WORKSHOP ON PRINCIPAL STRATIFICATION STANFORD UNIVERSITY, 2016

Luke W. Miratrix (Harvard University)
Lindsay C. Page (University of Pittsburgh)

Our team!

- Avi Feller (Berkeley)
- Jane Furey (Abt Associates)
- Todd Grindal (Abt Associates)
- Luke Miratrix (Harvard)
- Lindsay Page (University of Pittsburgh)
- Marie-Andree Somers (MDRC)
- Rebecca Unterman (MDRC)

Agenda

- Conceptual framework & key assumptions
- Break
- Small group exercise
  - Principal stratification set up
  - Defining estimands of interest
- Lunch
- Estimation & bounds
- Break
- Wrap up

Potential outcomes framework

ID	$Y_i (Z_i=1)$	$Y_i (Z_i=0)$	Z_i
1	105	?	1
2	90	?	1
3	87	?	1
4	111	?	1
5	?	92	0
6	?	80	0
7	?	103	0
8	?	97	0
…			

Each individual has the potential to experience assignment to the treatment condition or assignment to the control condition.

Therefore, the **potential outcomes** for an individual are what we would see under each condition.

Potential outcomes framework

ID	$Y_i (Z_i=1)$	$Y_i (Z_i=0)$	Z_i
1	105	?	1
2	90	?	1
3	87	?	1
4	111	?	1
5	?	92	0
6	?	80	0
7	?	103	0
8	?	97	0
…			

For outcome Y, the causal effect of the treatment (Z) for individual *i* is:

$$Y_i(1) - Y_i(0)$$

We can never compute this directly, but it can be estimated.

The Average Treatment Effect (ATE) for the sample is:

$$\frac{1}{n}\sum_i \left(Y_i(1) - Y_i(0)\right)$$

This also cannot be computed directly but can be estimated.

Potential outcomes framework: Key assumptions

Assumption 1: Ignorability. Whether we observe $Y_i(1)$ or $Y_i(0)$ for *individual i* is unrelated to the values of the potential outcomes. Random assignment allows us to make this assumption.

$$E[Y_i(1) - Y_i(0)]$$

$$= E[Y_i(1)|Z_i = 1] - E[Y_i(0)|Z_i = 0]$$

Assumption 2: Stable Unit Treatment Value Assumption (SUTVA). The observed outcome for each individual is a function of the potential outcomes and random assignment and is not dependent, for example, on the assignment of other individuals. **The potential outcomes are fixed pre-treatment variables.**

$$Y_{obs,i} = Z_i \times Y_i(1) + (1 - Z_i) \times Y_i\,(0)$$

Therefore,

$$E[Y_i(1)|Z_i = 1] - E[Y_i(0)|Z_i = 0] = E\left[Y_{obs,i}|Z_i = 1\right] - E[Y_{obs,i}|Z_i = 0]$$

Potential outcomes framework: ITT

$$ITT = E\left[Y_{obs,i}|Z_i = 1\right] - E\left[Y_{obs,i}|Z_i = 0\right]$$

The Intent to Treat (ITT) estimand is the causal impact on outcome *Y* of being assigned to the treatment.

Of course, we may be interested in assessing impacts among those who would actually take up the treatment. Therefore, we need to think about the post-treatment or intermediate "behaviors" (such as treatment compliance) that we care about.

Potential values of intermediate behavior

	Potential outcomes		Potential intermediate behavior		
ID	$Y_i(Z_i=1)$	$Y_i(Z_i=0)$	$D_i(Z_i=1)$	$D_i(Z_i=0)$	Z_i
1	105	?	✓	?	1
2	90	?	✓	?	1
3	87	?	✓	?	1
4	111	?	✓	?	1
5	?	92	?	✓	0
6	?	80	?	✓	0
7	?	103	?	✓	0
8	?	97	?	✓	0
…					

Potential values of intermediate behavior

ID	D_i (Z_i=1)	D_i (Z_i=0)	Z_i
1	1	?	1
2	1	?	1
3	0	?	1
4	0	?	1
5	?	0	0
6	?	1	0
7	?	0	0
8	?	1	0
…			

1. Suppose the intermediate behavior is binary – treatment take up.

Potential values of intermediate behavior

ID	D_i (Z_i=1)	D_i (Z_i=0)	Z_i
1	1	0	1
2	1	1	1
3	0	0	1
4	0	1	1
5	1	0	0
6	1	1	0
7	0	0	0
8	0	1	0
…			

1. Suppose the intermediate behavior is binary – treatment take up.

2. Suppose that we are able to observe both potential values of the intermediate behavior for each person.

Then, we could classify individuals according to the potential values of the intermediate behavior of interest (such as exposure to the treatment).

