Randomization Tests Under an Approximate Symmetry Assumption

Ivan A. Canay
Northwestern University

Joseph P. Romano
Stanford University

Azeem M. Shaikh
University of Chicago
Motivation

- Economic data often exhibits dependence and heterogeneity
  1. Time series
  2. Spatial models
  3. Panel data
  4. More generally: clusters

- Dependence and heterogeneity affects inference on parameters
Motivation

- Economic data often exhibits dependence and heterogeneity
  1. Time series
  2. Spatial models
  3. Panel data
  4. More generally: clusters

- Dependence and heterogeneity affects inference on parameters

- Standard practice typically involve
  1. Use appropriate LLN and CLT
  2. Estimate asymptotic variance $\Sigma$ with some form of HAC or CC.
  3. Use t-test (or Wald) coupled with standard asymptotics.
  4. Note: number of clusters needs to be large.
**This paper**

- **Randomization tests:**
  1. Assume data exhibits a **symmetry** under null hypothesis.
     - i.e., invariance of distr. of data to group of transformations.
  2. Construct tests that control **size exactly**.
  3. But symmetry may not hold in finite samples ...
This paper

- Randomization tests:
  1. Assume data exhibits a symmetry under null hypothesis.
     - i.e., invariance of distr. of data to group of transformations.
  2. Construct tests that control size exactly.
  3. But symmetry may not hold in finite samples ...

- Randomization Tests under Approximate Symmetry:
  1. Assume instead function of data satisfies symmetry approx.
     - i.e., fcn. of data converges weakly to distr. satisfying symmetry.
  2. Construct tests that control null rejection prob. asymptotically.
  3. Setting: data grouped into “clusters”
     - Small number of cluster with many obs. within clusters.
     - Can be heterogeneous and have dependence within/across clusters.
     - Parameter of interest is identified within each cluster.
Outline of Talk

1. Review of Randomization Tests
   A) Symmetric Location Example

2. Randomization Tests under Approximate Symmetry
   A) Asymptotic Results

3. Applications
   A) Time Series Regression
   B) Difference in Differences


5. Conclusions
Review of Randomization Tests

- Observe data $X \sim P \in \mathcal{P}$ on $\mathcal{X}$

- Hypotheses of interest:

  $H_0 : P \in \mathcal{P}_0$  vs.  $H_1 : P \in \mathcal{P} \setminus \mathcal{P}_0$.

- Reject $H_0$ for large values of a test statistic $T = T(X)$. 

Assumption $R_{G X} = X$ for $g \in G$ and $P \in \mathcal{P}$.
Review of Randomization Tests

- Observe data \( X \sim P \in \mathcal{P} \) on \( \mathcal{X} \)

- Hypotheses of interest:

  \[
  H_0 : P \in \mathcal{P}_0 \quad \text{vs.} \quad H_1 : P \in \mathcal{P} \setminus \mathcal{P}_0 .
  \]

- Reject \( H_0 \) for large values of a test statistic \( T = T(X) \).

**Assumption R**

\[
gX \overset{d}{=} X \quad \text{for} \quad g \in \mathbf{G} \quad \text{and} \quad P \in \mathcal{P}_0
\]

\( \mathbf{G} \) is a (finite) group of transformations from \( \mathcal{X} \) to \( \mathcal{X} \)

- **Intuition**: critical value by sampling from \( \mathbf{G} \).
Review of Randomization Tests (cont.)

- We let \( M = |G| \) and \( k = \lceil (1 - \alpha)M \rceil \)

- Compute **ordered values** of \( T(gX) \) for \( g \in G \):

\[
T^{(1)}(X) \leq T^{(2)}(X) \leq \cdots \leq T^{(k)}(X) \leq \cdots \leq T^{(M)}(X)
\]
Review of Randomization Tests (cont.)

- We let $M = |G|$ and $k = \lceil (1 - \alpha)M \rceil$

- Compute ordered values of $T(gX)$ for $g \in G$:

\[
T^{(1)}(X) \leq T^{(2)}(X) \leq \ldots \leq T^{(k)}(X) \leq \ldots \leq T^{(M)}(X)
\]
Review of Randomization Tests (cont.)

