On Theory and Identification:
When and Why We Need Theory for Causal Identification

Tara Slough*

May 6, 2021

Contents

*Assistant Professor, New York University. taraslough@nyu.edu

A1 Formal Exposition of Truncation by Death Problem

Consider the following graphical model:

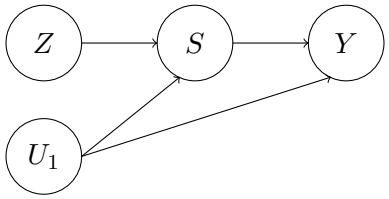

Figure A1: A graphical model depicting a causal process consistent with endogenous units of analysis if the potential outcome $Y(Z, S)$ is undefined for some units on the basis of the revelation of some $S(Z)$.

Treatment $Z_i \in \{0, 1\}$, is assigned such that the probability of assignment to treatment $Z = 1$ is $p \in (0, 1)$ for all units. A first outcome, $S(Z) \in \{0, 1\}$ indicates whether the a subject "survives." The dependent variable of interest $Y(S, Z)$ occurs subsequent to the realization of $S(Z)$. Define four causal types (principal strata): always survivors, if treated survivors, if untreated survivors, and never survivors. Table A1 defines these types, their shares in the population, and relevant potential outcomes.

Stratum	Weight	$S(Z=1)$	$S(Z=0)$	$\bar{Y}(S=1, Z=1)$	$\bar{Y}(S=1, Z=0)$	$\bar{Y}(S=0, Z=1)$	$\bar{Y}(S=0, Z=0)$
Always survivor	π_A	1	1	$\bar{Y}_A(1,1)$	$\bar{Y}_A(1,0)$	-	-
If **T**reated survivor	π_T	1	0	$\bar{Y}_T(1,1)$	-	-	$[\bar{Y}_T(0,0)]$
If **U**ntreated survivor	π_U	0	1	-	$\bar{Y}_U(1,0)$	$[\bar{Y}_U(0,1)]$	-
Never survivor	π_N	0	0	-	-	$[\bar{Y}_N(0,1)]$	$[\bar{Y}_N(0,0)]$

Table A1: Principal strata of an experiment with a binary treatment and binary survival variable. Elements in brackets indicate that a potential outcome is undefined. If defined, the outcome $Y(S, Z) \in \mathbb{R}^1$ and the last four columns indicate cell means.

Given the binary assignment to treatment and the binary survival variable, the ATE of Z on Y could ideally be written:

$$E[Y(Z=1)] - E[Y(Z=0)] = \begin{array}{l} \pi_A\bar{Y}_A(1,1) + \pi_T\bar{Y}_T(1,1) + \pi_U\underline{\bar{Y}_U(0,1)} + \pi_N\underline{\bar{Y}_N(0,1)} - \\ (\pi_A\bar{Y}_A(1,0) + \pi_T\underline{\bar{Y}_T(0,0)} + \pi_U\bar{Y}_U(1,0) + \pi_N\underline{\bar{Y}_N(0,0)}) \end{array} \quad (1)$$

However, because some of these quantities (underlined in red) are undefined, the expression (and hence the ATE) is undefined.

A2 Illustrative Examples in Social Science

This section expands upon the examples of truncation by death in the social sciences cited in Table 1. Some of these selection problems are discussed in the respective literature.

A3 Equilibrium Characterization, Proofs from Stylized Models

A3.1 Case #1: Always Crime

In this decision theoretic model, I assume that a crime occurred with probability 1. The bystander reports if the expected utility from reporting $E[U_B(r)]$ exceeds the expected utility from not reporting $E[U_B(\neg r)]$:

Table A2: Elaboration of truncation by death in literatures in Table 1

	Example	Truncation by death analogue	Mapping to formal framework
1	Effects of conflict	Individuals (militants or civilians) perish in conflict. This is quite literally truncation by death as in the medical setting.	The composition of actors, a, changes between treatment (exposure to conflict) and outcome measurement. The potential outcomes measuring attitudes and behaviors are undefined for deceased subjects. Further, if conflict is deadly, the composition of a is arguably endogenous to the treatment.
2	Downstream effects of shocks on political behavior.	There are various possible shocks. In the cited example, **?**, recipients of a 19th century land lottery in Georgia had, on average, more children than non-winners of the land lottery. (Alternatively, recipients of the land lottery had more surviving children.)	The composition of actors, a, changes between treatment (exposure to the wealth shock) and outcome measurement through differential rates of procreation. The potential outcomes measuring the behavior of children who would have been born (or survived) if their parents were treated are undefined.
3	Long-run effects of historical institutions.	Communities cease to exist or form after the imposition of (pre)-colonial institutions. While the prevalence of these issues are somewhat unclear, discussions of the inexact mapping between historic and current communities generally suggests these issues.	When communities are actors, the composition of actors, a, changes between treatment (imposition of [pre]-colonial) institutions and outcome measurement. The potential outcomes measuring community-level behavioral outcomes in communities that ceased to survive are undefined.
4	Email audit experiments	Subjects choose whether or not to respond to an email.	Suppose the subject can provide an accurate or inaccurate response to the query. Subsequent to a decision to respond to the email, the response quality strategy set is: $S_a = \{\text{accurate}, \text{inaccurate}\}$. Subsequent to a decision not to respond, $S_a = \emptyset$. As such the potential outcomes measuring the quality of information provided is undefined if the subject chooses not to respond.
5	Ideological positioning across elections as a function of electoral performance at time t.	Between elections t and $t+1$, a party disbands. One could view this as a change in the set of parties (actors) at election $t+1$ or a change in the strategies available to parties that are contesting elections vs. disbanded.	A party that chooses to disband does not choose a platform in election $t+1$. As such, the potential outcomes measuring platform choice in $t+1$ (or functions thereof) are undefined.
6	Incumbency (dis)advantage	Between election t and election $t+1$ a candidate chooses not to seek re-election. As such, voters in $t+1$ do not have the choice to vote for a candidate who is not contesting office.	The voter's strategy set – measuring the options on the ballot– is different if the incumbent and challenger contest office in $t+1$ than if they do not. If the incumbent does not run, for example, the potential outcome measuring voter's decision to re-elect the incumbent is undefined.
7	Police use of force	A police officer chooses not to stop a civilian. The decision to use force (subsequent to a stop) depends on whether or not a civilian was stopped or not.	The strategies available to an officer, S_a, are different subsequent to stopping a civilian versus not stopping a civilian. As such the potential outcome measuring police use of force is undefined if the citizen is not stopped.

$$E[U_B(r)] \geq E[U_B(\neg r)] \Rightarrow \iota_R \psi - c_R \geq \iota_N \psi$$

Solving for ψ, the citizen will report if:

$$\psi \geq \frac{c_R}{\iota_R - \iota_N} \qquad \blacksquare$$

Given $F_\psi(\cdot)$, the cdf from which ψ is drawn, the proportion of citizens that report a crime is $1 - F_\psi\left(\frac{c_R}{\iota_R - \iota_N}\right)$. With this rate of reporting, the ATE on reporting can be written:

$$\begin{aligned} ATE^1_{\mathcal{R}} &= 1 - F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N}\right) - \left(1 - F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N}\right)\right) \\ &= F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N}\right) - F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N}\right) > 0 \end{aligned}$$

This quantity is positive because $c_R^{Z=1} < c_R^{Z=0}$. Further, the ATE on incidence in the administrative record is:

$$\begin{aligned} ATE^1_{\mathcal{V}} &= \iota_R \underbrace{\left[1 - F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N}\right)\right]}_{\text{Reporting rate}} + \iota_N \underbrace{\left[F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N}\right)\right]}_{\text{Non-reporting rate}} - \iota_R \underbrace{\left[1 - F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N}\right)\right]}_{\text{Reporting rate}} - \iota_N \underbrace{\left[F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N}\right)\right]}_{\text{Non-reporting rate}} \\ &= (\iota_R - \iota_N)\left[F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N}\right) - F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N}\right)\right] > 0 \end{aligned}$$

This quantity is positive because $c_R^{Z=1} < c_R^{Z=0}$ and $\iota_R > \iota_N$. Because crime always occurs (there is no selection), the ATE is equivalent to the SACE in both cases. $\blacksquare$

