# What is a randomization test?

Qingyuan Zhao

Statistical Laboratory, University of Cambridge

March 23, 2023 @ University of Toronto

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).

# 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 in the literature.

# Outline


1. Related concepts and a real data example
2. What is a conditional randomization test (CRT)?
3. Examples
4. Multiple CRTs: A new unifying result
5. Discussion

# Randomization tests vs. Permutation tests

- Often used interchangeably. Some view randomization tests as a special case of permutation tests.
- 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**

Introduce 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.

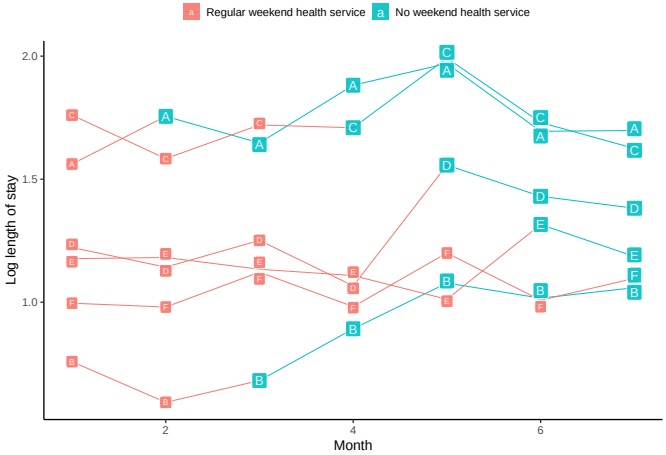
## What's the fundamental epistemic difference?

- 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).

# An example: The Australia weekend health services disinvestment trial

- "Stepped-wedge design:" Six wards in a hospital in Melbourne randomly crossovered to treatment (no weekend health services) in a staggered fashion (Haines *et al.* 2017).



# Which test?

## Sample raw data

| | gender | age | los | step | ward |
|---|---|---|---|---|---|
| 1: | Male | 32 | 3.937500 | 1 | 1 |
| 2: | Female | 23 | 1.425780 | 1 | 1 |
| --- | | | | | |
| 7985: | Male | 65 | 4.093750 | 7 | 6 |
| 7986: | Female | 87 | 1.121090 | 7 | 6 |

## Which variable(s) should we permute?

Time step? Ward? But where is the treatment?

- **Randomization test**: Permute crossover order (which induce an exposure status for each patient).
- **Quasi-randomization tests**: Permute time step, hospital ward, and/or crossover order.

# Results

- Test statistics $T_1$, $T_2$, $T_3$ are the coefficient of patient exposure status in different linear models. Tests are inverted to obtain 90% confidence intervals (CI).

| | $T_1$ (adjust for nothing) | | $T_2$ (adjust for ward) | | $T_3$ (adjust for ward & time) | |
|---|---|---|---|---|---|---|
| | p-value | CI | p-value | CI | p-value | CI |
| *Randomization test* | 0.0833 | [-0.09, 0.72] | 0.0042 | [0.06, 0.3] | 0.0069 | [0.06, 0.31] |
| *Quasi-Randomization tests* | | | | | | |
| Time | 0.0000 | [0.13, 0.24] | 0.0000 | [0.11, 0.21] | 0.0000 | [0.09, 0.23] |
| Ward | 0.0000 | [0.30, 0.42] | 0.0049 | [0.04, 0.19] | 0.0000 | [0.10, 0.25] |
| Time & ward | 0.0000 | [0.25, 0.33] | 0.0000 | [0.11, 0.22] | 0.0000 | [0.09, 0.23] |
| Crossover & time | 0.0029 | [0.08, 0.47] | 0.0000 | [0.12, 0.21] | 0.0000 | [0.09, 0.23] |
| Crossover & ward | 0.0000 | [0.29, 0.41] | 0.0093 | [0.02, 0.18] | 0.0001 | [0.08, 0.24] |
| Crossover, ward & time | 0.0000 | [0.24, 0.33] | 0.0000 | [0.11, 0.21] | 0.0001 | [0.09, 0.23] |
| *Linear model* | 0.0000 | [0.24, 0.34] | 0.0000 | [0.11, 0.22] | 0.0001 | [0.08, 0.24] |

- Quasi-randomization tests are **sensitive to model specification** and tend to **overstate significance**.
- Is it a good idea to permute ward or time?
- Probably not, because the wards have different specialties and some diseases are seasonal.

