Multiple conditional randomization tests

Qingyuan Zhao

Statistical Laboratory, University of Cambridge

Sep 29, 2021 @ UC Berkeley

Joint work with Yao Zhang (DAMTP, Cambridge);
Manuscripts: `https://arxiv.org/abs/2104.10618`;
Slides: `http://statslab.cam.ac.uk/~qz280/`.

The role of randomization tests became obscure

- Fisher (1935): To substitute t-test when normality is not true and to restore randomization as "the physical basis of the validity of the test".
- Extension by Pitman, Welch, Wilcoxon, Kempthorne, among many others.
- "Nonparametric tests"; Randomization tests = permutation tests.
- In Wikipedia, described on a page about "Resampling (statistics)" together with bootstrap, subsampling, and cross-validation.
- Cambridge dictionary of statistics: "procedures for determining statistical significance directly from data without recourse to some particular sampling distribution".

Rejuvenated interest in randomization tests

- Testing genomic associations;
- Testing conditional independence;
- Conformal inference for machine learning methods;
- Analysis of complex experimental designs;
- Evidence factors in observational studies;
- Causal inference with interference.

Advantage: randomization tests are distribution-free, so no need to derive the sampling distribution analytically.

Goals

This work tries to provide a unified framework that incorporates recent (or not recent) ideas

- Explicit conditioning on the counterfactual or potential outcomes of the experiment;
- Algebraic structure of permutation tests;
- Randomization model versus population model;
- Post-experiment conditioning and randomization;
- Using exchangeability to obtain distribution-free predictive intervals.

Goals

This work tries to provide a unified framework that incorporates recent (or not recent) ideas

- Explicit conditioning on the counterfactual or potential outcomes of the experiment;
- Algebraic structure of permutation tests;
- Randomization model versus population model;
- Post-experiment conditioning and randomization;
- Using exchangeability to obtain distribution-free predictive intervals.

Main thesis

Randomization inference should be precisely understood as what its name suggests: it is a mode of statistical inference that is **based on randomization and nothing more than randomization**.

Goals

This work tries to provide a unified framework that incorporates recent (or not recent) ideas

- Explicit conditioning on the counterfactual or potential outcomes of the experiment;
- Algebraic structure of permutation tests;
- Randomization model versus population model;
- Post-experiment conditioning and randomization;
- Using exchangeability to obtain distribution-free predictive intervals.

Main thesis

Randomization inference should be precisely understood as what its name suggests: it is a mode of statistical inference that is **based on randomization and nothing more than randomization**.

A trichotomy of randomness in data

1. Randomness in nature (counterfactual variables);
2. Randomness introduced by the experimenter (drawing balls, using a pseudo-RNG, etc.);
3. Randomness introduced by the analyst (optional).

Outline

1. A single CRT

2. Multiple CRTs

3. Some CRTs in the literature

4. Discussion

Outline

1. A single CRT
2. Multiple CRTs
3. Some CRTs in the literature
4. Discussion

Setup

- N "experiment" "units"; "treatment" $\boldsymbol{Z} \in \mathcal{Z}$ is randomized.
- Example: $\boldsymbol{Z} = (Z_1, \dots, Z_N)$ collects a common attribute of the units. But this is not required.
- Potential or counterfactual "outcomes": $(Y_1(\boldsymbol{z}), \dots, Y_N(\boldsymbol{z}) \mid \boldsymbol{z} \in \mathcal{Z})$. Observed or factual outcome: $Y_i = Y_i(\boldsymbol{Z})$.
- No interference/SUTVA is treated as part of the null hypothesis instead of an assumption.
- Vector notation: $\boldsymbol{Y}(\boldsymbol{z}) = (Y_1(\boldsymbol{z}), \dots, Y_N(\boldsymbol{z})) \in \mathcal{Y} \subseteq \mathbb{R}^n$ and $\boldsymbol{Y} = (Y_1, \dots, Y_N) \in \mathcal{Y}$.
- Let $\boldsymbol{W} = (\boldsymbol{Y}(\boldsymbol{z}) : \boldsymbol{z} \in \mathcal{Z}) \in \mathcal{W}$ be the potential outcomes schedule.[1] $\mathcal{W}$ contains all functions from $\mathcal{Z}$ to $\mathcal{Y}$.
- Observed covariates $\boldsymbol{X}$ are always conditioned on.

Assumption 1: Randomized experiment

We assume $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{W}$ and the density function $\pi(\cdot)$ of $\boldsymbol{Z}$ is known and positive everywhere.

[1] This terminology is due to David Freedman.

Null hypothesis

A typical (partially) sharp null hypothesis assumes that certain potential outcomes are equal or related.

- Example 1: no interference $H_0 : Y_i(\boldsymbol{z}) = Y_i(\boldsymbol{z}^*)$ whenever $z_i = z_i^*$;
- Example 2: constant treatment effect τ (on top of no interference) $H_0 : Y_i(1) - Y_i(0) = \tau$.