- We let $M = |G|$ and $k = \lceil (1 - \alpha)M \rceil$

- Compute ordered values of $T(gX)$ for $g \in G$:

$$T^{(1)}(X) \leq T^{(2)}(X) \leq \cdots \leq T^{(k)}(X) \leq \cdots \leq T^{(M)}(X)$$

- Define

$$M^+(X) \equiv |\{ m : T^{(m)}(X) > T^{(k)}(X) \}|$$
$$M^0(X) \equiv |\{ m : T^{(m)}(X) = T^{(k)}(X) \}|$$

**Definition 1 (Randomization Test)**

$$\phi(X) \equiv \begin{cases} 
1 & T(X) > T^{(k)}(X) \\
\alpha(X) & T(X) = T^{(k)}(X), \quad \text{for} \quad \alpha(X) = \frac{M\alpha - M^+}{M^0} \\
0 & T(X) < T^{(k)}(X) 
\end{cases}$$
Theorem 1

If Assumption R holds, then

\[ E_P[\phi(X)] = \alpha \quad \text{for all } P \in \mathbf{P}_0. \]

Key Idea:

\[ T(X) | T^{(1)}(X), \ldots, T^{(M)}(X) \sim \text{Unif}\{ T^{(1)}(X), \ldots, T^{(M)}(X) \} \]

Remark:

- \( M \) may be too big.
- Can replace with a stochastic approx. without affecting exactness.
- Use \( g_1 = \text{identity} \) and \( g_2, \ldots, g_B \) i.i.d. Unif(\( G \)).
Consider $X = (X_1, \ldots, X_q)$ where

- $X_j$ are independent r.v. on $\mathbb{R}^d$
- $X_j \sim P_j$: $P_j$ is symmetric about $\mu \in \mathbb{R}^d$
- Hypotheses: $H_0: \mu = 0$ vs. $H_1: \mu \neq 0$.

- Let $G = \{-1, 1\}^q$ be the group of sign changes.
- Define $gX = (g_1X_1, \ldots, g_qX_q)$
- Assumption R holds: $\phi(X)$ controls size in finite sample.
Symmetric Location Example

Consider $X = (X_1, \ldots, X_q)$ where

- $X_j$ are independent r.v. on $\mathbb{R}^d$
- $X_j \sim P_j : P_j$ is symmetric about $\mu \in \mathbb{R}^d$
- Hypotheses: $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$.

- Let $G = \{-1, 1\}^q$ be the group of sign changes.
- Define $gX = (g_1 X_1, \ldots, g_q X_q)$
- Assumption R holds: $\phi(X)$ controls size in finite sample.

- If $d = 1$ (i.e. scalar) and $X_j \sim N(\mu, \sigma_j^2)$ or mixture of normals, then
  - t-test valid for certain $\alpha$: $\alpha \leq 8\%$ or $\alpha \leq 10\%$ and $q \leq 14$.
  - but possibly quite conservative.

- See Bakirov & Szekely (2005) and Ibragimov & Müller (2010).
**Symmetric Location Example - Normal**

\[
X_j \sim N(0, 1) \text{ for } j \leq q/2 \text{ and } X_j \sim N(0, a^2) \text{ for } j > q/2
\]

\[t\text{-test as in IM vs. rand. test using } T = |t\text{-stat}|\]
**Symmetric Location Example - Normal**

\[ X_j \sim N(0, 1) \text{ for } j \leq q/2 \text{ and } X_j \sim N(0, a^2) \text{ for } j > q/2 \]

*t-test* as in IM vs. rand. test using \( T = |t-\text{stat}| \)

**Figure:** Rejection rates. *t-test* and randomization test. Parameter values are the following: \( q = 8 \) (left panel) and \( q = 16 \) (right panel), \( \alpha = 0.05 \), and 100,000 MC.
Comments

**Randomization test:**

- valid for all $\alpha$ ($\Rightarrow$ p-values)
RANDOMIZATION TEST:

- valid for all $\alpha$ ($\Rightarrow$ p-values)
- valid for $d > 1$
Randomization test:

