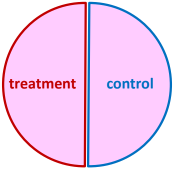
Sensitivity analysis for an unobserved effect modifier in RCT-to-target-population generalization of treatment effect

Trang Quynh Nguyen (tnguye28@jhu.edu)
(joint work with Elizabeth Stuart, Cyrus Ebnesajjad & Stephen Cole)

presentation at LEGACY, August 13, 2015

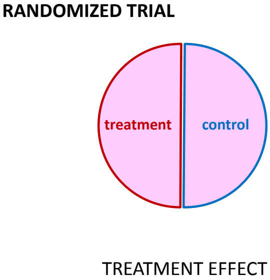
Main idea: have an RCT...

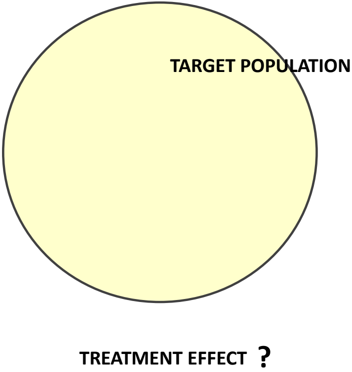
RANDOMIZED TRIAL



TREATMENT EFFECT

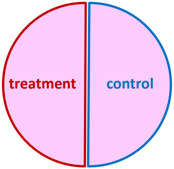
...but interested in a target population...





...but there are differentialy distributed effect modifiers...

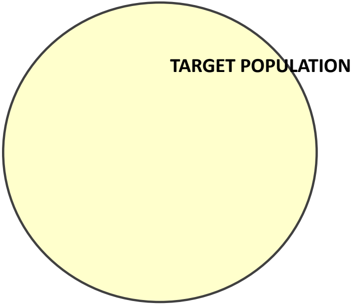
RANDOMIZED TRIAL



TREATMENT EFFECT

70% male, 40% female

50% college educated

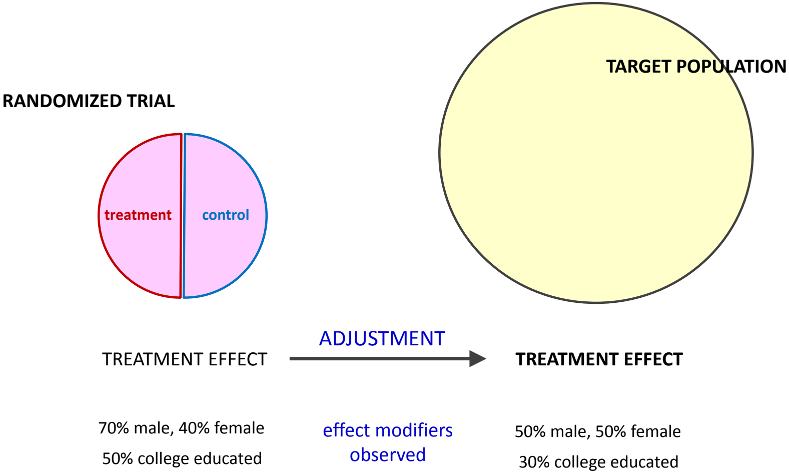


TREATMENT EFFECT ?