A3.2 Case #2: Exogenous Crime and Exogenous Investigation

This model directly follows from Case #1. However, in the $1 - \rho$ proportion of cases (precincts) in which there is no crime perpetrated, the reporting outcome is undefined. As such, $ATE_{\mathcal{R}}$ and $ATE_{\mathcal{V}}$ are undefined. In the ρ proportion of cases in which there is crime, the SACE follows from the calculation of the ATE from Model A3.1. Thus:

$$\begin{aligned} SACE_{\mathcal{R}} &= F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N}\right) - F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N}\right) > 0 \\ SACE_{\mathcal{V}} &= (\iota_R - \iota_N)\left[F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N}\right) - F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N}\right)\right] > 0 \end{aligned}$$

Now consider the quantities estimated by a difference-in-means, $\Delta_{\mathcal{R}}$ and $\Delta_{\mathcal{N}}$:

$$\begin{aligned} \Delta_{\mathcal{R}} &= \rho\left(1 - F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N}\right)\right) - \rho\left(1 - F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N}\right)\right) \\ &= \rho SACE_{\mathcal{R}} \\ \Delta_{\mathcal{V}} &= (\iota_R - \iota_N)\left[\rho F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N}\right) - \rho F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N}\right)\right] \\ &= \rho SACE_{\mathcal{V}} \end{aligned}$$

A naive comparison of treatment and control beats will yield the quantities $\rho SACE_{\mathcal{R}}$ and $\rho SACE_{\mathcal{V}}$, respectively. Both quantities are posititve. ■

A3.3 Case #3: Endogenous Crime and Exogenous Investigation

I characterize a subgame perfect equilibrium in pure strategies by backward induction. As such, I begin with the citizen's decision whether or not to report a crime in the subgame in which a crime has occurred. This is equivalent to the citizen's calculation in subsection A3.1. The citizen reports if and only if:

$$\psi \geq \frac{c_R}{\iota_R - \iota_N}$$

Now consider the suspect's choice. He will commit a crime if the expected utility from reporting $E[U_S(v)]$ exceeds the expected utility from not reporting $E[U_S(\neg v)]$:

$$\begin{aligned} \lambda - p\left[\iota_R\left[1 - F_\psi\left(\frac{c_R}{\iota_R - \iota_N}\right)\right] + \iota_N F_\psi\left(\frac{c_R}{\iota_R - \iota_N}\right)\right] &\geq 0 \\ p\left[\iota_R + (\iota_N - \iota_R)F_\psi\left(\frac{c_R}{\iota_R - \iota_N}\right)\right] &\leq \lambda \end{aligned}$$

In the unique subgame perfect equilibrium, thus, the suspect commits a crime if:

$$\lambda \geq p\left[\iota_R + (\iota_N - \iota_R)F_\psi\left(\frac{c_R}{\iota_R - \iota_N}\right)\right]$$

Upon observing the crime, the bystander reports if $\psi \geq \frac{c_R}{\iota_R - \iota_N}$. ■

As in the previous case, the ATEs are undefined because some crimes do not occur. To compute the SACE, first it is useful to define two thresholds of λ which define crime occurrence under treatment and control:

$$\begin{aligned} \tilde{\lambda} &= p\left[\iota_R + (\iota_N - \iota_R)F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N}\right)\right] \\ \underset{\sim}{\lambda} &= p\left[\iota_R + (\iota_N - \iota_R)F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N}\right)\right] \end{aligned}$$

Because $c_R^{Z=0} > c_R^{Z=1}$, $\tilde{\lambda} > \underset{\sim}{\lambda}$. This implies that any crime that would occur if a unit is treated would occur if the unit is untreated. The "always survivor" stratum is thus defined by any suspect for whom $\lambda > \tilde{\lambda}$. The SACEs are thus given by:

$$\begin{aligned} SACE_{\mathcal{R}} &= F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N} \mid \lambda > \tilde{\lambda}\right) - F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N} \mid \lambda > \tilde{\lambda}\right) > 0 \\ SACE_{\mathcal{V}} &= (\iota_R - \iota_N)\left[F_\psi\left(\frac{c_R^{Z=0}}{\iota_R - \iota_N} \mid \lambda > \tilde{\lambda}\right) - F_\psi\left(\frac{c_R^{Z=1}}{\iota_R - \iota_N} \mid \lambda > \tilde{\lambda}\right)\right] > 0 \end{aligned}$$

