Multiple conditional randomization tests

Qingyuan Zhao

Statistical Laboratory, University of Cambridge

June 27, 2022 @ IMS Annual Meeting

Joint work with Yao Zhang (DAMTP, Cambridge);
Manuscripts: 2203.10980 (single CRT); arXiv:2104.10618 (multiple CRT);
Slides: `http://statslab.cam.ac.uk/~qz280/`.

The meaning of randomization tests has become 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, Kempthorne, among many others.
- Also known as (none of them is very accurate):
  - **Nonparametric** tests;
  - **Permutation** tests;
  - **Rerandomization** tests.
- In Wikipedia, described in 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 (Efron *et al.* 2001; Bates *et al.* 2020);
- Testing conditional independence (Candès *et al.* 2018; Berrett *et al.* 2020);
- Conformal predictive inference for machine learning methods (Vovk *et al.* 2005; Lei *et al.* 2013);
- Analyses of complex experimental designs (Morgan and Rubin 2012; Ji *et al.* 2017);
- Evidence factors in observational studies (Rosenbaum 2017);
- Causal inference with interference (Athey *et al.* 2018; Basse *et al.* 2019).

Randomization tests vs. Permutation tests

- Often used interchangeably.
- But the semantics are clearly different:
  - **Randomization** tests emphasize on the basis of inference (probabilistic).
  - **Permutation** tests emphasize on the computational algorithm (non-probabilistic).
- Over decades, many authors pointed out that they are based on different assumptions. But the terms are still rarely distinguished in practice/classroom.
- Why? The simplest randomization test (for $1/2$ treated $1/2$ control) is a permutation test.
- How should we resolve this?

Our proposal

Use a new term—**quasi-randomization tests**.

Randomization tests vs. Quasi-randomization tests

- Quasi: “used to show that something is almost, but not completely, the thing described.”
- **Quasi-randomization means that we pretend (parts of) the data are randomized**, even though no physical actions of randomization took place.
- We do this all the time: i.i.d., exchangeablity, infinite population. But they are still assumptions.

What’s the fundamental epistemic difference?

- Randomization tests rely on **human action**—randomness introduced by an experiment.
- Quasi-randomization tests rely on **human perception**—randomness we cannot explain and thus believe is part of nature.

- Closely related is **randomized experiment** vs. **quasi-experiment** (termed by Donald Campbell in social science = observational study in statistics).

This talk

This talk has two goals:

1. To clarify what a "randomization test" means and distinguish it from related concepts.
2. To provide a unifying framework that incorporates many old and new ideas about multiple conditional randomization tests.

Outline

1. Single CRT: Theory
2. Single CRT: Examples
3. Multiple CRTs: Theory
4. Multiple CRTs: Examples

Setup

- N units, treatment $\boldsymbol{Z} \in \mathcal{Z}$ is randomized.
- Potential outcomes $\boldsymbol{Y}(\boldsymbol{z}) = (Y_1(\boldsymbol{z}), \ldots, Y_N(\boldsymbol{z}))$; Consistency: $\boldsymbol{Y} = (Y_1, \ldots, Y_N) = Y(\boldsymbol{Z})$.
- Po. outcomes schedule $\boldsymbol{W} = (\boldsymbol{Y}(\boldsymbol{z}) : \boldsymbol{z} \in \mathcal{Z}) \in \mathcal{W}$.

Assumption (Randomization)

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

Null hypothesis

A typical 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 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 $[N] = \{1, \ldots, N\}$ 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**.

- Example 1: No interference + constant treatment effect is fully sharp.
- Example 2: In crossover designs, hypotheses about a particular lagged effect is partially sharp.

Conditional randomization tests (CRT)

- Requries a partition $\mathcal{R} = \{\mathcal{S}_m\}_{m=1}^M$ of $\mathcal{Z}$ and test statistics $(T_m(\cdot,\cdot))_{m=1}^M$, where $T_m : \mathcal{Z} \times \mathcal{W} \to \mathbb{R}$.
- $\mathcal{R}$ defines an equivalent relation $\equiv_{\mathcal{R}}$ (and vice versa).
- Let $\mathcal{S}_{\boldsymbol{z}}$ denote the equivalence class containing $\boldsymbol{z}$. Let $T_{\boldsymbol{z}}(\cdot,\cdot)$ be the corresponding test statistic.
- The *p-value* of the CRT is given by

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

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

Properties of CRT

Valid?

- Theorem: $\mathbb{P}\{P(\boldsymbol{Z}, \boldsymbol{W}) \le \alpha \mid \boldsymbol{Z} \in \mathcal{S}_{\boldsymbol{z}}, \boldsymbol{W}\} \le \alpha, \ \forall \alpha \in [0, 1], \boldsymbol{z} \in \mathcal{Z}$.
- Proof: Apply probability integral transform (Basse *et al.* 2019)

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 $\boldsymbol{Y}$ (we say it's **computable**).
- Remark: without randomization (Assumption 1), the distribution of $\boldsymbol{Z}^* \mid \boldsymbol{W} \overset{d}{=} \boldsymbol{Z} \mid \boldsymbol{W}$ is unknown.

