studies, this means that the treatment group might "contaminate" the control group (or vice versa)
- Which of the 4 methods does this cause a concern for?

Externalities

- Social benefits (costs) not the same as private benefits (costs)
- For impact evaluation studies, this means that the treatment group might "contaminate" the control group (or vice versa)
- Which of the 4 methods does this cause a concern for?
  - ALL OF THEM

Externalities

- Social benefits (costs) not the same as private benefits (costs)
- For impact evaluation studies, this means that the treatment group might "contaminate" the control group (or vice versa)
- Which of the 4 methods does this cause a concern for?
  - ALL OF THEM
- Need to think carefully about potential for externalities when designing randomized evaluations
  - you might actually be able to do something to avoid them, whereas no control over data for other 3 methods
- Difficult to measure, but can sometimes use variation in exposure to treatment
- At the very least, need to think about how externalities might bias your results
  - can go in either direction (overstate or understate program effect depending on situation)

ITT vs. TOT

ITT vs. TOT

- If there's imperfect compliance (to either treatment or control), you'll end up estimating the effect of the intention to treat (ITT)

- Sometimes (but not always) we're only interested in the effect of the treatment on the treated (TOT)
  - can use the Wald Estimator (ratio of differences in outcomes to differences in compliance for T & C)
  - assignment to treatment as an instrumental variable (affects your outcome only through its effect on whether or not you got treated)

Logistics

- Attrition (only matters for bias if it's "differential"; anybody dropping out is bad for power)

Logistics

- Attrition (only matters for bias if it's "differential"; anybody dropping out is bad for power)
- Data quality

Logistics

- Attrition (only matters for bias if it's "differential"; anybody dropping out is bad for power)
- Data quality
  - Is the respondent answering the same question you think you're asking?
    - translation issues
  - Is the respondent giving a truthful answer?
    - particular concern when asking about sensitive topics
    - "courtesy bias"
  - Does the respondent even know the right answer?
    - can use multiple measures to check "reliability ratio"

Statistical Significance

- All we can do is *estimate* the effect of treatment for the population
  - as long as we're working with a sample, we'll never know the true effect for the whole population
- How likely is it that we would have observed this difference between treatment and control even if in truth there really wasn't any difference? (type I error)
- The effect we estimated could be either too big, too small, or just right

Statistical Significance

- If we repeated the whole study with a bunch of different samples, most of the estimates will be about right and only a few will be outliers, but we don't know how our particular estimate compares to the truth
- How far the outliers are likely to be from the truth depends on
  - how big the sample is
  - how much variation there is in the outcome among those in the sample
- The standard error of our estimate tells us how spread out the distribution of all potential estimates (based on different samples) would be

Statistical Significance

- Instead of looking at just our estimate of average outcome among T & C (the "point estimate"), do analysis based on "confidence interval"
  - takes point estimate as the middle and factors in the standard error
- If the confidence intervals for average outcomes among T & C don't overlap, we feel "confident" that the difference between the two groups is probably real and not just due to chance
- Confidence intervals for average outcomes among T & C are less likely to overlap if:
  - the standard errors are small
  - the difference between point estimates for T & C is big

Statistical Power

- If you're designing a randomized evaluation, how big a sample size should you use?
  - goal is to be able to detect a treatment effect (in a statistical sense, using the confidence intervals) whenever T & C are truly different (avoid type II error)
  - but bigger samples cost more money, so no need to overdo it

Statistical Power

- When choosing sample size, factor in:
  - what hypothesis you're trying to test (treatment has no effect versus two treatments differ)
  - what confidence level you want
  - how much variation there is in the comparison group
  - how big the effect size is likely to be
  - inter-group correlation when randomization at cluster level
- Use the (free) Optimal Design software to figure out your options based on the answers to these questions (yes, it's a bit of a guessing game)
  - `http://sitemaker.umich.edu/group-based/optimal_design_software`

Internal Validity

- Are we willing to accept the identifying assumption?

Internal Validity

- Are we willing to accept the identifying assumption?
  - randomization: no systematic difference in unobservables btwn T & C
  - RD: no systematic difference in unobservables btwn T & C *at the threshold*
  - matching: *after controlling for observables*, no systematic difference in unobservables btwn T & C
  - diff-in-diff: *no change in* unobservable differences btwn T & C *over time*

Internal Validity

- Are we willing to accept the identifying assumption?
  - randomization: no systematic difference in unobservables btwn T & C
  - RD: no systematic difference in unobservables btwn T & C *at the threshold*
  - matching: *after controlling for observables*, no systematic difference in unobservables btwn T & C
  - diff-in-diff: *no change in* unobservable differences btwn T & C *over time*
- Were externalities appropriately accounted for?

Internal Validity

- Are we willing to accept the identifying assumption?
  - randomization: no systematic difference in unobservables btwn T & C
  - RD: no systematic difference in unobservables btwn T & C *at the threshold*
  - matching: *after controlling for observables*, no systematic difference in unobservables btwn T & C
  - diff-in-diff: *no change in* unobservable differences btwn T & C *over time*
- Were externalities appropriately accounted for?
- And what about all the other concerns we discussed today?
  - attrition, data quality, etc.

External Validity

- To what extent are we willing to generalize the study's results to another context?

- Relative to the rest of the world (or whatever other context you're interested in):
  - How representative are the participants?
  - How representative is their environment? (physical & cultural)
  - How representative is the intervention?
  - How representative is the scale of the intervention?
    - "general equilibrium effects"

Cost Effectiveness & Cost-Benefit Analysis

- Great, so your estimate of the effect of some program is both internally and externally valid
- Now what do you do with it?
- Ultimately, with scarce resources, want to pick the programs with the biggest bang for the buck
- Cost-effectiveness is the first step
    - compare cost of program to what it actually accomplished (usually in terms of one outcome)
- Cost-benefit analysis factors in all the changes that result from a program for a more holistic estimate of whether or not the program was "worth it"
- Hard to do, but ideally cost-benefit analysis lets us compare various programs to choose the best one

Mechanisms

- So the program had an effect. The next obvious question is *why*?
  - Were the effects on the intensive or the extensive margin?
- If we understand why, we can design other programs that address that mechanism that are likely to be successful
  - again, internally valid? externally valid?
- Harder to design randomized evaluations that address these sort of questions, but very valuable
  - even harder to answer why using other methods, but worth trying (might be able to at least provide suggestive evidence)

Now, put this all to use...

- Ethics
- Externalities
- ITT vs TOT
- Logistics (attrition & data quality)
- Statistical significance and power
- Internal & external validity
- Cost-effectiveness & cost-benefit analysis
- Mechanisms