# Setup

- Consider $N$ "experimental" "units". The "treatment" $\boldsymbol{Z} \in \mathcal{Z}$ is randomized.
- Example: $\boldsymbol{Z} = (Z_1, \ldots, Z_N)$ collects a common attribute of the units. But this is not required.
- Potential "outcomes": $\boldsymbol{Y}(\boldsymbol{z}) = (Y_1(\boldsymbol{z}), \ldots, Y_N(\boldsymbol{z}))$.
- Consistency of the observed outcome: $\boldsymbol{Y} = (Y_1, \ldots, Y_N) = Y(\boldsymbol{Z})$.
- No interference/SUTVA is treated as part of the null hypothesis instead of an assumption.
- Let $\boldsymbol{W} = (\boldsymbol{Y}(\boldsymbol{z}) : \boldsymbol{z} \in \mathcal{Z}) \in \mathcal{W}$ be the potential outcomes schedule.[1]
- Observed covariates $\boldsymbol{X}$ are always conditioned upon.

**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 adapted from Freedman (2009).

# 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 $\tau$ (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**. p

- 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)

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

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

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

# Properties of CRT

## Valid?

- Theorem: $\mathbb{P}\{P(\boldsymbol{Z}, \boldsymbol{W}) \leq \alpha \mid \boldsymbol{Z} \in \mathcal{S}_{\boldsymbol{z}}, \boldsymbol{W}\} \leq \alpha, \ \forall \alpha \in [0,1], \boldsymbol{z} \in \mathcal{Z}$.
- 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: Always valid, but not always computable.

# Alternative viewpoints

## Condition on a function of the treatment (Hennessy *et al.* 2016)

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

## Condition on a $\sigma$-algebra

- Condition on the $\sigma$-algebra $\mathcal{G} = \sigma\left(\{\boldsymbol{Z} \in \mathcal{S}_m\}_{m=1}^{\infty}\right)$ or $\sigma(G)$.

## Most general

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

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

# How to construct a computable CRT?

Suppose, as in most applications, $T_{\boldsymbol{z}}(\boldsymbol{z}^*, \boldsymbol{W}) = T_{\boldsymbol{z}}(\boldsymbol{z}^*, \boldsymbol{Y}(\boldsymbol{z}^*))$.

- The challenge is that 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 $\mathcal{R} \Longrightarrow$ more treatment assignments but fewer experimental units.
- How to choose $\mathcal{R}$? This can be simplified by assuming that $H$ has a **level-set structure** in the sense that there exists **exposure functions** $D_i(\boldsymbol{z}), i \in [N]$ such that

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

- This has attracted a lot of attention in the causal intereference literature (Aronow 2012; Athey *et al.* 2018; Puelz *et al.* 2019). See also our paper.

# Examples

Next we go over some examples and try to classify them according to

- **Basis of inference:** Randomization vs. Quasi-randomization;
- **Usage of conditioning:** Conditional vs. Unconditional;
- **Computation:** Permutation vs. more general resampling.

# Fisher's exact test for $2 \times 2$ contingency tables

| | | Outcome $Y$ | | |
|---|---|---|---|---|
| | | 0 | 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: $N_{00} \sim \mathrm{Bin}(N_{0\cdot}, \pi_0)$, $N_{10} \sim \mathrm{Bin}(N_{1\cdot}, \pi_1)$, 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

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

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

# Permutation tests for independence

- Suppose we observed i.i.d. variables $(Z_1, Y_1), \ldots, (Z_n, Y_n)$ and would like to test $H_0 : Z_1 \perp\!\!\!\perp Y_1$.
- The permutation test is clearly a **conditional quasi-randomization test**.

## Is this intrisincally different from the causal inference problem?

- Some would say **yes** (Lehmann 1975; Ernst 2004).
- Our answer is **no**. They are **two sides of the same coin.**

Recall CRT is valid and computable if Assumption 1 (randomized experiment) and $H$ are both true.

- 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 says $\boldsymbol{Z} \perp\!\!\!\perp \boldsymbol{Y}$.

- Can be extended to test conditional independence (Candès *et al.* 2018; Berrett *et al.* 2020). See also our paper.