50% male, 50% female

30% college educated

well, if observe them, adjust for them



RANDOMIZED TRIAL

treatment

control

TARGET POPULATION

TREATMENT EFFECT

ADJUSTMENT

TREATMENT EFFECT

70% male, 40% female

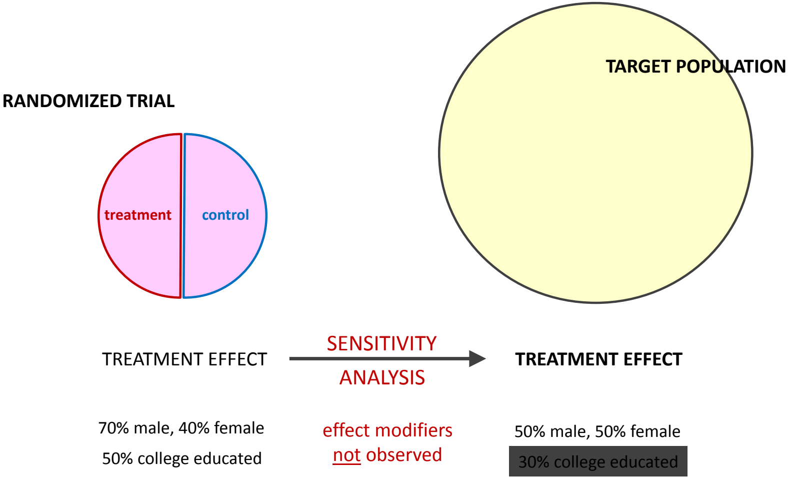
50% college educated

effect modifiers observed

50% male, 50% female

30% college educated

if don’t observe them, conduct sensitivity analyses



RANDOMIZED TRIAL

TARGET POPULATION

TREATMENT EFFECT

SENSITIVITY

ANALYSIS

TREATMENT EFFECT

70% male, 40% female

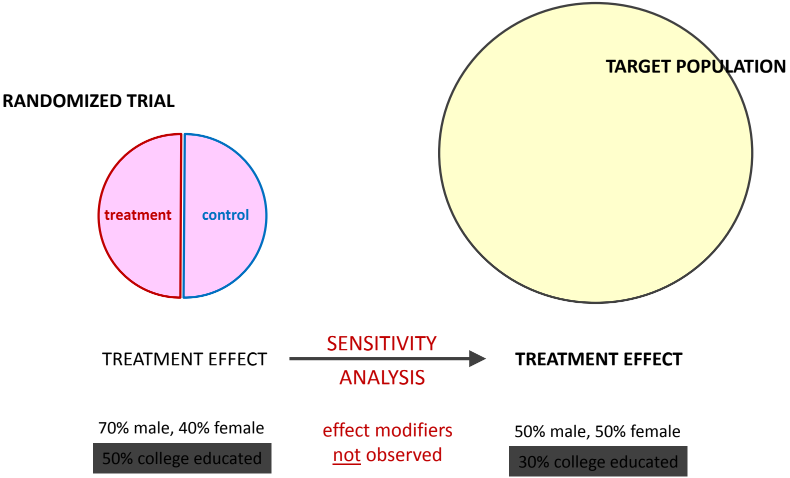
50% college educated

effect modifiers
not observed

50% male, 50% female

30% college educated

if don't observe them, conduct sensitivity analyses



RANDOMIZED TRIAL

treatment

control

TARGET POPULATION

TREATMENT EFFECT

SENSITIVITY

ANALYSIS

TREATMENT EFFECT

70% male, 40% female

50% college educated

effect modifiers

not observed

50% male, 50% female

30% college educated

Notation

T: treatment (0,1), randomized in the RCT
Y: outcome
Y^t: potential outcome under treatment $t, t = 0, 1$

Two datasets: RCT and a dataset representing the population

S: sample membership (1=study/RCT, 0=target population)

Two average treatment effects (ATEs):

$$\begin{array}{rl} \text{Study/RCT ATE:} & \text{SATE} = \text{E}[Y^1 - Y^0 | S = 1] \\ \text{Target population ATE:} & \text{TATE} = \text{E}[Y^1 - Y^0 | S = 0] \end{array}$$

Notation, cont'd

X: non-effect-modifying covariates

Z: effect modifiers, observed in both samples

U: effect modifier, observed in the RCT but not in the target population

V: effect modifier, not observed in both samples

X, Z, U, V may be associated with S.

1. All effect modifiers observed in both samples: the case with only Z

Assume the following model for the potential outcomes

$$E[Y_i^t] = \beta_0 + \beta_T t + \beta_X X_i + \beta_Z Z_i + \beta_{ZT} Z_i t.$$

$$\text{SATE} = \beta_T + \beta_{ZT} E[Z|S = 1]$$

$$\text{TATE} = \beta_T + \beta_{ZT} E[Z|S = 0]$$

assmptn: model holds in target population, no undue extrapolation

1. Only Z, cont'd

Option 1: Assess Δ, the difference between SATE and TATE:

$$\widehat{\Delta} = \hat{\beta}_{ZT}\{\hat{\mathrm{E}}[Z|S=1] - \hat{\mathrm{E}}[Z|S=0]\},$$

and get an adjusted point estimate of TATE:

$$\widehat{\mathrm{TATE}} = \widehat{\mathrm{SATE}} - \widehat{\Delta}.$$

1. Only Z, cont'd

Option 2: weighting-based TATE estimation

1. stack the two samples; fit a model regressing sample membership S on effect modifiers Z
2. predict odds of being in the target population sample, $W_i = \frac{\text{P}(S=0|Z_i)}{\text{P}(S=1|Z_i)}$, and reweight the RCT sample using W_i
   - the weighted RCT sample resembles the target population sample with respect to Z!
3. use the weighted RCT sample to estimate TATE

assmptn: positivity

Cole, S. R., & Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. American Journal of Epidemiology, 172(1), 107-15. doi:10.1093/aje/kwq084
Kern, H. L., Stuart, E. A., Hill, J. L., & Green, D. P. (In Press). Assessing methods for generalizing experimental impact estimates to target populations. Journal of Research on Educational Effectiveness.