A difference-in-means estimator estimates:

$$\begin{aligned}
\Delta_{\mathcal{R}} =& F_\lambda(\tilde{\lambda}) \left[F_\psi \left(\frac{c_R^{Z=0}}{\iota_R - \iota_N} \mid \lambda > \tilde{\lambda} \right) - F_\psi \left(\frac{c_R^{Z=1}}{\iota_R - \iota_N} \mid \lambda > \tilde{\lambda} \right) \right] - \\
& (F_\lambda(\tilde{\lambda}) - F_\lambda(\underset{\sim}{\lambda})) F_\psi \left(\frac{c_R^{Z=0}}{\iota_R - \iota_N} \mid \lambda \in (\underset{\sim}{\lambda}, \tilde{\lambda}] \right) \\
=& SACE_{\mathcal{R}} - (F_\lambda(\tilde{\lambda}) - F_\lambda(\underset{\sim}{\lambda})) F_\psi \left(\frac{c_R^{Z=0}}{\iota_R - \iota_N} \mid \lambda \in (\underset{\sim}{\lambda}, \tilde{\lambda}] \right) \\
\Delta_{\mathcal{V}} =& F_\lambda(\tilde{\lambda}) \left[(\iota_R - \iota_N) \left[F_\psi \left(\frac{c_R^{Z=0}}{\iota_R - \iota_N} \mid \lambda > \tilde{\lambda} \right) - F_\psi \left(\frac{c_R^{Z=1}}{\iota_R - \iota_N} \mid \lambda > \tilde{\lambda} \right) \right] \right] - \\
& (F_\lambda(\tilde{\lambda}) - F_\lambda(\underset{\sim}{\lambda})) \left[(\iota_R - \iota_N) F_\psi \left(\frac{c_R^{Z=0}}{\iota_R - \iota_N} \mid \lambda \in (\underset{\sim}{\lambda}, \tilde{\lambda}] \right) \right] \\
=& SACE_{\mathcal{V}} - (F_\lambda(\tilde{\lambda}) - F_\lambda(\underset{\sim}{\lambda})) \left[(\iota_R - \iota_N) F_\psi \left(\frac{c_R^{Z=0}}{\iota_R - \iota_N} \mid \lambda \in (\underset{\sim}{\lambda}, \tilde{\lambda}] \right) \right]
\end{aligned}$$

The sign of the difference-in-means estimator is ambiguous for both outcomes. While both SACEs are positive, the second term in both expressions is negative. ■

A3.4 Case #4: Strategic Policing

I characterize a subgame perfect equilibrium in pure strategies by backward induction. As such, I begin with the officer's decision of whether or not to investigate a crime, conditional on whether the crime was reported:

$$\begin{aligned}
E[u_O(i|r)] &\geq E[u_O(\neg i|r)] \\
-\kappa &\geq -\alpha_r \\
\kappa &\leq \alpha_r
\end{aligned}
\qquad\qquad
\begin{aligned}
E[u_O(i|\neg r)] &\geq E[u_O(\neg i|\neg r)] \\
-\kappa &\geq -\alpha_{\neg r} \\
\kappa &\leq \alpha_{\neg r}
\end{aligned}$$

When the bystander evaluates the likelihood of reporting, the probability that a crime is investigated is thus given by $1 - F_\kappa(\alpha_r)$ (if reported) and $1 - F_\kappa(\alpha_{\neg r})$. Plugging these into the bystander's expected utility, the bystander reports if and only if:

$$\psi \geq \frac{c_R}{F_\kappa(\alpha_{\neg r}) - F_\kappa(\alpha_r)}$$

Now consider the suspect's choice. He will commit a crime if the expected utility from reporting $E[U_S(v)]$ exceeds the expected utility from not reporting $E[U_S(\neg v)]$. Denote the threshold above which a crime occurs as $\hat{\lambda}$.

