Causality and randomization

Maximilian Kasy

November 2, 2018

Introduction

- This talk is based on

  Kasy, M. (2016). Why experimenters might not always want to randomize, and what they could do instead. *Political Analysis*, 24(3):324–338.

- Causality is often defined by reference to Randomized Controlled Trials (RCTs).

- To what extent is randomization important? Are RCTs the best way to learn about causal effects?

Introduction

Some intuitions

1. We don't add random noise to estimators or tests
   – why add random noise to treatment assignments?
2. Identification requires controlled trials (CTs),
   but not randomized controlled trials (RCTs).
3. Goal of treatment assignment is to
   "compare apples with apples."
   $\Rightarrow$ Balance covariate distribution.
   (Not just balance of means!)

Introduction

Somewhat more formally

- Treatment assignment in an experiment is a decision problem.
- General result: For any decision problem, randomized procedures perform worse than deterministic procedures.
- More specific result:
    - Suppose the goal is to assign treatment to minimize the mean squared error of estimators of average treatment effects.
    - Then (non-random) assignments which make treatment and control groups as similar as possible (in terms of a well-defined metric) are optimal.
    - Random assignment generates unnecessary imbalances.

Roadmap

1. **Review of definitions**
2. Decision problems
3. Optimal treatment assignments
4. Arguments for randomization
5. Conclusion

Review of definitions

A made-up history of causality

1. Pure probability theory:
   - Does not allow to talk about causality,
   - only joint distributions.
2. Causality in the sciences ("Gallilei"):
   Controlled experiments.
   - Additional concept: **Exogenous variation**.
   - Do the same thing
     $\Rightarrow$ same thing happens to the outcomes you measure.
   - Variation in experimental circumstances
     $\Rightarrow$ difference in observed outcomes $\approx$ causal effect.

Review of definitions

A made-up history of causality, continued

3. Causality in econometrics, biostatistics,... ("Fisher"):
    - Additional concept: **Unobserved heterogeneity**
      $\Rightarrow$ Can never replicate experimental circumstances fully.
    - But we can still create experimental circumstances which are the same in expectation.
      $\Rightarrow$ Randomized experiments (or "quasi-experiments").
4. Most experiments in social science (and this talk):
    - Additional concept: **Observed heterogeneity**.
    - Random treatment assignment makes treatment and control group the same in expectation.
    - But they might randomly be very different ex-post.
    - We can do better: Make them similar in terms of observables!

Review of definitions

Identification

1. Learning about underlying structures, causal mechanisms
2. from a population distribution.
3. Example:
   Identify a causal effect
   by a difference in expectations
   if we have a randomized experiment.

- Identification inverts the mapping
- from underlying structures to a population distribution
- implied by a model and identifying assumptions.

Review of definitions

Structural objects

- Contested notion; my preferred definition:
- An object is structural, if it is **invariant** across relevant counterfactuals.
- Example: Dropping a ball from the tower of Pisa.
    - Acceleration is the same, no matter which floor you drop it from,
    - and also the same if you do this on the Eiffel tower.
    - Time to ground would not be the same,
    - and acceleration is not the same if you do this on the moon.

Review of definitions

Treatment effects and potential outcomes

- I will focus without loss of generality on two "treatments:" $D = 0$ or $D = 1$.
- Units i, potential outcomes Y_i^0 and Y_i^1, realized outcomes Y_i.
- Treatment effect for unit i: $Y_i^1 - Y_i^0$.
- Average treatment effect:

$$ATE = E[Y^1 - Y^0].$$

- Expectation averages over the population of interest.