Summary: Randomization guarantees validity, but the test is not always computable.

Further theory

See our paper for

- Alternative viewpoints: Conditioning on a function of the treatment, a σ-algebra, or a post-randomized variable.
- A review of methods to construct computable CRTs (Aronow 2012; Athey *et al.* 2018; Puelz *et al.* 2019).

Fisher's exact test for 2×2 contingency tables

		Outcome Y 0	Outcome Y 1	Total
Treatment A	0	N_{00}	N_{01}	$N_{0\cdot}$
	1	N_{10}	N_{11}	$N_{1\cdot}$
	Total	$N_{\cdot 0}$	$N_{\cdot 1}$	N

Fisher observed that the null probability of observing $(N_{00}, N_{01}, N_{10}, N_{11})$ **given the marginal totals** is given by the hypergeometric distribution. An exact test can then be immediately derived.

- This is a **unconditional randomization test** if the randomization fixes $N_{0\cdot}$ and $N_{1\cdot}$ (as in the famous tea-tasting example).
- This is a **conditional randomization test** if the treatments are assigned by Bernoulli trials.
- This is a **conditional quasi-randomization test** in the "two Binomials" setup: $N_{00} \sim \text{Bin}(N_{0\cdot}, \pi_0)$, $N_{10} \sim \text{Bin}(N_{1\cdot}, \pi_1)$, and the null hypothesis is $H_0 : \pi_0 = \pi_1$.
- This is a permutation test, although resampling is not needed.

Permutation tests for treatment effect in randomized experiments

- This generalizes Fisher's exact test to continuous outcomes or discrete outcomes with more levels.
- This is a **conditional randomization test** that conditions on the order statistics of $\boldsymbol{Z}$, or

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

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

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

  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 $\boldsymbol{z}^* \in \mathcal{S}_{\boldsymbol{z}}$.

Further examples

See our paper for discussion on

- Quasi-randomization tests for (conditional) independence;
- Conformal prediction.

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: $P^{(1)}(\boldsymbol{Z}, \boldsymbol{W}), \dots, P^{(K)}(\boldsymbol{Z}, \boldsymbol{W})$.
- Question: When can we treat them as **independent pieces of evidence**?

A new unifying result

- For any $\mathcal{J} \subseteq [K]$, we define the *union*, *refinement* and *coarsening* of the conditioning sets as

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

- 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

$$\underline{\mathcal{R}}^{\{j,k\}} \subseteq \mathcal{R}^{\{j,k\}}, \quad \forall j,k \in [K], j \neq k. \tag{1}$$

$$T_{\boldsymbol{Z}}^{(j)}(\boldsymbol{Z}, \boldsymbol{W}),\ j \in \mathcal{J} \text{ are independent given } \underline{\mathcal{G}}^{\mathcal{J}}, \boldsymbol{W}, \quad \forall \mathcal{J} \subseteq [K]. \tag{2}$$

Then for any $0 < \alpha^{(1)}, \ldots, \alpha^{(K)} < 1$,

$$\mathbb{P}\Big\{ P^{(1)}(\boldsymbol{Z}, \boldsymbol{W}) \leq \alpha^{(1)}, \ldots, P^{(K)}(\boldsymbol{Z}, \boldsymbol{W}) \leq \alpha^{(K)} \mid \overline{\mathcal{G}}^{[K]}, \boldsymbol{W} \Big\} \leq \prod_{k=1}^{K} \alpha^{(k)}.$$

Special cases

To simplify, suppose $T_m^{(j)} = T^{(j)}$ does not depend on m.

Independent treatment variables

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 $j \neq k$.

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 $k > j$.

Remark: This does not require knowing the distribution $\pi(\cdot)$ of $\boldsymbol{Z}$.

A direct proof for sequential CRTs with $K = 2$

1. $\mathcal{S}_{\boldsymbol{z}}^{(1)} \supseteq \mathcal{S}_{\boldsymbol{z}}^{(2)}$ for all $\boldsymbol{z} \in \mathcal{Z}$, **which implies $\mathcal{G}^{(1)} \subseteq \mathcal{G}^{(2)}$**; and
2. $T^{(1)}(\boldsymbol{z}, \boldsymbol{W})$ does not depend on $\boldsymbol{z}$ when $\boldsymbol{z} \in \mathcal{S}_m^{(2)}$ for all m, **which implies $T^{(1)}(\boldsymbol{Z}, \boldsymbol{w})$ is $\mathcal{G}^{(2)}$-measurable** (and is thus independent of $T^{(2)}(\boldsymbol{Z}, \boldsymbol{w})$ given $\mathcal{G}^{(2)}$).

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

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

The general proof requires a much more careful consideration of the structure of conditioning events.