$$\begin{aligned}
&\lambda - p \left[(1 - F_\kappa(\alpha_r)) \left[1 - F_\psi \left(\frac{c_R}{F_\kappa(\alpha_{\neg r}) - F_\kappa(\alpha_r)} \right) \right] + (1 - F_\kappa(\alpha_{\neg r})) F_\psi \left(\frac{c_R}{F_\kappa(\alpha_{\neg r}) - F_\kappa(\alpha_r)} \right) \right] \geq 0 \\
&\hat{\lambda} \geq p \left[1 - F_\kappa(\alpha_r) + (F_\kappa(\alpha_r) - F_\kappa(\alpha_{\neg r})) F_\psi \left(\frac{c_R}{F_\kappa(\alpha_{\neg r}) - F_\kappa(\alpha_r)} \right) \right]
\end{aligned}$$

In the unique subgame perfect equilibrium, thus, the suspect commits a crime iff $\lambda > \hat{\lambda}$. Upon observing the crime, the bystander reports if $\psi \geq \frac{c_R}{\iota_R - \iota_N}$; and upon recieving the report, the officer investigates if $\kappa \leq \alpha_r$

but upon not receiving the report, the officer investigates iff $\kappa \leq \alpha_{\neg r}$. ■

This case is identical to the previous case except that $\iota_R \equiv 1 - F_\kappa(\alpha_R)$ and $\iota_N \equiv 1 - F_\kappa(\alpha_{\neg R})$. Substituting these expressions and redefining $\tilde{\lambda}$ and $\underset{\sim}{\lambda}$ as:

$$\tilde{\lambda} = p\left[1 - F_\kappa(\alpha_r) + (F_\kappa(\alpha_r) - F_\kappa(\alpha_{\neg r}))\, F_\psi\left(\frac{c_R^{Z=1}}{F_\kappa(\alpha_{\neg r}) - F_\kappa(\alpha_r)}\right)\right]$$
$$\underset{\sim}{\lambda} = p\left[1 - F_\kappa(\alpha_r) + (F_\kappa(\alpha_r) - F_\kappa(\alpha_{\neg r}))\, F_\psi\left(\frac{c_R^{Z=0}}{F_\kappa(\alpha_{\neg r}) - F_\kappa(\alpha_r)}\right)\right]$$

the remark follows directly from the proof to Remark 3. ■

A4 Proof of Proposition 1

Suppose that an experiment manipulates a single treatment Z. Assume:

1. Treatment assignment is ignorable: $Y(Z) \perp Z, Pr(Z = z) \in (0, 1)$.
2. SUTVA: $Y_i(z_i) = Y_i(z_i, \mathbf{z}_{\neg i}) \forall i$.

Consider a dynamic model for which $h^\emptyset \neq h^T$. Index sets of non-terminal histories, $h \in H \backslash H^T$ by the cardinality of the set of past actions. In this notation, $h^\emptyset \equiv h^0$. The subsequent histories are represented by $h \in H^1$. etc. In this notation, a dynamic model implies that $\exists H^1$.

With this notation, strategy set symmetry, as defined in Definition 1, implies that for any $h \in H^j$, the actor and set of strategies available for all elements H^{j+1} are equivalent, for all $j \in \{0, 1, ..., T-1\}$.

Consider an action, a, in the strategy set of arbitrary node $h \in H$. Denote a variable measuring this action as $\mathcal{A}$. The ATE can be written:

$$\sum_{h \in H^j} Pr(h|Z = 1) E[\mathcal{A}|Z = 1, h = h] - \sum_{h \in H^j} Pr(h|Z = 0) E[\mathcal{A}|Z = 0, h = h] \tag{2}$$

First, consider the first post-treatment action, $j = 0$. Both expectations in Equation 2 are defined. The ATE is both defined and identifed given Assumptions 1 and 2 and standard arguments (i.e. **?** Equation 2.3 or **?** Section 2.2).
Now, consider some $j > 0$. Consider two cases:

1. If a is in the strategy set for all $h \in H^j$, the expression $E[\mathcal{A}|Z = z, h = h]$ is defined. The ATE is both defined and identified.
2. If a is *not* in the strategy set for any $h \in H^j$, the expression $E[\mathcal{A}|Z = z, h = h]$ is undefined for some h. The ATE is undefined, and thus unidentified.

By Definition 1, if a model is strategy set symmetric, it follows from the case of $j = 0$ and Case #1 above that the ATE is identified for all actions. Further, if the model is not strategy set symmetric, it follows from the case of $j = 0$ and Case #2 that the ATE must be identified for at least one outcome (at h^0) and must be unidentified for at least one outcome. ■