Definition

A (partially) sharp null hypothesis H defines an **imputability mapping**

$$\begin{aligned} \mathcal{H} :\ \mathcal{Z} \times \mathcal{Z} &\to 2^{[N]}, \\ (\boldsymbol{z}, \boldsymbol{z}^*) &\mapsto \mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*), \end{aligned}$$

where $\mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*)$ is the largest subset of $[N] = \{1, \dots, N\}$ such that $\boldsymbol{Y}_{\mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*)}(\boldsymbol{z}^*)$ is imputable from $\boldsymbol{Y}(\boldsymbol{z})$ under H.

- **Fully sharp** means that $\mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*) \equiv [N]$. Otherwise **partially sharp**.

Conditional randomization tests for discrete treatments

A **conditional randomization test** (CRT) for a discrete treatment $\boldsymbol{Z}$ is defined by

1. A *partition* $\mathcal{R} = \{\mathcal{S}_m\}_{m=1}^M$ of $\mathcal{Z}$; and
2. A collection of *test statistics* $(T_m(\cdot,\cdot))_{m=1}^M$, where $T_m : \mathcal{Z} \times \mathcal{W} \to \mathbb{R}$.

Importance of partitioning

- Any partition $\mathcal{R}$ defines an equivalent relation $\equiv_{\mathcal{R}}$ (and vice versa).
- Let $\mathcal{S}_{\boldsymbol{z}}$ denote the equivalence class containing $\boldsymbol{z}$.
- For any $\boldsymbol{z} \in \mathcal{Z}$ and $\boldsymbol{z}^* \in \mathcal{S}_{\boldsymbol{z}}$, we have $\boldsymbol{z} \in \mathcal{S}_{\boldsymbol{z}}$, $\mathcal{S}_{\boldsymbol{z}^*} = \mathcal{S}_{\boldsymbol{z}}$ and $T_{\boldsymbol{z}^*}(\cdot,\cdot) = T_{\boldsymbol{z}}(\cdot,\cdot)$.

Definition: *p*-value

$$\begin{aligned} P(\boldsymbol{Z},\boldsymbol{W}) &= \mathbb{P}^*\{T_{\boldsymbol{Z}}(\boldsymbol{Z}^*,\boldsymbol{W}) \le T_{\boldsymbol{Z}}(\boldsymbol{Z},\boldsymbol{W}) \mid \boldsymbol{Z}^* \in \mathcal{S}_{\boldsymbol{Z}}, \boldsymbol{W}\} \\ &= \mathbb{P}^*\{T_{\boldsymbol{Z}}(\boldsymbol{Z}^*,\boldsymbol{W}) \le T_{\boldsymbol{Z}}(\boldsymbol{Z},\boldsymbol{W}) \mid \boldsymbol{Z}^* \equiv_{\mathcal{R}} \boldsymbol{Z}, \boldsymbol{W}\}. \end{aligned}$$

where $\boldsymbol{Z}^*$ is an independent copy of $\boldsymbol{Z}$ conditional on $\boldsymbol{W}$.

Properties of CRT

Computable?

- $T_{\boldsymbol{z}}(\cdot,\cdot)$ is said to be **imputable** under H if for all $\boldsymbol{z}^* \in \mathcal{S}_{\boldsymbol{z}}$, $T_{\boldsymbol{z}}(\boldsymbol{z}^*, \boldsymbol{W})$ only depends on $\boldsymbol{W}$ through its imputable part $\boldsymbol{Y}_{\mathcal{H}(\boldsymbol{z},\boldsymbol{z}^*)}(\boldsymbol{z}^*)$.
- Lemma: Suppose Assumption 1 is satisfied and $T_{\boldsymbol{z}}(\cdot,\cdot)$ is imputable for all $\boldsymbol{z} \in \mathcal{Z}$. Then $P(\boldsymbol{Z}, \boldsymbol{W})$ only depends on $\boldsymbol{Z}$ and $\boldsymbol{Y}$.
- In this case we say the p-value is **computable** under H and denote it by $P(\boldsymbol{Z}, \boldsymbol{Y})$.

Valid?

- Theorem: $\mathbb{P}\left\{P(\boldsymbol{Z}, \boldsymbol{W}) \le \alpha \mid \boldsymbol{Z} \in \mathcal{S}_{\boldsymbol{z}}, \boldsymbol{W}\right\} \le \alpha,\ \forall \alpha \in [0,1], \boldsymbol{z} \in \mathcal{Z}$.
- Moreover, given Assumption 1 and a partially sharp null H, if $P(\boldsymbol{Z}, \boldsymbol{W})$ is computable, then $\mathbb{P}\left\{P(\boldsymbol{Z}, \boldsymbol{Y}) \le \alpha\right\} \le \alpha,\ \forall \alpha \in [0,1]$.
- Proof: Apply probability integral transform.
- Remark: Validity does not depend on computability.