- valid for all $\alpha$ ($\Rightarrow$ p-values)
- valid for $d > 1$
- valid for any test statistic
**Randomization test:**

- valid for all $\alpha$ ($\Rightarrow$ p-values)
- valid for $d > 1$
- valid for any test statistic
- similar ($\Rightarrow$ better power)
Randomization test:

- valid for all $\alpha$ ($\Rightarrow$ p-values)
- valid for $d > 1$
- valid for any test statistic
- similar ($\Rightarrow$ better power)
- valid for other symmetric distributions
Outline

1. Review of Randomization Tests
   A) Symmetric Location Example

2. Randomization Tests under Approximate Symmetry
   A) Asymptotic Results

3. Applications
   A) Time Series Regression
   B) Difference in Differences


5. Conclusions
RT UNDER APPROXIMATE SYMMETRY

- Observe data $X^{(n)} \sim P_n \in P_n$ on $\mathcal{X}_n$

- Hypotheses of interest:

$$H_0 : P_n \in P_{n,0} \quad \text{vs.} \quad H_1 : P_n \in P_n \setminus P_{n,0}.$$
RT UNDER APPROXIMATE SYMMETRY

- Observe data $X^{(n)} \sim P_n \in \mathbb{P}_n$ on $\mathcal{X}_n$

- Hypotheses of interest:
  
  $H_0 : P_n \in \mathbb{P}_{n,0}$ \hspace{1em} vs. \hspace{1em} $H_1 : P_n \in \mathbb{P}_n \setminus \mathbb{P}_{n,0}$.

**Assumption RW**

Let $S_n : \mathcal{X}_n \rightarrow S (\subseteq \text{Euclidean Space})$ be a function of the data.

1. $S_n = S_n(X^{(n)}) \rightsquigarrow S$ under $P_n$ when $P_n \in \mathbb{P}_{n,0}$

2. $gS \overset{d}{=} S$ for all $g \in G$

  $G$ is a (finite) group of transformations from $S$ to $S$
Observe data $X^{(n)} \sim P_n \in \mathcal{P}_n$ on $\mathcal{X}_n$

Hypotheses of interest:

$H_0 : P_n \in \mathcal{P}_{n,0}$ vs. $H_1 : P_n \in \mathcal{P}_n \setminus \mathcal{P}_{n,0}$

**Assumption RW**

Let $S_n : \mathcal{X}_n \to S (\subseteq \text{Euclidean Space})$ be a function of the data.

1. $S_n = S_n(X^{(n)}) \sim S$ under $P_n$ when $P_n \in \mathcal{P}_{n,0}$
2. $gS \overset{d}{=} S$ for all $g \in G$
   
   $G$ is a (finite) group of transformations from $S$ to $S$

Reject $H_0$ for large values of a test statistic $T = T(S_n)$.

As before, we let $M = |G|$ and $k = \lceil (1 - \alpha)M \rceil$. 
Definition 2 (RT under Weak Convergence)

\[
\phi(S_n) \equiv \begin{cases} 
1 & T(S_n) > T^{(k)}(S_n) \\
 a(S_n) & T(S_n) = T^{(k)}(S_n) \\
0 & T(S_n) < T^{(k)}(S_n) 
\end{cases} \quad \text{for} \quad a(S_n) = \frac{M\alpha - M^+(S_n)}{M^0(S_n)}.
\]
\[ T^{(1)}(S_n) \leq T^{(2)}(S_n) \leq \cdots \leq T^{(k)}(S_n) \leq \cdots \leq T^{(M)}(S_n) \]

**Definition 2 (RT under Weak Convergence)**

\[
\phi(S_n) \equiv \begin{cases} 
1 & T(S_n) > T^{(k)}(S_n) \\
\alpha(S_n) & T(S_n) = T^{(k)}(S_n), \quad \text{for} \quad \alpha(S_n) = \frac{M\alpha - M^+(S_n)}{M^0(S_n)} \\
0 & T(S_n) < T^{(k)}(S_n) \end{cases}
\]

Remarks:

- Exactly as before, but with \( S_n \) in place of \( X \)
- Test does not use the data beyond \( S_n \)
- No invariance assumption on \( P_n \)
Main Result

Theorem 2

If Assumption RW holds,

A. $T : S \to \mathbb{R}$ is continuous,

B. $g : S \to S$ is continuous, and

C. For any two distinct elements $g \in G$ and $g' \in G$,

either $T(gs) = T(g's) \forall s \in S$

or $\Pr(T(gs) \neq T(g's)) = 1$.