Toy example: A smoking reduction intervention

	RCT sample			Target population
OBSERVED DATA:	Treatment	Control	Full	sample
	(n=200)	(n=200)	sample	(n=10,000)
Covariates				
Years of education: mean (SD)	12.06 (1.64)	12.11 (1.58)	12.08 (1.61)	11.02 (1.52)
Gender: percent female	49.50	50.50	50.00	19.86
Years smoked: mean (SD)	7.36 (2.57)	7.50 (2.45)	7.43 (2.51)	7.98 (2.72)
Outcome				
Cigarettes per week: mean (SD)	97.42 (6.00)	101.80 (5.29)	99.61 (6.06)	

Models fit to the RCT sample:

$$\widehat{\texttt{smoke}} = 120.31 - 2.02(\texttt{edu}) - 4.36(\texttt{female}) + 1.09(\texttt{smkyrs}) - 4.39(\texttt{treat})$$

$$\begin{aligned}\widehat{\texttt{smoke}} = 120.81 &- 2.03(\texttt{edu}) - 2.74(\texttt{female}) + 0.93(\texttt{smkyrs}) - 5.11(\texttt{treat}) \\ &- 3.27(\texttt{female} * \texttt{treat}) + 0.32(\texttt{smkyrs} * \texttt{treat}).\end{aligned}$$

$\widehat{\text{SATE}} = -4.39$, 95% CI=$(-5.05, -3.73)$

Formula-based adjustment: $\widehat{\text{TATE}} = -3.23$

Weighting-based estimation: $\widehat{\text{TATE}} = -3.36$, 95% CI=$(-4.11, -2.60)$

2. An effect modifier observed in RCT but not in target population: the case with U and Z

Assume the following potential outcomes model:

$$E[Y_i^t] = \beta_0 + \beta_T t + \beta_X X_i + \beta_Z Z_i + \beta_{ZT} Z_i t + \beta_U U_i + \beta_{UT} U_i t.$$

$$\text{SATE} = \beta_T + \beta_{ZT}\text{E}[Z|S=1] + \beta_{UT}\text{E}[U|S=1]$$

$$\text{TATE} = \beta_T + \beta_{ZT}\text{E}[Z|S=0] + \beta_{UT}\text{E}[U|S=0]$$

assmptn: no three-way TZU interaction

2. U and Z, cont'd

Option 1: Bias-formula-based sensitivity analysis

$$\begin{aligned}\mathsf{SATE} - \mathsf{TATE} = &\beta_{ZT}\{\mathsf{E}[Z|S=1] - \mathsf{E}[Z|S=0]\} + \\ &\beta_{UT}\{\mathsf{E}[U|S=1] - \mathsf{E}[U|S=0]\}.\end{aligned}$$

$$\begin{aligned}\widehat{\mathsf{TATE}} = \widehat{\mathsf{SATE}} &- \hat{\beta}_{ZT}\{\hat{\mathsf{E}}[Z|S=1] - \hat{\mathsf{E}}[Z|S=0]\} \\ &- \hat{\beta}_{UT}\{\hat{\mathsf{E}}[U|S=1] - \mathsf{E}[U|S=0]\}.\end{aligned}$$

$\Longrightarrow$ Specify a plausible range for $\mathsf{E}[U|S=0]$, and get a range for the point estimate of TATE.

2. U and Z, cont'd

Option 2: Weighting-based sensitivity analysis

We wish to weight the RCT sample by the odds of being in the target population given U and Z, but these odds are unknown.

But

$$W_i = \frac{\mathrm{P}(S=0|Z_i,U_i)}{\mathrm{P}(S=1|Z_i,U_i)} = \frac{\mathrm{P}(S=0|Z_i)}{\mathrm{P}(S=1|Z_i)} \cdot \frac{\mathrm{P}(U=U_i|S=0,Z_i)}{\mathrm{P}(U=U_i|S=1,Z_i)}.$$

$\Longrightarrow$ Estimate the distribution of U given Z in the RCT sample, and specify a plausible range for the distribution of U given Z in the target population. For each instance of this distribution, construct W_i, reweight the RCT sample and estimate TATE.