How to construct a CRT? I

Method 1: Condition on a function of the treatment (Hennessy *et al.* 2016)

- Condition on $g(\boldsymbol{Z})$ or equivalently the set $\mathcal{S}_{\boldsymbol{z}} = \{\boldsymbol{z}^* \in \mathcal{Z} : g(\boldsymbol{z}^*) = g(\boldsymbol{z})\}$.

Method 2: Imputable intersection

- Challenge: In many problems, $T_{\boldsymbol{z}}(\boldsymbol{z}^*, \boldsymbol{W}) = T_{\boldsymbol{z}}(\boldsymbol{z}^*, \boldsymbol{Y}(\boldsymbol{z}^*))$. However, only the sub-vector $\boldsymbol{Y}_{\mathcal{H}(\boldsymbol{z},\boldsymbol{z}^*)}(\boldsymbol{z}^*)$ is imputable under H.
- Natural solution: Use test statistics $T_m(\boldsymbol{z}, \boldsymbol{Y}_{\mathcal{H}_m}(\boldsymbol{z}))$ where $\mathcal{H}_m = \bigcap_{\boldsymbol{z},\boldsymbol{z}^* \in \mathcal{S}^m} \mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*)$.
- Tradeoff: Coarser $\mathcal{R} \Longrightarrow$ larger subset of treatment assignments but smaller subset of experimental units.
- Problem becomes: How to choose $\mathcal{R}$? Difficult with complex $\mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*)$.

How to construct a CRT? II

Method 3: Focal units (Aronow 2012; Athey *et al.* 2018)

- Suppose H has a **level-set structure** (Athey *et al.* 2018) in the sense that it exists **exposure functions** $D_i : \mathcal{Z} \to \mathcal{D}$, such that (Manski 2013; Ugander *et al.* 2013; Aronow and Samii 2017)

$$\mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*) = \{i \in [N] : D_i(\boldsymbol{z}) = D_i(\boldsymbol{z}^*)\}.$$

- Consequently, $\mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*)$ is symmetric, and

$$\mathcal{H}_m = \bigcap_{\boldsymbol{z}, \boldsymbol{z}^* \in \mathcal{S}_m} \mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*) = \{i \in [N] : D_i(\boldsymbol{z}) \text{ is a constant over } \boldsymbol{z} \in \mathcal{S}_m\}.$$

- The other direction is often easier: choose $\mathcal{R}$ such that $\mathcal{H}_m = \mathcal{I}$ ("focal units") for all m.
- Specifically, the conditioning set is

$$\mathcal{S}_{\boldsymbol{z}} = \{\boldsymbol{z}^* \in \mathcal{Z} : \boldsymbol{D}_{\mathcal{I}}(\boldsymbol{z}^*) = \boldsymbol{D}_{\mathcal{I}}(\boldsymbol{z})\},$$

where $\boldsymbol{D}_{\mathcal{I}}(\cdot) = (D_i(\cdot) : i \in \mathcal{I})$. Easy to verify that this is a partition.

How to construct a CRT? III

Method 4: Bipartite graph representation (Puelz *et al.* 2019)

- Suppose the imputability mapping admits the form (suppose $0 \in \mathcal{D}$)

$$\mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*) = \{i \in [N] : D_i(\boldsymbol{z}) = D_i(\boldsymbol{z}^*) = 0\}.$$

- Then

$$\mathcal{H}_m = \bigcap_{\boldsymbol{z}, \boldsymbol{z}^* \in \mathcal{S}_m} \mathcal{H}(\boldsymbol{z}, \boldsymbol{z}^*) = \{i \in [N] : D_i(\boldsymbol{z}) = 0, \forall \boldsymbol{z} \in \mathcal{S}_m\}.$$

- This can be represented as a bipartite **null exposure graph** with vertex set $\mathcal{V} = [N] \cup \mathcal{Z}$ and edge set $\mathcal{E} = \{(i, \boldsymbol{z}) \in [N] \times \mathcal{Z} : D_i(\boldsymbol{z}) = 0\}$.
- Key insight in Puelz *et al.* (2019): $\mathcal{H}_m \cup \mathcal{S}_m$ forms a biclique (complete bipartite subgraph).
- So the problem is reduced to finding a collection of large bicliques $\mathcal{H}_m \cup \mathcal{S}_m$ in the graph such that $\{\mathcal{S}_m\}_{m=1}^M$ partitions $\mathcal{Z}$.

How to construct a CRT? IV

More general viewpoint

- Condition on the σ-algebra: $\mathcal{G} = \sigma\left(\{\boldsymbol{Z} \in \mathcal{S}_m\}_{m=1}^{\infty}\right)$.
- This allows us to consider continuous treatments and more complicated conditioning events.