Then

$E_{P_n}[\phi(S_n)] \to \alpha$,

as $n \to \infty$ when $P_n \in P_{n,0}$ for all $n \geq 1$. 

Several Challenges:

1. Arguments valid for finite $n$ may not hold even approx. for large $n$.
   $$ T(S_n)|T^{(1)}(S_n), \ldots, T^{(M)}(S_n) \not\sim \text{Unif} \{T^{(1)}(S_n), \ldots, T^{(M)}(S_n)\} $$

2. Mimicking traditional proof doesn’t seem fruitful …
   … but will use it indirectly.

3. Earlier large-sample results for $S_n = \text{identity}$ not useful either.

4. Rely on almost sure representation theorem and $\neq$ arguments.
Several Challenges:

1. Arguments valid for finite \( n \) may not hold even approx. for large \( n \).
   
   \[ T(S_n) | T^{(1)}(S_n), \ldots, T^{(M)}(S_n) \not\sim \text{Unif}\{|T^{(1)}(S_n), \ldots, T^{(M)}(S_n)|\} \]

2. Mimicking traditional proof doesn’t seem fruitful …
   
   … but will use it indirectly.

3. Earlier large-sample results for \( S_n = \text{identity} \) not useful either.

4. Rely on almost sure representation theorem and \( \neq \) arguments.

Ties requirement satisfied for several statistics:

1. It holds for \( t \)-stat and absolute value of \( t \)-stat

2. It holds for wald-type test statistics
Outline

1. Review of Randomization Tests
   A) Symmetric Location Example

2. Randomization Tests under Approximate Symmetry
   A) Asymptotic Results

3. Applications
   A) Time Series Regression
   B) Difference in Differences


5. Conclusions
FRAMEWORK

\[ X^{(n)} \sim P_n : \\
H_0 : P_n \in P_{n,0} \]
$X^{(n)} \sim P_n :$

$H_0 : P_n \in P_{n,0}$

$X_1^{(n)} \sim P_{n,1}$

$X_2^{(n)} \sim P_{n,2}$

$X_3^{(n)} \sim P_{n,3}$

$X_4^{(n)} \sim P_{n,4}$

$X_q^{(n)} \sim P_{n,q}$

$(S_n, q) =\begin{cases} I & \text{if } T(S_n) > \text{cv}_n, 1 \\ g & \end{cases}$
\[ X^{(n)} \sim P_n : \\
H_0 : P_n \in P_{n,0} \]

**FRAMEWORK**

\[ X_1^{(n)} \sim P_{n,1} \]
\[ X_2^{(n)} \sim P_{n,2} \]
\[ X_3^{(n)} \sim P_{n,3} \]
\[ X_4^{(n)} \sim P_{n,4} \]

\[ S_{n,1}(X_{n,1}) \]
\[ S_{n,2}(X_{n,2}) \]
\[ S_{n,3}(X_{n,3}) \]
\[ S_{n,4}(X_{n,4}) \]
\[ S_{n,q}(X_{n,q}) \]
FRAMEWORK

\[ X^{(n)} \sim P_n : \]
\[ H_0 : P_n \in P_{n,0} \]

\[ S_{n,1}(X_{n,1}) \]
\[ S_{n,2}(X_{n,2}) \]
\[ S_{n,3}(X_{n,3}) \]
\[ S_{n,4}(X_{n,4}) \]
\[ S_{n,q}(X_{n,q}) \]

\[ \phi(S_n) = I\{T(S_n) > cv_{n,1-\alpha}\} \]
RELATE TO CLUSTERS

- Suppose $P_{n,0} = \{P_n \in \mathbb{P}_n : \theta(P_n) = \theta_0\}$.