Review of definitions

The fundamental problem of causal inference

- **We never observe both Y^0 and Y^1 at the same time**
- One of the potential outcomes is always missing from the data.
- Treatment D determines which of the two we observe.

$$Y = D \cdot Y^1 + (1 - D) \cdot Y^0.$$

- Selection problem: In general

$$E[Y|D=1] = E[Y^1|D=1] \neq E[Y^1],$$
$$E[Y|D=0] = E[Y^0|D=0] \neq E[Y^0],$$
$$E[Y|D=1] - E[Y|D=0] \neq E[Y^1 - Y^0] = ATE.$$

Review of definitions

Randomization

- No selection $\Leftrightarrow$ D is random

$$(Y^0, Y^1) \perp D.$$

- In this case, the ATE is **identified**.

$$E[Y|D=1] = E[Y^1|D=1] = E[Y^1]$$
$$E[Y|D=0] = E[Y^0|D=0] = E[Y^0]$$
$$E[Y|D=1] - E[Y|D=0] = E[Y^1 - Y^0] = ATE.$$

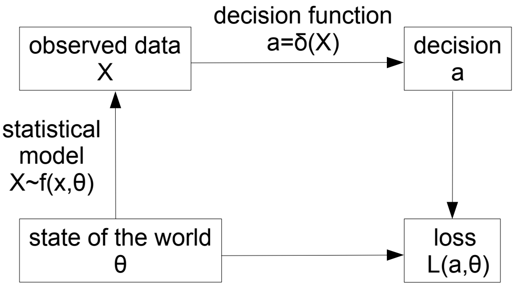
- Can ensure this by actually randomly assigning D
- Independence $\Rightarrow$ comparing treatment and control actually compares "apples with apples" (ex ante).
- This gives **empirical content to** the notion of **potential outcomes**!

Roadmap

1. Review of definitions
2. **Decision problems**
3. Optimal treatment assignments
4. Arguments for randomization
5. Conclusion

Decision problems

General setup



Decision problems

Notions of risk

- **Risk function:** Expected loss, averaging over sampling distribution, function of state of the world:

$$R(\delta, \theta) = E_\theta[L(\delta(X), \theta)].$$

- **Bayes risk:** Average of risk function over some prior distribution (i.e., decision weights):

$$R(\delta, \pi) = \int R(\delta, \theta)\pi(\theta)d\theta.$$

- **Worst case risk:** Maximum of risk function, over some set of θ, given $\delta(\cdot)$:

$$\overline{R}(\delta) = \sup_{\theta \in \Theta} R(\delta, \theta).$$

Decision problems

Randomized decision procedures

- We can allow δ to depend on some randomization device U: $a = \delta(X, U)$, where $P(U = u|\theta, X) = p_u$ for $u = 1, \ldots, k$.
- Denote δ^u the deterministic decision rule $a = \delta(X, u)$.
- It follows from the definitions that

$$\begin{aligned} R(\delta,\theta) &= p_1 \cdot R(\delta^1,\theta) + \ldots + p_k \cdot R(\delta^k,\theta), \\ R(\delta,\pi) &= p_1 \cdot R(\delta^1,\pi) + \ldots + p_k \cdot R(\delta^k,\pi) \\ \overline{R}(\delta) &= p_1 \cdot \overline{R}(\delta^1) + \ldots + p_k \cdot \overline{R}(\delta^k). \end{aligned}$$

  (Worst case risk is somewhat subtle – we will return.)
- Averages (over U) are not as good as best cases. Thus

$$R(\delta,\pi) \geq \min_u R(\delta^u,\pi),$$
$$\overline{R}(\delta) \geq \min_u \overline{R}(\delta^u).$$

- We just proved the following theorem.

Theorem (Optimality of deterministic decisions)

Consider a general decision problem.
Let $R^*(\cdot)$ *equal* $R(\cdot, \pi)$ *or* $\overline{R}(\cdot)$. *Then:*

1. *The optimal risk* $R^*(\delta^*)$, *when considering only deterministic procedures is no larger than the optimal risk when allowing for randomized procedures.*
2. *If the optimal deterministic procedure is unique, then it has strictly lower risk than any non-trivial randomized procedure.*

Roadmap

1. Review of definitions
2. Decision problems
3. **Optimal treatment assignments**
4. Arguments for randomization
5. Conclusion

Optimal treatment assignments

Setup

1. *Sampling:*
   Random sample of n units
   baseline survey $\Rightarrow$ vector of covariates X_i
2. *Treatment assignment:*
   binary treatment assigned by $D_i = d_i(\boldsymbol{X}, U)$
   $\boldsymbol{X}$ matrix of covariates; U randomization device
3. *Realization of outcomes:*
   $Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$
4. *Estimation:*
   estimator $\widehat{\beta}$ of the (conditional) average treatment effect,
   $\beta = \frac{1}{n} \sum_i E[Y_i^1 - Y_i^0 | X_i, \theta]$

- The theorem implies:
  The optimal $\boldsymbol{d}(\boldsymbol{X}, U)$ does not depend on U.
- But how do we get the optimal $\boldsymbol{d}$?