Potential values of intermediate behavior

ID	D_i (Z_i=1)	D_i (Z_i=0)	Z_i
1	1	0	1
2	1	1	1
3	0	0	1
4	0	1	1
5	1	0	0
6	1	1	0
7	0	0	0
8	0	1	0
…			

Compliers:
Take up treatment under assignment to treatment but not under assignment to control

Potential values of intermediate behavior

ID	D_i (Z_i=1)	D_i (Z_i=0)	Z_i
1	1	0	1
2	1	1	1
3	0	0	1
4	0	1	1
5	1	0	0
6	1	1	0
7	0	0	0
8	0	1	0
…			

Compliers:
Take up treatment under assignment to treatment but not under assignment to control

Always-takers:
Always take up treatment

Potential values of intermediate behavior

ID	D_i (Z_i=1)	D_i (Z_i=0)	Z_i
1	1	0	1
2	1	1	1
3	0	0	1
4	0	1	1
5	1	0	0
6	1	1	0
7	0	0	0
8	0	1	0
…			

Compliers:
Take up treatment under assignment to treatment but not under assignment to control

Always-takers:
Always take up treatment

Never-takers:
Never take up treatment

Potential values of intermediate behavior

ID	D_i (Z_i=1)	D_i (Z_i=0)	Z_i
1	1	0	1
2	1	1	1
3	0	0	1
4	0	1	1
5	1	0	0
6	1	1	0
7	0	0	0
8	0	1	0
…			

Compliers:
Take up treatment under assignment to treatment but not under assignment to control

Always-takers:
Always take up treatment

Never-takers:
Never take up treatment

Defiers:
Take up treatment under assignment to control but not under assignment to treatment

Potential values of intermediate behavior

Principal Strata
(Frangakis & Rubin, 2002)

Stratum membership is defined by answers to two questions:

1. What were individuals' intermediate behaviors, given their randomized treatment assignments?

2. *What would their intermediate behaviors have been,* had they been assigned to the counterfactual experimental condition?

Compliers:
Comply with treatment under assignment to treatment but not under assignment to control

Always-takers:
Always take up treatment

Never-takers:
Never take up treatment

Defiers:
Take up treatment under assignment to control but not under assignment to treatment

Potential values of intermediate behavior

Recall **Stable Unit Treatment Value Assumption (SUTVA).** The observed outcome for each individual is a function of the potential outcomes and random assignment and are not dependent, for example, on the assignment of other individuals.

$$Y_{obs,i} = Z_i \times Y_i(1) + (1 - Z_i) \times Y_i\,(0)$$

The same assumption applies to values of intermediate outcomes:

$$D_{obs,i} = Z_i \times D_i(1) + (1 - Z_i) \times D_i\,(0)$$

Another way of saying this is that how an individual responds to the offer or lack thereof of a particular treatment is a characteristic of that individual and is not, for example, influenced by the random assignment of others.

Compliance type, as defined by the joint values of the intermediate behaviors, **is a pre-treatment covariate** (even though it is not fully observed).

Subgroup-specific ITT estimates

$$Overall\ ITT = E\left[Y_{obs,i}|Z_i = 1\right] - E\left[Y_{obs,i}|Z_i = 0\right]$$

Because treatment assignment is uncorrelated with baseline characteristics, we can consider subgroup-specific ITT effects. For example:

$$Overall\ ITT\ for\ males = E\left[Y_{obs,i}|Z_i = 1, male_i = 1\right] - E\left[Y_{obs,i}|Z_i = 0, male_i = 1\right]$$

$$Overall\ ITT\ for\ females = E\left[Y_{obs,i}|Z_i = 1, male_i = 0\right] - E\left[Y_{obs,i}|Z_i = 0, male_i = 0\right]$$

The ITT effect within each subgroup is a causal effect of treatment. HOWEVER, the comparison of causal effects across the two subgroups is a descriptive exercise.