# Conformal prediction

- Suppose $(X_1, Y_1), \ldots, (X_N, Y_N)$ are exchangeable and $Y_N$ is unobserved. The goal is 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 $H_0 : Y_N = y$.
- Example: fit any regression to $(X_1, Y_1), \ldots, (X_{N-1}, Y_{N-1}), (X_N, y)$ and let the $p$-value be the percentile of the absolute residual for $(X_N, y)$. Small $p$-value means poor conformity (Vovk *et al.* 2005; Lei *et al.* 2013).
- This is a **conditional quasi-randomization test** by **pretending sampling is randomized**.
- 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

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

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

# 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 $H^{(k)}$, $k = 1, \ldots, K$.
- Corresponding $p$-values: $P^{(1)}(\boldsymbol{Z}, \boldsymbol{W}), \ldots, P^{(K)}(\boldsymbol{Z}, \boldsymbol{W})$.
- 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

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

- Generated $\sigma$-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

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

# 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 $\boldsymbol{w} \in \mathcal{W}$,

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

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

# Example: Evidence factors for observational studies

- In sensitivity analysis 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 $\Pi$ is the set of allowed distributions of $\boldsymbol{Z}$ (Rosenbaum 2002).
- Rosenbaum (2017) shows that the bounding $p$-values for multiple permutation tests are "nearly independent" when the permutation groups have a **knit product** structure.
- A more general viewpoint: $P^{(k)}(\boldsymbol{Z}, \boldsymbol{Y}; \pi), k \in [K]$ are constructed by sequential CRTs (which, crucially, do not depend on $\pi$). So $P^{(k)}(\boldsymbol{Z}, \boldsymbol{Y}; \pi), k \in [K]$ are "nearly independent" for all $\pi$.
- Then for all $\pi^* \in \Pi$, we have

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

# Example: Testing lagged treatment effects in stepped-wedge designs

- In cross-over designs, evidence for causation is scattered over time.
- If cleverly constructed, CRTs are "nearly independent" and can be combined by global/multiple testing methods.
- See our paper for more detail.

# Discussion

- Randomization tests are based entirely on randomization.
- This is made precise by trichotomizing the randomness in data into
  1. Randomness in nature (potential outcomes);
  2. **Randomness introduced by the experimenter (e.g. drawing balls or using a pseudo-RNG);**
  3. Randomness and conditioning introduced by the analyst (optional).
- If we follow this simple principle, randomization tests are always valid.
- What's challenging is to construct computable tests and make them "nearly independent".

> *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)*

- Take-aways message: **Do not call it a randomization test if there is no randomization.**

# 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. M. D. Ernst, *Statistical Science* **19**, 676–685 (2004).
9. R. A. Fisher, *Philosophical Transactions of the Royal Society of London. Series A* **222**, 309–368 (1922).
10. D. A. Freedman, *Statistical Models: Theory and Practice*, (Cambridge University Press, 2009).
11. T. P. Haines *et al.*, *PLoS medicine* **14**, e1002412 (2017).
12. J. Hennessy, T. Dasgupta, L. Miratrix, C. Pattanayak, P. Sarkar, *Journal of Causal Inference* **4**, 61–80 (2016).
13. X. Ji, G. Fink, P. J. Robyn, D. S. Small, *et al.*, *The Annals of Applied Statistics* **11**, 1–20 (2017).
14. E. L. Lehmann, *Nonparametrics: statistical methods based on ranks.* (Holden-day, Inc., 1975).
15. J. Lei, J. Robins, L. Wasserman, *Journal of the American Statistical Association* **108**, 278–287 (2013).
16. K. L. Morgan, D. B. Rubin, *Annals of Statistics* **40**, 1263–1282 (2012).
17. D. Puelz, G. Basse, A. Feller, P. Toulis, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (2019).
18. P. R. Rosenbaum, *Observational studies*, (Springer, 2002).
19. P. R. Rosenbaum, *Statistical Science* **32**, 514–530 (2017).
20. L. K. Southworth, S. K. Kim, A. B. Owen, *Journal of Computational Biology* **16**, 625–638 (2009).
21. R. J. Tibshirani, R. Foygel Barber, E. Candes, A. Ramdas, presented at the Advances in Neural Information Processing Systems, vol. 32.
22. V. Vovk, A. Gammerman, G. Shafer, *Algorithmic learning in a random world*, (Springer, 2005).