Method 5: Randomized CRTs (Basse *et al.* 2019)

- Motivation: We may have several ways to partition $\mathcal{Z}$ (e.g. multiple sets of focal units).
- Key idea: condition on $G = g(\boldsymbol{Z}, V)$ where V is randomized by the analyst, so $V \perp\!\!\!\perp \boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{W}$.
- Post-randomization and conditioning change the density of $\boldsymbol{Z}$:

$$\pi(\boldsymbol{z} \mid g) = \frac{\mathbb{P}(G = g \mid \boldsymbol{Z} = \boldsymbol{z})\pi(\boldsymbol{z})}{\int \mathbb{P}(G = g \mid \boldsymbol{Z} = \boldsymbol{z})\pi(\boldsymbol{z})\mathrm{d}\boldsymbol{z}}.$$

- The **randomized p-value** is defined as

$$P(\boldsymbol{Z}, \boldsymbol{W}; G) = \mathbb{P}^* \left\{T_G(\boldsymbol{Z}^*, \boldsymbol{W}) \leq T_G(\boldsymbol{Z}, \boldsymbol{W}) \mid G, \boldsymbol{W}\right\}.$$

Outline

1 A single CRT

2 Multiple CRTs

3 Some CRTs in the literature

4 Discussion

Setup

- K conditional randomization tests, defined by partitions $\mathcal{R}^{(k)} = \left\{\mathcal{S}_m^{(k)}\right\}_{m=1}^{\infty}$ and test statistics $(T_m^{(k)}(\cdot,\cdot))_{m=1}^{\infty}$, for K possibly different hypotheses $H^{(k)}$, $k = 1, \dots, K$.
- Corresponding p-values: $P^{(1)}(\boldsymbol{Z}, \boldsymbol{W}), \dots, P^{(K)}(\boldsymbol{Z}, \boldsymbol{W})$.
- For any subset of tests $\mathcal{J} \subseteq [K]$, we define the *union*, *refinement* and *coarsening* of the conditioning sets as

$$\mathcal{R}^{\mathcal{J}} = \bigcup_{k \in \mathcal{J}} \mathcal{R}^{(k)}, \quad \underline{\mathcal{R}}^{\mathcal{J}} = \left\{ \bigcap_{j \in \mathcal{J}} \mathcal{S}_{\boldsymbol{z}}^{(j)} : \boldsymbol{z} \in \mathcal{Z} \right\}, \quad \text{and} \quad \overline{\mathcal{R}}^{\mathcal{J}} = \left\{ \bigcup_{j \in \mathcal{J}} \mathcal{S}_{\boldsymbol{z}}^{(j)} : \boldsymbol{z} \in \mathcal{Z} \right\}.$$

- Generated σ-algebras: $\mathcal{G}^{(k)}$, $\mathcal{G}^{\mathcal{J}}$, $\underline{\mathcal{G}}^{\mathcal{J}}$, $\overline{\mathcal{G}}^{\mathcal{J}}$.

Main theorem

Suppose the following two conditions are satisfied for all $j, k \in [K]$, $j \neq k$:

$$\underline{\mathcal{R}}^{\{j,k\}} \subseteq \mathcal{R}^{\{j,k\}}, \tag{1}$$

$$T_{\boldsymbol{Z}}^{(j)}(\boldsymbol{Z}, \boldsymbol{W}) \perp\!\!\!\perp T_{\boldsymbol{Z}}^{(k)}(\boldsymbol{Z}, \boldsymbol{W}) \mid \underline{\mathcal{G}}^{\{j,k\}}, \boldsymbol{W}. \tag{2}$$

Then we have

$$\mathbb{P}\left\{P^{(1)}(\boldsymbol{Z}, \boldsymbol{W}) \leq \alpha^{(1)}, \dots, P^{(K)}(\boldsymbol{Z}, \boldsymbol{W}) \leq \alpha^{(K)} \mid \overline{\mathcal{G}}^{[K]}, \boldsymbol{W}\right\} \leq \prod_{k=1}^{K} \alpha^{(k)}, \ \forall \alpha^{(1)}, \dots, \alpha^{(K)} \in [0,1].$$

Moreover, given Assumption 1 and the null hypotheses $H^{(1)}, \dots, H^{(K)}$, if the CRTs are computable, then

$$\mathbb{P}\left\{P^{(1)}(\boldsymbol{Z}, \boldsymbol{Y}) \leq \alpha^{(1)}, \dots, P^{(K)}(\boldsymbol{Z}, \boldsymbol{Y}) \leq \alpha^{(K)}\right\} \leq \prod_{k=1}^{K} \alpha^{(k)}, \quad \forall \alpha^{(1)}, \dots, \alpha^{(K)} \in [0,1].$$

A simple case: $K = 2$

What does the theorem say?