- $X^{(n)}$ can be grouped into $q$ clusters: $X_j^{(n)}$.

- Compute $q$ estimators: $\hat{\theta}_{n,j} = \hat{\theta}_{n,j}(X_j^{(n)})$ for $1 \leq j \leq q$. 

Asymptotic Normality and sign-changes

Define $S_n = (S_{n,1}, \ldots, S_{n,q})$ with $S_{n,j} = \sqrt{n}(\hat{\theta}_{n,j} - 0)$.

Suppose $S_n \xrightarrow{d} N(0, \text{diag}(\Sigma_1, \ldots, \Sigma_q))$.

Special case of Assumption RW with $G = f_1, \ldots, g_q$.

Our results apply more generally (i.e. rates, distributions, etc).

Applications: specify $X^{(n)}_j$ and $\hat{\theta}_{n,j}$. Check convergence.
**Relate to Clusters**

- Suppose $P_{n,0} = \{P_n \in \mathcal{P}_n : \theta(P_n) = \theta_0\}$.

- $X^{(n)}$ can be grouped into $q$ clusters: $X^{(n)}_j$.

- Compute $q$ estimators: $\hat{\theta}_{n,j} = \hat{\theta}_{n,j}(X^{(n)}_j)$ for $1 \leq j \leq q$.

---

**Asymptotic Normality and sign-changes**

Define $S_n = (S_{n,1}, \ldots, S_{n,q})$ with

$$S_{n,j} = \sqrt{n}(\hat{\theta}_{n,j} - \theta_0).$$

Suppose

$$S_n \sim N(0, \text{diag}(\Sigma_1, \ldots, \Sigma_q)).$$

Special case of Assumption RW with $G = \{-1, 1\}^q$. 
**Relate to Clusters**

- Suppose \( P_{n,0} = \{P_n \in \mathcal{P}_n : \theta(P_n) = \theta_0 \} \).

- \( X^{(n)} \) can be grouped into \( q \) clusters: \( X^{(n)}_j \).

- Compute \( q \) estimators: \( \hat{\theta}_{n,j} = \hat{\theta}_{n,j}(X^{(n)}_j) \) for \( 1 \leq j \leq q \).

---

**Asymptotic Normality and Sign-changes**

Define \( S_n = (S_{n,1}, \ldots, S_{n,q}) \) with

\[
S_{n,j} = \sqrt{n}(\hat{\theta}_{n,j} - \theta_0).
\]

Suppose

\[
S_n \overset{\text{d}}{\rightarrow} N(0, \text{diag}(\Sigma_1, \ldots, \Sigma_q)).
\]

Special case of Assumption RW with \( G = \{-1, 1\}^q \).

- Our results apply more generally (i.e. rates, distributions, etc)

- **Applications**: specify \( X^{(n)}_j \) and \( \hat{\theta}_{n,j} \). Check convergence.
Application - Time Series Regression

Time series linear regression as in BCH.

\[ Y_t = \theta Z_t + \epsilon_t , \quad E[\epsilon_t Z_t] = 0 , \quad t = 1, \ldots, n . \]

Two DGPs (N and H) for \( Z_t \) and \( \epsilon_t \),

\[ Z_t = 1 + \rho Z_{t-1} + \nu_{1,t} , \]
\[ \epsilon_t = \rho \epsilon_{t-1} + \nu_{2,t} . \]

**Design N:** \( \nu_{1,t} \) and \( \nu_{2,t} \) are independent \( N(0, 1) \) random variables.
APPLICATION - TIME SERIES REGRESSION

Time series linear regression as in BCH.

\[ Y_t = \theta Z_t + \epsilon_t, \quad E[\epsilon_t Z_t] = 0, \quad t = 1, \ldots, n. \]

Two DGPs (N and H) for \( Z_t \) and \( \epsilon_t \),

\[ Z_t = 1 + \rho Z_{t-1} + \nu_{1,t}, \]
\[ \epsilon_t = \rho \epsilon_{t-1} + \nu_{2,t}. \]