We could say descriptively, for example, that the treatment effect was larger for males than it was for females, but we can't say that gender was the cause of the differential in treatment effect, because we have not experimentally manipulated gender.

Subgroup-specific ITT estimates

$$Overall\ ITT = E\left[Y_{obs,i}|Z_i = 1\right] - E\left[Y_{obs,i}|Z_i = 0\right]$$

Just as we can express ITT effects for fully observed subgroups, we can also express ITT effects for partially observed (or latent) subgroups.

$$ITT_C = E\left[Y_{obs,i}|Z_i = 1, s_i = c\right] - E\left[Y_{obs,i}|Z_i = 0, s_i = c\right]$$

$$ITT_{AT} = E\left[Y_{obs,i}|Z_i = 1, s_i = at\right] - E\left[Y_{obs,i}|Z_i = 0, s_i = at\right]$$

$$ITT_{NT} = E\left[Y_{obs,i}|Z_i = 1, s_i = nt\right] - E\left[Y_{obs,i}|Z_i = 0, s_i = nt\right]$$

$$ITT_D = E\left[Y_{obs,i}|Z_i = 1, s_i = d\right] - E\left[Y_{obs,i}|Z_i = 0, s_i = d\right]$$

These are causal effects. Comparison of ITT effects across subgroups is a descriptive exercise, as we have not randomly manipulated compliance type.

Principal strata

		Z = 0	
		D = 1	**D = 0**
Z = 1	**D = 1**	Always takers	Compliers
	D = 0	Defiers	Never takers

Principal strata

While the defined strata are the subgroups of interest, they are not fully observed.

		Z = 0	
		D = 1	**D = 0**
Z = 1	**D = 1**	Always takers	Compliers
	D = 0	Defiers	Never takers

And so, outcome distributions for groups defined by treatment assignment and treatment take up (e.g., what we are actually able to observe) are **mixtures of distributions** across our strata of interest.

Therefore, we apply sets of assumptions (justified by science, field knowledge and/or theory) to make progress and simplify an otherwise very hard estimation problem.

Removing boxes: the Monotonicity Assumption

$$D_i(1) \geq D_i(0)$$

		Z = 0	
		D = 1	**D = 0**
Z = 1	**D = 1**	Always takers	Compliers
	D = 0	Defiers	Never takers

In IV, monotonicity is a standard assumption which states that the potential value of the intermediate variable under assignment to treatment is at least as large as the potential value of the intermediate variable under assignment to control. For example, assignment to treatment can't decrease tendency to take treatment.

Therefore, applying the monotonicity assumption leads us to "assume away" defiers.

Constraining the outcomes: the Exclusion Restriction

$$ITT_{AT} = 0 \qquad ITT_{NT} = 0$$

		Z = 0	
		D = 1	**D = 0**
Z = 1	**D = 1**	Always takers	Compliers
	D = 0	Defiers	Never takers

In IV, the exclusion restriction states that random assignment can only impact the outcome via change induced in the level of the intermediate variable (i.e., that this intermediate variable is the only causal pathway between randomization and the outcome.)

IV, therefore, relies on monotonicity and the exclusion restriction, but we could imagine circumstances where we would want to relax one or both of these assumptions.

Generalizing beyond compliance

- We can use the same framework and process for articulating a variety of types of research questions:
  - Naturally occurring variation in control setting (ex. Impact of ECHS by quality of high school otherwise attended)
  - "Truncation-due-to-death" (ex. Quality of life indicator only observable for subjects who remain alive)
  - Surrogate outcome (ex. Long run impacts of career academy high school opportunity on earnings realized primarily by those who experienced a significant rise in labor market exposure during high school)
- Process to formulate and set up the estimation problem is similar in each case

Setting up principal strata & defining treatment effects of interest

- To utilize the principal stratification framework, ask:

1. What is the treatment that was randomized?
2. What is the intermediate behavior of interest?
3. What are the strata?
4. What is / are the outcome(s) of interest?
5. In which strata is the treatment effect well defined?
6. What are justifiable assumptions that can reasonably be applied to:
    a. reduce the number of strata, and/or
    b. constrain the treatment effects within certain strata?
        in order to make the problem more tractable?