Evidence factors for observational studies

- In Rosenbaum's or other sensitivity analyses for observational studies, it is common to use the upper bounding p-value

$$P(\boldsymbol{Z}, \boldsymbol{Y}) = \sup_{\pi \in \Pi} P(\boldsymbol{Z}, \boldsymbol{Y}; \pi)$$

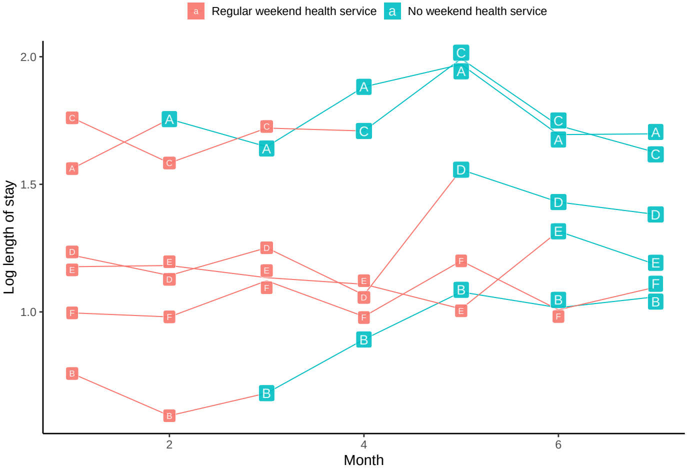
  where Π is the set of allowed distributions of $\boldsymbol{Z}$.
- Suppose $P^{(k)}(\boldsymbol{Z}, \boldsymbol{Y}; \pi), k \in [K]$ are constructed by sequential CRTs.
- Then for all $\pi^* \in \Pi$, we have

$$\begin{aligned} & \mathbb{P}_{\pi^*}\left(P^{(1)}(\boldsymbol{Z}, \boldsymbol{Y}) \le \alpha^{(1)}, \dots, P^{(K)}(\boldsymbol{Z}, \boldsymbol{Y}) \le \alpha^{(K)}\right) \\ \le & \mathbb{P}_{\pi^*}\left(P^{(1)}(\boldsymbol{Z}, \boldsymbol{Y}; \pi^*) \le \alpha^{(1)}, \dots, P^{(K)}(\boldsymbol{Z}, \boldsymbol{Y}; \pi^*) \le \alpha^{(K)}\right) \\ \le & \prod_{k=1}^{K} \alpha^{(k)}. \end{aligned}$$

- This generalizes the "knit product" structure for multiple permutation tests (Rosenbaum 2017).

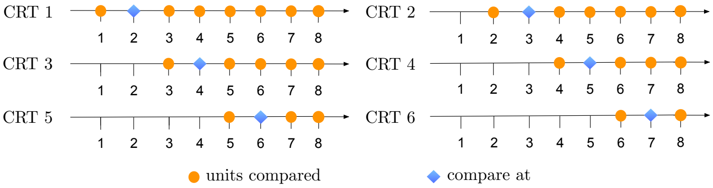
Stepped-wedge design

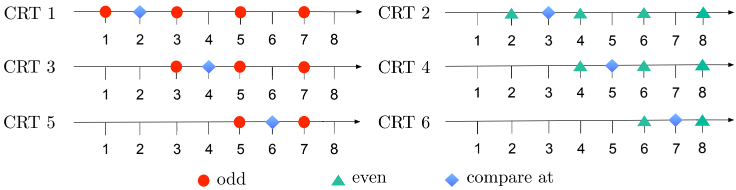
- In a stepped-wedge randomized trial, units/clusters cross over from control to treatment at random times (“staggered adoption”).



Testing lagged treatment effects in stepped-wedge design

- Evidence for (lagged) treatment effect is scattered over time.
- If cleverly constructed, CRTs are “nearly independent” and can be combined by global/multiple testing methods.
- Example below: lag $= 1$.





References

1. P. M. Aronow, *Sociological Methods & Research* **41**, 3–16 (2012).
2. S. Athey, D. Eckles, G. W. Imbens, *Journal of the American Statistical Association* **113**, 230–240 (2018).
3. G. Basse, A Feller, P Toulis, *Biometrika* **106**, 487–494 (2019).
4. S. Bates, M. Sesia, C. Sabatti, E. Candès, *Proceedings of the National Academy of Sciences* **117**, 24117–24126 (2020).
5. 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).
6. E. Candès, Y. Fan, L. Janson, J. Lv, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**, 551–577 (2018).
7. B. Efron, R. Tibshirani, J. D. Storey, V. Tusher, *Journal of the American Statistical Association* **96**, 1151–1160 (2001).
8. X. Ji, G. Fink, P. J. Robyn, D. S. Small, *et al.*, *The Annals of Applied Statistics* **11**, 1–20 (2017).
9. J. Lei, J. Robins, L. Wasserman, *Journal of the American Statistical Association* **108**, 278–287 (2013).
10. K. L. Morgan, D. B. Rubin, *Annals of Statistics* **40**, 1263–1282 (2012).
11. D. Puelz, G. Basse, A. Feller, P. Toulis, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (2019).
12. P. R. Rosenbaum, *Statistical Science* **32**, 514–530 (2017).
13. L. K. Southworth, S. K. Kim, A. B. Owen, *Journal of Computational Biology* **16**, 625–638 (2009).
14. V. Vovk, A. Gammerman, G. Shafer, *Algorithmic learning in a random world*, (Springer, 2005).