**Design N:** \( \nu_{1,t} \) and \( \nu_{2,t} \) are independent \( N(0, 1) \) random variables.

**Design H:** \( \nu_{1,t} = a_t \tilde{\zeta}_{1,t} \) and \( \nu_{2,t} = b_t \tilde{\zeta}_{2,t} \), where

\[ \tilde{\zeta}_{k,t} \sim \frac{1}{3} N(-1, 1/2) + \frac{1}{3} N(0, 1/2) + \frac{1}{3} N(1, 1/2). \]

The constant \( a_t \) and \( b_t \) have a jump at \( t = n/2 \).
How to Apply our Method

- **Clusters:** $q$ blocks of consecutive obs. $\rightarrow X_j^{(n)}$

- **Estimators:** LS on each block $\rightarrow \hat{\theta}_{n,j}$

- Under weak assumptions,

$$S_n(X^{(n)}) = \sqrt{n}(\hat{\theta}_{n,j} - \theta_0, \ldots, \hat{\theta}_{n,q} - \theta_0)' \overset{d}{\rightarrow} N(0, \Sigma),$$

where $\Sigma$ is diagonal.

- See, e.g., Jenish & Prucha (2009) and BCH: holds for Models N and H.
How to Apply our Method

- **Clusters:** $q$ blocks of consecutive obs. $\rightarrow X_j^{(n)}$

- **Estimators:** LS on each block $\rightarrow \hat{\theta}_{n,j}$

- Under weak assumptions,

  $$S_n(X^{(n)}) = \sqrt{n} (\hat{\theta}_{n,j} - \theta_0, \ldots, \hat{\theta}_{n,q} - \theta_0)' \overset{d}{\rightarrow} N(0, \Sigma),$$

  where $\Sigma$ is diagonal.

- See, e.g., Jenish & Prucha (2009) and BCH: holds for Models N and H.

- **Our method:** $G = \{-1, 1\}^q$ and $T = |t\text{-stat}|$.

  $$T(S_n) = \frac{|\bar{\theta}_q - \theta_0|}{s_\theta / \sqrt{q}}, \quad \bar{\theta}_q = \frac{1}{q} \sum_{j=1}^{q} \hat{\theta}_{n,j}, \quad s_\theta^2 = \frac{1}{q-1} \sum_{j=1}^{q} (\hat{\theta}_{n,j} - \bar{\theta}_q)^2.$$

- **Note:** rate of convergence drops-out.
**ALTERNATIVE METHODS**

1. “Standard” approaches (large $q$)

   (I) Compute $\hat{\theta}_n^F$ with OLS using the **full sample**.

   (II) Reject for large values of

   \[
   \sqrt{n}\frac{|\hat{\theta}_n^F - \theta_0|}{\text{sandwich}}.
   \]

   Critical value depends on asym. framework.

   var. est. is consistent (Newey & West (1987), Andrews (1991))

   var. est. is inconsistent (Keifer & Vogelsang (2005))
**Alternative Methods**

1. “Standard” approaches (large $q$)
   
   (i) Compute $\hat{\theta}_n^F$ with OLS using the **full sample**.
   
   (ii) Reject for large values of
   
   $$\frac{\sqrt{n}|\hat{\theta}_n^F - \theta_0|}{\text{sandwich}}$$
   
   Critical value depends on asym. framework.
   
   var. est. is consistent (Newey & West (1987), Andrews (1991))
   
   var. est. is inconsistent (Keifer & Vogelsang (2005))

2. Bias Reduced Linearization (BRL)

   (i) **Unbiased** cluster covariance estimator (CCE) in “sandwich”.

   (ii) **Dof correction** to match first two moments of chi-square

   (Bell and McCaffrey, 2002, Imbens and Kolesar, 2012)
**ALTERNATIVE METHODS**

1. “Standard” approaches (large $q$)
   
   (i) Compute $\hat{\theta}_n^F$ with OLS using the **full sample**.