- Condition (1) assumes a nested structure between the partitions:

$$\mathcal{S}_{\boldsymbol{z}}^{(1)} \cap \mathcal{S}_{\boldsymbol{z}}^{(2)} = \mathcal{S}_{\boldsymbol{z}}^{(1)} \text{ or } \mathcal{S}_{\boldsymbol{z}}^{(2)} \text{ for all } \boldsymbol{z} \in \mathcal{Z}.$$

- It allows $\mathcal{S}_{\boldsymbol{z}}^{(1)} \subseteq \mathcal{S}_{\boldsymbol{z}}^{(2)}$ for some $\boldsymbol{z}$ and $\mathcal{S}_{\boldsymbol{z}^*}^{(1)} \supseteq \mathcal{S}_{\boldsymbol{z}^*}^{(2)}$ for another $\boldsymbol{z}^* \neq \boldsymbol{z}$.
- Condition (2) assumes the conditional independence

$$T_{\boldsymbol{z}}^{(1)}(\boldsymbol{Z}, \boldsymbol{W}) \perp\!\!\!\perp T_{\boldsymbol{z}}^{(2)}(\boldsymbol{Z}, \boldsymbol{W}) \mid \boldsymbol{Z} \in \mathcal{S}_{\boldsymbol{z}}^{(1)} \cap \mathcal{S}_{\boldsymbol{z}}^{(2)}, \boldsymbol{W}, \quad \forall \boldsymbol{z} \in \mathcal{Z}.$$

- The main conclusion of the Theorem is that

$$\mathbb{P}\left\{P^{(1)}(\boldsymbol{Z}, \boldsymbol{W}) \leq \alpha^{(1)}, P^{(2)}(\boldsymbol{Z}, \boldsymbol{W}) \leq \alpha_2 \mid \boldsymbol{Z} \in \mathcal{S}_{\boldsymbol{z}}^{(1)} \cup \mathcal{S}_{\boldsymbol{z}}^{(2)}, \boldsymbol{W}\right\} \leq \alpha^{(1)}\alpha^{(2)}$$

for all $\boldsymbol{z} \in \mathcal{Z}$ and $\alpha^{(1)}, \alpha^{(2)} \in [0, 1]$.

A simple case: $K = 2$

Proof under strong conditions

- Consider a stronger version of condition (1): $\mathcal{S}_{\boldsymbol{z}}^{(1)} \supseteq \mathcal{S}_{\boldsymbol{z}}^{(2)}$ for all $\boldsymbol{z} \in \mathcal{Z}$. Then $\mathcal{G}^{(1)} \subseteq \mathcal{G}^{(2)}$.
- Furthermore, suppose $T^{(1)}(\boldsymbol{z}, \boldsymbol{w})$ only depends on $\boldsymbol{z}$ through the indicators $1\{\boldsymbol{z} \in \mathcal{S}_m^{(2)}\}, m = 1, \dots$. In other words, $T^{(1)}(\boldsymbol{Z}, \boldsymbol{w})$ is $\mathcal{G}^{(2)}$-measurable, which implies the conditional independence (2).
- Let $\psi^{(k)}(\boldsymbol{Z}, \boldsymbol{W}) = 1\{P^{(k)}(\boldsymbol{Z}, \boldsymbol{W}) \le \alpha^{(k)}\}, k = 1, 2$ be the test functions.

Then by the law of iterated expectation, for any $\boldsymbol{w} \in \mathcal{W}$,

$$\begin{aligned}
\mathbb{P}\left\{P^{(1)}(\boldsymbol{Z}, \boldsymbol{w}) \le \alpha^{(1)}, P^{(2)}(\boldsymbol{Z}, \boldsymbol{w}) \le \alpha^{(2)} \mid \mathcal{G}^{(1)}\right\} &= \mathbb{E}\left\{\psi^{(1)}(\boldsymbol{Z}, \boldsymbol{w})\psi^{(2)}(\boldsymbol{Z}, \boldsymbol{w}) \mid \mathcal{G}^{(1)}\right\} \\
&= \mathbb{E}\left\{\mathbb{E}\left[\psi^{(1)}(\boldsymbol{Z}, \boldsymbol{w})\psi^{(2)}(\boldsymbol{Z}, \boldsymbol{w}) \mid \mathcal{G}^{(2)}\right] \mid \mathcal{G}^{(1)}\right\} \\
&= \mathbb{E}\left\{\psi^{(1)}(\boldsymbol{Z}, \boldsymbol{w})\mathbb{E}\left[\psi^{(2)}(\boldsymbol{Z}, \boldsymbol{w}) \mid \mathcal{G}^{(2)}\right] \mid \mathcal{G}^{(1)}\right\} \\
&\le \alpha^{(2)}\mathbb{E}\left\{\psi^{(1)}(\boldsymbol{Z}, \boldsymbol{w}) \mid \mathcal{G}^{(1)}\right\} \\
&\le \alpha^{(1)}\alpha^{(2)}.
\end{aligned}$$