2. U and Z, cont'd

Option 3: Hybrid method (from-SATE-to-zATE-to-TATE)

1. Weight the RCT sample using the weights $W_i^{|Z} = \frac{P(S=0|Z_i)}{P(S=1|Z_i)}$, and use it to estimate a Z-adjusted ATE (zATE)
2. Conduct sensitivity analysis on U using the formula

$$\widehat{\text{TATE}} = \widehat{\text{zATE}} - \hat{\beta}_{UT}\{\hat{\mathrm{E}}[U|S=1, W^{|Z}] - \mathrm{E}[U|S=0]\}$$

where $\hat{\mathrm{E}}[U|S=1, W^{|Z}]$ is the weighted RCT mean U and $\mathrm{E}[U|S=0]$ is the unknown target population mean U

Toy example, cont'd

Now we do not observe # years smoked in the target population.

OBSERVED DATA:	RCT sample Treatment (n=200)	 Control (n=200)	 Full sample	Target population sample (n=10,000)
Covariates				
Years of education: mean (SD)	12.06 (1.64)	12.11 (1.58)	12.08 (1.61)	11.02 (1.52)
Gender: percent female	49.50	50.50	50.00	19.86
Years smoked: mean (SD)	7.36 (2.57)	7.50 (2.45)	7.43 (2.51)	
Outcome				
Cigarettes per week: mean (SD)	97.42 (6.00)	101.80 (5.29)	99.61 (6.06)	

Models fit to the RCT sample:

$$\widehat{\texttt{smoke}} = 120.31 - 2.02(\texttt{edu}) - 4.36(\texttt{female}) + 1.09(\texttt{smkyrs}) - 4.39(\texttt{treat})$$

$$\widehat{\texttt{smoke}} = 120.81 - 2.03(\texttt{edu}) - 2.74(\texttt{female}) + 0.93(\texttt{smkyrs}) - 5.11(\texttt{treat})$$
$$-3.27(\texttt{female} * \texttt{treat}) + 0.32(\texttt{smkyrs} * \texttt{treat}).$$

$\widehat{\text{SATE}} = -4.39$, 95% CI$=(-5.05, -3.73)$

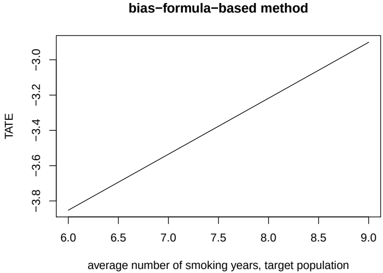
Toy example, cont'd

Bias-formula-based and hybrid method sensitivity analyses are straightforward. We use a range of 6-to-9 years for the mean number of years smoked in the target population.

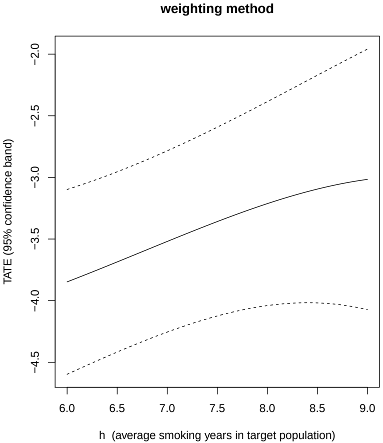
With weighting-based sensitivity analyses, for the variable number of years smoked (U),

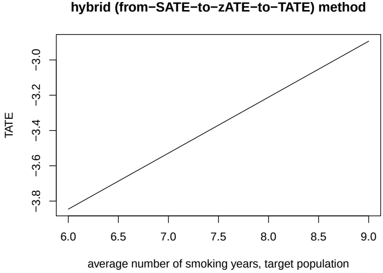
- with the RCT sample, informed by data, we assume and estimate a normal distribution conditional on gender;
- for the target population, we specify a normal distribution not conditional on gender, with a moving mean as the sensitivity parameter.

Toy example, cont'd


bias−formula−based method

three sensitivity analyses


weighting method


hybrid (from−SATE−to−zATE−to−TATE) method

3. When concerned about an unobserved (or unknown) effect modifier: the case with V and Z

We consider a generic V that is independent of X, Z,

and assume the same potential outcomes model:

$$E[Y_i^t] = \beta_0 + \beta_T t + \beta_X X_i + \beta_Z Z_i + \beta_{ZT} Z_i t + \beta_V V_i + \beta_{VT} V_i t.$$

$$\begin{aligned} \text{SATE} &= \beta_T + \beta_{ZT}\text{E}[Z|S=1] + \beta_{VT}\text{E}[V|S=1] \\ \text{TATE} &= \beta_T + \beta_{ZT}\text{E}[Z|S=0] + \beta_{VT}\text{E}[V|S=0] \end{aligned}$$

3. V and Z, cont'd

Modified option 1: Bias-formula-based sensitivity analysis

$$\begin{aligned}\widehat{\text{TATE}} = \widehat{\text{SATE}} &- \hat{\beta}_{ZT}\{\hat{\text{E}}[Z|S=1] - \hat{\text{E}}[Z|S=0]\} \\ &- \beta_{VT}\{\text{E}[V|S=1] - \text{E}[V|S=0]\}.\end{aligned}$$

Because V is independent of Z, β_{ZT} can be estimated without bias via a regression model that includes X, Z, ZT and leave out V.