   (ii) Reject for large values of
   
   $$\frac{\sqrt{n} |\hat{\theta}_n^F - \theta_0|}{\text{sandwich}}.$$  

   Critical value depends on asym. framework.
   
   var. est. is consistent (Newey & West (1987), Andrews (1991))
   var. est. is inconsistent (Keifer & Vogelsang (2005))

2. Bias Reduced Linearization (BRL)

   (i) **Unbiased** cluster covariance estimator (CCE) in “sandwich”.

   (ii) **Dof correction** to match first two moments of chi-square
   
   (Bell and McCaffrey, 2002, Imbens and Kolesar, 2012)

3. Ibragimov & Müller (2010):

   Same test statistic we use here: *but* with cv from $t$—distribution.

   Same comments as before apply.
**Alternative Methods**

   
   (I) Cluster Covariance Estimator (CCE).
   
   (II) Reject for large values of
   
   $\sqrt{n} \left| \hat{\theta}_n^F - \theta_0 \right|$
   
   \[
   \frac{\text{sandwich}}{.}
   \]
   
   Critical value from $t$-dist. with $q - 1$ dof.
Alternative Methods

   (I) Cluster Covariance Estimator (CCE).
   (II) Reject for large values of

\[
\frac{\sqrt{n} | \hat{\theta}^F_n - \theta_0 |}{\text{sandwich}}.
\]

Critical value from \( t \)-dist. with \( q - 1 \) dof.

A) BCH: all conditions from IM are needed.

Our test: Keeps all advantages over IM here.

(I) Cluster Covariance Estimator (CCE).

(II) Reject for large values of

$$\frac{\sqrt{n}|\hat{\theta}_n^F - \theta_0|}{\text{sandwich}}.$$  

Critical value from $t$-dist. with $q - 1$ dof.

A) BCH: all conditions from IM are needed.

Our test: Keeps all advantages over IM here.

B) BCH allows heterogeneity on $z$ but needs $n^{-1}Z_jZ'_j \rightarrow^p \Gamma_j = \Gamma$.

Our test: Robust to $\Gamma_j \neq \Gamma_{ji}$. 

Alternative Methods
Alternative Methods

   (I) Cluster Covariance Estimator (CCE).
   (II) Reject for large values of

   $$\sqrt{n} |\hat{\theta}_n^F - \theta_0|$$

   sandwich

   Critical value from t-dist. with $q - 1$ dof.

   A) BCH: all conditions from IM are needed.

   **Our test:** Keeps all advantages over IM here.

   B) BCH allows heterogeneity on $z\epsilon$ but needs $n^{-1} Z_j Z_j' \rightarrow^p \Gamma_j = \Gamma$.

   **Our test:** Robust to $\Gamma_j \neq \Gamma_j'$.

   C) BCH: only for t-test - no easy F-test analog.

   **Our test:** multivariate parameters/ other test statistics.
   **Our test:** No “sandwich” estimator needed.
Size - Time Series Regression

\[ Y_t = \theta Z_t + \epsilon_t, \quad Z_t = 1 + \rho Z_{t-1} + \nu_t, \quad \epsilon_t = \rho \epsilon_{t-1} + u_t, \quad H_0 : \theta = 1. \]

<table>
<thead>
<tr>
<th>q</th>
<th>Rand</th>
<th>BCH</th>
<th>BRL</th>
<th>Rand</th>
<th>BCH</th>
<th>BRL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>5.0</td>
<td>5.2</td>
<td>5.2</td>
<td>5.1</td>
<td>5.1</td>
<td>5.3</td>
</tr>
<tr>
<td>0.5</td>
<td>5.4</td>
<td>6.1</td>
<td>8.2</td>
<td>18.1</td>
<td>17.6</td>
<td>18.8</td>
</tr>
<tr>
<td>0.8</td>
<td>5.8</td>
<td>8.2</td>
<td>16.2</td>
<td>4.9</td>
<td>4.9</td>
<td>5.0</td>
</tr>
<tr>
<td>0.95</td>
<td>5.4</td>
<td>5.2</td>
<td>7.5</td>
<td>4.9</td>
<td>5.3</td>
<td>5.8</td>
</tr>
</tbody>
</table>

TABLE: Null rejection probabilities: \( n = 100, \alpha = 5\% \), and \( M = \min\{1,000, 2^q\} \). 100,000 MC replications (se. \( \approx 0.0002 \)).