General case: Main ideas of the proof

1. The conditioning events $\mathrm{R}^{[K]}$ can be partially ordered by set inclusion. This induces a directed acyclic graph (so-called Hasse diagram): $\mathcal{S} \to \mathcal{S}'$ if $\mathcal{S} \supset \mathcal{S}'$ and there is no other $\mathcal{S}''$ such that $\mathcal{S} \supset \mathcal{S}'' \supset \mathcal{S}'$.
2. Let $\mathcal{K}(\mathcal{S}) = \{k \in [K] : \mathcal{S} \in \mathcal{R}^{(k)}\}$. The nested events condition (1) implies certain properties of the Hasse diagram. For example, $\{\mathcal{K}(\mathrm{an}(\mathcal{S})), \mathcal{K}(\mathcal{S}), \mathcal{K}(\mathrm{de}(\mathcal{S}))\}$ forms a partition of $[K]$.
3. The conditions (1) and (2) imply that for any $\mathcal{S} \in \mathcal{R}^{[K]}$, $j \in \mathcal{K}(\mathcal{S})$, and $k \in \mathcal{K}(\mathrm{an}(\mathcal{S}) \cup \{\mathcal{S}\}) \setminus \{j\}$,

$$P^{(j)}(\boldsymbol{Z}, \boldsymbol{W}) \perp\!\!\!\perp P^{(k)}(\boldsymbol{Z}, \boldsymbol{W}) \mid \boldsymbol{Z} \in \mathcal{S}, \boldsymbol{W}.$$

4. Using this and induction, we can show a stronger result than the main theorem: for any $\mathcal{S} \in \mathcal{R}^{[K]}$,

$$\begin{aligned} &\mathbb{P}\left\{P^{(1)}(\boldsymbol{Z}, \boldsymbol{W}) \le \alpha^{(k)}, \ldots, P^{(K)}(\boldsymbol{Z}, \boldsymbol{W}) \le \alpha^{(K)} \mid \boldsymbol{Z} \in \mathcal{S}, \boldsymbol{W}\right\} \\ \le\ &\mathbb{P}\left\{P^{(k)}(\boldsymbol{Z}, \boldsymbol{W}) \le \alpha^{(k)} \text{ for } k \in \mathcal{K}(\mathrm{an}(\mathcal{S})) \mid \boldsymbol{Z} \in \mathcal{S}, \boldsymbol{W}\right\} \prod_{j \in \mathcal{K}(\{\mathcal{S}\} \cup \mathrm{de}(\mathcal{S}))} \alpha_j. \end{aligned}$$

How to construct ”nearly independent” CRTs? I

Method 1: Independent treatment variables

Proposition: the conditions (1) and (2) are satisfied if

1. The tests are unconditional: $\mathcal{S}_{\boldsymbol{z}}^{(k)} = \mathcal{Z}$ for all k and $\boldsymbol{z}$; and
2. $T^{(k)}(\boldsymbol{Z}, \boldsymbol{W})$ only depends on $\boldsymbol{Z}$ through $\boldsymbol{Z}^{(k)} = h^{(k)}(\boldsymbol{Z})$ for all k and $\boldsymbol{Z}^{(j)} \perp\!\!\!\perp \boldsymbol{Z}^{(k)}$ for all $j \neq k$.

This can be easily extended to the case where $\mathcal{R}^{(1)} = \cdots = \mathcal{R}^{(K)}$.

Method 2: Sequential CRTs

Proposition: the conditions (1) and (2) are satisfied if

1. $\mathcal{S}_{\boldsymbol{z}}^{(1)} \supseteq \cdots \supseteq \mathcal{S}_{\boldsymbol{z}}^{(K)}$ for all $\boldsymbol{z} \in \mathcal{Z}$; and
2. $T^{(j)}(\boldsymbol{z}, \boldsymbol{W})$ does not depend on $\boldsymbol{z}$ when $\boldsymbol{z} \in \mathcal{S}_{m}^{(k)}$ for all m and $k > j$.

How to construct "nearly independent" CRTs? II

Method 3: Post-randomization (Bates *et al.* 2020)

- Suppose the test statistics are $T^{(k)}(\boldsymbol{Z}^{(k)}, \boldsymbol{W})$ and there exists a U such that

$$\boldsymbol{Z}^{(1)} \perp\!\!\!\perp \cdots \perp\!\!\!\perp \boldsymbol{Z}^{(K)} \mid U.$$

  U is unobserved but the joint distribution of $(U, \boldsymbol{Z})$ is known.

- The key idea of Bates *et al.* (2020) is that we can construct a post-randomization $G = g(\boldsymbol{Z}, V)$ such that $G \overset{d}{=} U \mid \boldsymbol{Z}$. Then

$$\boldsymbol{Z}^{(1)} \perp\!\!\!\perp \cdots \perp\!\!\!\perp \boldsymbol{Z}^{(K)} \mid G.$$

  We can then condition on G (this changes the distribution of $\boldsymbol{Z}$) and use Method 1.