$\Longrightarrow$ Specifiy a range for the degree of effect modification by V (β_{VT}) and a range for the difference in mean/prevalance between the RCT and target population ($\text{E}[V|S=1] - \text{E}[V|S=0]$), and get a surface for the point estimate of TATE.

3. V and Z, cont'd

Modified option 3: Hybrid (from-SATE-to-xzATE-to-TATE) method

1. Weight the RCT sample using $W_i^{|X,Z} = \frac{\mathrm{P}(S=0|X_i,Z_i)}{\mathrm{P}(S=1|X_i,Z_i)}$, and use it to estimate an X-and-Z-adjusted ATE (xzATE)
2. Conduct sensitivity analysis on V using the formula

$$\widehat{\text{TATE}} = \widehat{\text{xzATE}} - \beta_{VT}\{\mathrm{E}[V|S=1] - \mathrm{E}[V|S=0]\}$$

by specifying two ranges for $\mathrm{E}[V|S=1] - \mathrm{E}[V|S=0]$ and β_{VT}, and getting one surface for TATE point estimates plus two surfaces for confidence limits.

Toy example, cont'd

Now for covariates, we only observe gender and education. We are concerned about unobserved effect modifiers.

	RCT sample			Target population
OBSERVED DATA:	Treatment	Control	Full	sample
	(n=200)	(n=200)	sample	(n=10,000)
Covariates				
Years of education: mean (SD)	12.06 (1.64)	12.11 (1.58)	12.08 (1.61)	11.02 (1.52)
Gender: percent female	49.50	50.50	50.00	19.86
Outcome				
Cigarettes per week: mean (SD)	97.42 (6.00)	101.80 (5.29)	99.61 (6.06)	

Models fit to the RCT sample:

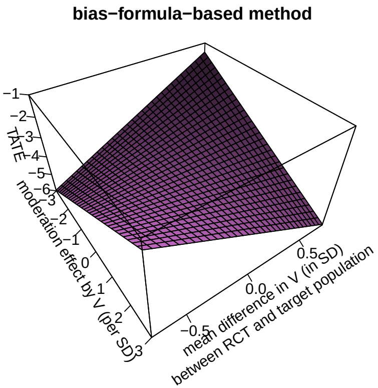
$$\widehat{\texttt{smoke}} = xxx - xxx(\texttt{edu}) - xxx(\texttt{female}) - 4.53(\texttt{treat})$$

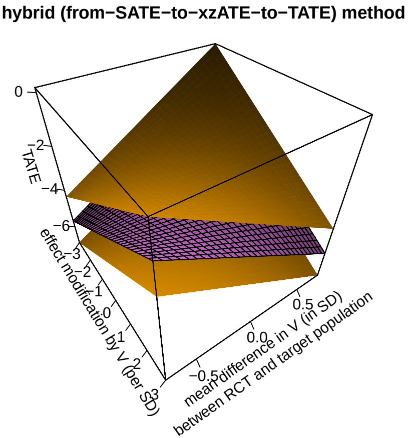
$$\widehat{\texttt{smoke}} = 127.50 - 2.04(\texttt{edu}) - 1.98(\texttt{female}) - 3.16(\texttt{treat}) - 2.74(\texttt{female} * \texttt{treat}).$$

$$\widehat{\text{SATE}} = -4.53, \text{ 95\% CI}=(-5.37, -3.69)$$

Toy example, cont'd

two sensitivity analyses





Summary of sensitivity analysis methods

	for a specific U observed in the RCT but not in the target population	for a generic V not observed in either sample independent of X, Z
1.	bias-formula-based method	bias-formula-based method
2.	weighting-based method	
3.	hybrid method (via zATE)	hybrid method (via xzATE)

Two real data examples

- effect of an anti-retroviral regimen on CD4 count
- effect of a job training intervention on earnings

Things to address/consider: Your inputs appreciated!

- note the difference from the RCT sample is a subset of the target population sample
- weighting adjustment for Z only or for X, Z
- potential applications
- future directions

Thank you!