Optimal treatment assignments

Sketch of solution

- Key object: Conditional expectation of potential outcomes,

$$f(x,d) = E[Y^d | X = x].$$

- Bayesian approach: Prior distribution over $f(\cdot,\cdot)$.
  Possibly informed by earlier data.
- Estimator: E.g. difference in means,

$$\widehat{\beta} = \frac{1}{n_1} \sum_i D_i Y_i - \frac{1}{n_0} \sum_i (1 - D_i) Y_i.$$

- Loss: Squared estimation error,

$$(\widehat{\beta} - \beta)^2.$$

Optimal treatment assignments

Discrete optimization

- Risk $R(\boldsymbol{d},\beta|\boldsymbol{X})$: Expected loss, i.e. mean squared error.
- Straightforward to write down in closed form.
  Formalizes the notion of "balance."
- The optimal design solves

$$\max_{\mathbf{d}} R(\boldsymbol{d},\beta|\boldsymbol{X}).$$

- With continuous or many discrete covariates, the optimum is unique, and thus randomization is strictly dominated.
- Absent covariates, all units look the same. In this case, the optimum is not unique, and randomization does not hurt.
- Possible optimization algorithms:
  1. Search over random $\mathbf{d}$,
  2. greedy algorithm,
  3. simulated annealing.

Roadmap

1. Review of definitions
2. Decision problems
3. Optimal treatment assignments
4. **Arguments for randomization**
5. Conclusion

Arguments for randomization

Identification

- In the beginning I showed identification of the ATE with random assignment.
- Is the ATE still identified without randomization?
- Yes, for controlled assignment!

Proposition (Conditional independence)

Suppose that (X_i, Y_i^0, Y_i^1) *are i.i.d. draws from the population of interest, which are independent of* U*. Then any treatment assignment of the form* $D_i = d_i(X_1, \ldots, X_n, U)$ *satisfies conditional independence,*

$$(Y_i^0, Y_i^1) \perp D_i | X_i.$$

This is true, in particular, for deterministic treatment assignments of the form $D_i = d_i(X_1, \ldots, X_n)$*.*

Arguments for randomization

Adversarial audience

- I did not formally define worst-case risk for randomized procedures before. The definition I implicitly used was

$$\bar{R}(\delta, U) = \sup_{\theta \in \Theta} R(\delta(\cdot, U), \theta).$$

  Worst-case θ is chosen "after" realization of U.
- Possible alternative definition:

$$\bar{R}(\delta) = \sup_{\theta \in \Theta} \left(\sum_{u=1}^{k} p_u \cdot R(\delta(\cdot, u), \theta) \right).$$

  - Worst-case θ is chosen "before" realization of U.
  - In this case, random strategies can be optimal.
  - Has been justified by reference to adversarial audience.
  - Assumes that audience doesn't care about imbalanced covariates, as long as they are the product of randomness.
  - Note: Conditional on knowledge of audience, experimental estimates are biased!

Arguments for randomization

Randomization inference

- Randomization inference requires randomization.
- Randomization inference tests strong null hypotheses of the form $Y_i^1 = Y_i^0$ for all i.
- By our theorem, randomization inference can not be the solution to any decision problem.
- Compromise approach: Randomize only among treatment assignments that yield low expected mean squared error.

Conclusion

- Causality requires exogenous variation.
- In social and life sciences, there is unobserved heterogeneity.
- Randomization makes treatment and control groups the same *in expectation*.
- In practice there is also observed heterogeneity.
- We get better estimates of causal effects by balancing covariate distributions.
- Identification of causal effects relies on controlled trials (CTs), not randomized controlled trials (RCTs).

A web-app for implementing the proposed optimal designs is available at

`https://maxkasy.github.io/home/treatmentassignment/`

Thank you!