- This might seem magical at first, but notice that the power of the test will depend on how G resembles U.

Outline

1. A single CRT

2. Multiple CRTs

3. Some CRTs in the literature

4. Discussion

Permutation tests for treatment effect

- Equivalent to a CRT that conditions on the order statistics of $\boldsymbol{Z}$. Equivalently,

$$\mathcal{S}_{\boldsymbol{z}} = \{(z_{\sigma(1)}, \ldots, z_{\sigma(N)}) : \sigma \text{ is a permutation of } [N]\}.$$

- What if we condition on more events? Efron *et al.* (2001) consider a "balanced" permutation test

$$\mathcal{S}_{\boldsymbol{z}} = \{\boldsymbol{z}^* : \boldsymbol{z}^* \text{ is a permutation of } \boldsymbol{z} \text{ and } \boldsymbol{z}^T \boldsymbol{z}^* = N/4\},$$

  when $\boldsymbol{Z}$ is randomized uniformly over $\mathcal{Z} = \{\boldsymbol{z} \in \{0,1\}^N : \boldsymbol{z}^T \mathbf{1} = N/2\}$.
- A counterexample with inflated type I error is provided by Southworth *et al.* (2009), who argued that the problem is that $\mathcal{S}_{\boldsymbol{z}}$ is not a group under balanced permutations (nor is $\mathcal{S}_{\boldsymbol{z}} \cup \{\boldsymbol{z}\}$).
- In view of our theory, the problem is that this violates the invariance: $\mathcal{S}_{\boldsymbol{z}^*} = \mathcal{S}_{\boldsymbol{z}}$ whenever $\boldsymbol{z}^* \in \mathcal{S}_{\boldsymbol{z}}$.

Permutation tests for independence

- Suppose we observed i.i.d. variables $(Z_1, Y_1), \dots, (Z_n, Y_n)$ and would like to test $H_0 : Z_1 \perp\!\!\!\perp Y_1$.
- In classical treatment, the key is to establish the **permutation principle** under the null:

$$(Z_1, \dots, Z_N, Y_1, \dots, Y_N) \overset{d}{=} (Z_{\sigma(1)}, \dots, Z_{\sigma(N)}, Y_1, \dots, Y_N) \text{ for all permutations } \sigma \text{ of } [N].$$

- Lehmann (1975) refers to this as the **population model** and the causal inference problem as the **randomization model**. Ernst (2004) argues that the reasoning behind these two models is fundamentally different.

Two sides of the same coin

CRT is valid if Assumption 1 (randomized experiment) and H are both satisfied.

- In causal inference, Assumption 1 is given, so CRT tests H.
- In independence testing, suppose we "define" the potential outcomes as $\boldsymbol{Y}(\boldsymbol{z}) = \boldsymbol{Y}$ for all $\boldsymbol{z} \in \mathcal{Z}$. The "causal" null hypothesis $H_0 : \boldsymbol{Y}(\boldsymbol{z}) = \boldsymbol{Y}(\boldsymbol{z}^*), \forall \boldsymbol{z}, \boldsymbol{z}^* \in \mathcal{Z}$ is automatically satisfied, so CRT tests Assumption 1 which in this case says $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{Y}$.

Randomization tests for conditional independence

- Observing i.i.d. $(Z_1, Y_1, X_1), \ldots, (Z_n, Y_n, X_n)$, we would like to test $Z_1 \perp\!\!\!\perp Y_1 \mid X_1$.
- Randomization tests (as introduced) can be easily applied by treating $\boldsymbol{X} = (X_1, \ldots, X_n)$.
- What should we call this, randomization tests or conditional randomization tests? We prefer the former, but some others have chosen the latter (Candès *et al.* 2018; Berrett *et al.* 2020).

Evidence factors for observational studies

- In sensitivity analysis for unmeasured confounders, it is common to use the upper bounding p-value
$$P(\boldsymbol{Z}, \boldsymbol{Y}) = \sup_{\pi \in \Pi} P(\boldsymbol{Z}, \boldsymbol{Y}; \pi)$$
where Π contains the set of allowed distributions of $\boldsymbol{Z}$.
- Rosenbaum derives analytic forms of $P(\boldsymbol{Z}, \boldsymbol{Y})$ for signed-score tests under his Γ-sensitivity model.
- Rosenbaum (2017) demonstrates that when there are multiple CRTs, the upper-bounding p-value are "nearly independent" when the permutation groups have a knit product structure.
- We think the key insight lies in the construction of sequential CRTs. Specifically, the conditions below does not depend on the distribution of $\boldsymbol{Z}$, so $P^{(1)}(\boldsymbol{Z}, \boldsymbol{Y}; \pi), \dots, P^{(K)}(\boldsymbol{Z}, \boldsymbol{Y}; \pi)$ are "nearly independent" for all π.