Agenda

- Conceptual framework & key assumptions
- Break
- Small group exercise
  - Principal stratification set up
  - Defining estimands of interest
- Lunch
- Estimation & bounds
- Break
- Wrap up

Agenda

- Conceptual framework & key assumptions
- Break
- Small group exercise
  - Principal stratification set up
  - Defining estimands of interest
- Lunch
- Estimation & bounds
- Break
- Wrap up

Application 1: Truncation-by-Death

- Outcome of interest is not defined for all sample members
- Not censoring: there is no underlying value to the outcome
- Examples:
  - Causal effect of post-stroke procedure on a quality-of-life metric can only be assessed for those who would remain alive with or without procedure
  - Lee (2009): In the study of a given job training program, is the positive effect on total earnings a function of increased employment (extensive margin) or an increased wage rate (intensive margin) or both? Any improvement in wage rate can only be assessed for those who would be employed with or without the intervention.

Application 1: Truncation-by-Death

- **Research Question:** In the study of a given job training program, is the positive effect on total earnings a function of increased employment or an increased wage rate or both?

Treatment	→ Offer of job training program
Outcome(s)	→ Hourly wage
Key intermediate behavior	→ Employment

Lee (2009): "Wages are only observed for individuals who are employed. Thus, even if there is random assignment of the 'treatment' of a training program, there may not only be an effect on wages but also on the probability that a person's wage will even be observed. Even a randomized experiment cannot guarantee that treatment and control individuals will be comparable *conditional on being employed*."

Application 1: Truncation-by-Death

- Defining all possible principal strata

		Job training offer = 0	
		Employed	**Not employed**
Job training offer = 1	**Employed**		
	Not employed		

Application 1: Truncation-by-Death

- Defining all possible principal strata

		Job training offer = 0	
		Employed	**Not employed**
Job training offer = 1	**Employed**	Always employed	Employed if treated
	Not employed	Employed if not treated	Never employed

Application 1: Truncation-by-Death

- What assumptions can be applied to reduce the number of strata?

		Job training offer = 0	
		Employed	**Not employed**
Job training offer = 1	**Employed**	Always employed	Employed if treated
	Not employed	Employed if not treated	Never employed

Is it a reasonable assumption that individuals are more likely to be employed under treatment than under control? Or, might the training raise the lowest wage certain workers are willing to accept, thereby potentially lowering employment for some if it is not possible to attain a job with this higher acceptable wage?

Application 1: Truncation-by-Death

- In which stratum / strata is the treatment effect well defined?

		Job training offer = 0	
		Employed	**Not employed**
Job training offer = 1	**Employed**	Always employed	Employed if treated
	Not employed	Employed if not treated	Never employed

In these strata, the treatment effect on hourly wages is not defined, because hourly wage is not defined for those who are not employed. In the *employed if treated* stratum, we observe hourly wages only under treatment and in the *never employed* stratum, we observe wages under neither treatment nor control.

Application 1: Truncation-by-Death

- In which stratum / strata is the treatment effect well defined?

		Job training offer = 0	
		Employed	**Not employed**
Job training offer = 1	**Employed**	Always employed	Employed if treated
	Not employed	Employed if not treated	Never employed

Individuals in this stratum are employed regardless of the intervention, so for this stratum only can we estimate the causal effect of the intervention on hourly wages. In the language of truncation-by-death, the treatment effect in this stratum is referred to as the Survivor Average Causal Effect (SACE).

Small group exercise

Role of assumptions

- Common assumptions
  - Exclusion restriction(s) – constraining outcomes
  - Monotonicity – removing boxes
  - Irrelevant alternatives – removing boxes

- Key take-away regarding assumptions:
  - Critical part of analytic set up
  - The more groups we can eliminate, the better / more precise our impact estimates will be
  - Assumptions are unverifiable, so they must be substantively justified

Questions / Discussion / Wrap-up

Agenda

- Conceptual framework & key assumptions
- Break
- Small group exercise
  - Principal stratification set up
  - Defining estimands of interest
- Lunch
- Estimation & bounds
- Break
- Wrap up