Recall the Proposition for sequential CRTs

The conditions (1) and (2) are satisfied if

1. $\mathcal{S}_{\boldsymbol{z}}^{(1)} \supseteq \cdots \supseteq \mathcal{S}_{\boldsymbol{z}}^{(K)}$ for all $\boldsymbol{z} \in \mathcal{Z}$; and
2. $T^{(j)}(\boldsymbol{z}, \boldsymbol{W})$ does not depend on $\boldsymbol{z}$ when $\boldsymbol{z} \in \mathcal{S}_m^{(k)}$ for all m and $k > j$.

Conformal prediction

- Suppose $(X_1, Y_1), \dots, (X_N, Y_N)$ are exchangeable and Y_N is unobserved.
- Would like to construct a prediction interval $\hat{\mathcal{C}}(X_N)$ such that

$$\mathbb{P}(Y_N \in \hat{\mathcal{C}}(X_N)) \leq 1 - \alpha.$$

- Key idea: invert the permutation test for $H_0 : Y_N = y$.
- For example, we may fit any regression model to $(X_1, Y_1), \dots, (X_{N-1}, Y_{N-1}), (X_N, y)$ and let the p-value be the percentile of the residual for (X_N, y). Small p-value means (X_N, y) "conforms" poorly with other observations.
- Our main point: this is a randomization test by viewing **random sampling as a kind of randomization**.
- Suppose there is a (potentially infinite) super-population $(X_i, Y_i)_{i \in \mathcal{I}}$. "Treatment" $Z : [N] \to \mathcal{I}$ selects which units are observed and the order. "Potential outcomes" are given by

$$\boldsymbol{Y}(z) = \left((X_{z(1)}, Y_{z(1)}), \dots, (X_{z(N)}, Y_{z(N)})\right).$$

- We can use a CRT for $H_0 : Y_N = y$ by conditioning on the unordered Z. This can be extended to allow "covariate shift" (i.e. the distribution of $Z(N)$ differs from the rest) (Tibshirani *et al.* 2019).

Outline

1. A single CRT

2. Multiple CRTs

3. Some CRTs in the literature

4. Discussion

Discussion

- Main thesis: Randomization inference is simple—It is based on randomization and nothing more than randomization.
- This is made precise by trichotomizing the randomness into those introduced by nature, experimenter, and analyst.
- To appreciate the flexibility of CRTs, the paper has another example about testing lagged treatment effect in stepped-wedge randomized trials.
- To understand clever proposals, sometimes we need to redefine "potential outcomes" or "randomization". Or perhaps this is the responsibility of the methodologists?

> *The postulate of randomness thus resolves itself into the question, 'Of what population is this a random sample?' which must frequently be asked by every practical statistician.*
>
> *—Fisher "On the Mathematical Foundations of Theoretical Statistics" (1922)*

References

1. P. M. Aronow, *Sociological Methods & Research* **41**, 3–16 (2012).
2. P. M. Aronow, C. Samii, *The Annals of Applied Statistics* **11**, 1912–1947 (2017).
3. S. Athey, D. Eckles, G. W. Imbens, *Journal of the American Statistical Association* **113**, 230–240 (2018).
4. G. Basse, A Feller, P Toulis, *Biometrika* **106**, 487–494 (2019).
5. S. Bates, M. Sesia, C. Sabatti, E. Candès, *Proceedings of the National Academy of Sciences* **117**, 24117–24126 (2020).
6. T. B. Berrett, Y. Wang, R. F. Barber, R. J. Samworth, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82**, 175–197 (2020).
7. E. Candès, Y. Fan, L. Janson, J. Lv, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**, 551–577 (2018).
8. B. Efron, R. Tibshirani, J. D. Storey, V. Tusher, *Journal of the American statistical association* **96**, 1151–1160 (2001).
9. M. D. Ernst, *Statistical Science* **19**, 676–685 (2004).
10. R. A. Fisher, *Philosophical Transactions of the Royal Society of London. Series A* **222**, 309–368 (1922).
11. J. Hennessy, T. Dasgupta, L. Miratrix, C. Pattanayak, P. Sarkar, *Journal of Causal Inference* **4**, 61–80 (2016).
12. E. L. Lehmann, *Nonparametrics: statistical methods based on ranks.* (Holden-day, Inc., 1975).
13. C. F. Manski, *The Econometrics Journal* **16**, S1–S23 (2013).
14. D. Puelz, G. Basse, A. Feller, P. Toulis (2019).
15. P. R. Rosenbaum, *Statistical Science* **32**, 514–530 (2017).
16. L. K. Southworth, S. K. Kim, A. B. Owen, *Journal of Computational Biology* **16**, 625–638 (2009).
17. R. J. Tibshirani, R. Foygel Barber, E. Candes, A. Ramdas, presented at the Advances in Neural Information Processing Systems, vol. 32.
18. J. Ugander, B. Karrer, L. Backstrom, J. Kleinberg, presented at the Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 329–337.