Note: BCH shown to perform better than stand. approach (HAC and KV).
**Power - Time Series Regression**

\[ Y_t = \theta Z_t + \epsilon_t, \quad Z_t = 1 + \rho Z_{t-1} + \upsilon_t, \quad \epsilon_t = \rho \epsilon_{t-1} + \upsilon_t, \quad H_0 : \theta = 1. \]

**Figure:** Size Adjusted Power Curves. **BCH-test** and **randomization test**. Parameter values are the following: Model A (left panel) and model B (right panel), \( \alpha = 0.05, \rho = 0.8, \) and 100,000 MC.
1. Review of Randomization Tests
   A) Symmetric Location Example

2. Randomization Tests under Approximate Symmetry
   A) Asymptotic Results

3. Applications
   A) Time Series Regression
   B) Difference in Differences


5. Conclusions
**Application - Difference in Differences**

Observed data \( X_{(n)} = \{(Y_{j,t}, D_{j,t}) : j \in J_0 \cup J_1, t \in T_0 \cup T_1\} \), where

\[
T_0 = \text{pre-treatment time periods} \\
T_1 = \text{post-treatment time periods} \\
J_0 = \text{control units} \\
J_1 = \text{treatment units}
\]

Model:

\[
Y_{j,t} = \theta D_{j,t} + \eta_j + \gamma_t + \epsilon_{j,t} \text{ with } E[\epsilon_{j,t}|(D_{j,t} : t \in T)] = 0
\]
APPLICATION - DIFFERENCE IN DIFFERENCES

Observed data $X_{(n)} = \{(Y_{j,t}, D_{j,t}) : j \in J_0 \cup J_1, t \in T_0 \cup T_1\}$, where

- $T_0 =$ pre-treatment time periods
- $T_1 =$ post-treatment time periods
- $J_0 =$ control units
- $J_1 =$ treatment units

Model:

$$Y_{j,t} = \theta D_{j,t} + \eta_j + \gamma_t + \epsilon_{j,t} \text{ with } E[\epsilon_{j,t} | (D_{j,t} : t \in T)] = 0$$

ASSUMPTION

1. $\min\{|T_0|, |T_1|\} \to \infty$ and $|J_0| \to \infty$
   - but $|J_1|$ is fixed
   - i.e., small number of treated units (as in Conley & Taber, 2011)

2. Processes $\{\epsilon_{j,t} : t \in T_0 \cup T_1\}$ indep. across $j$. 
How to Apply our Method

1. Clusters: $X_j^{(n)} = \{(Y_k,t, D_k,t) : k \in \{j\} \cup J_0, t \in T_0 \cup T_1\}$ for $j \in J_1$.

2. Estimators: $\hat{\theta}_{n,j}$ by LS with fixed effects using data $X_j^{(n)}$ for $j \in J_1$. 
How to Apply our Method

1. **Clusters:** \( X_j^{(n)} = \{(Y_k,t, D_k,t) : k \in \{j\} \cup J_0, t \in T_0 \cup T_1\} \) for \( j \in J_1 \).

2. **Estimators:** \( \hat{\theta}_{n,j} \) by LS with fixed effects using data \( X_j^{(n)} \) for \( j \in J_1 \).

3. Equivalent to

   \[
   \hat{\theta}_{n,j} = \Delta_{n,j} - \frac{1}{|J_0|} \sum_{k \in J_0} \Delta_{n,k},
   \]

   where

   \[
   \Delta_{n,j} = \frac{1}{|T_1|} \sum_{t \in T_1} Y_{j,t} - \frac{1}{|T_0|} \sum_{t \in T_0} Y_{j,t}.
   \]
How to Apply our Method

1. **Clusters:** \( X_j^{(n)} = \{(Y_k,t, D_k,t) : k \in \{j\} \cup J_0, t \in T_0 \cup T_1\} \) for \( j \in J_1 \).

2. **Estimators:** \( \hat{\theta}_{n,j} \) by LS with fixed effects using data \( X_j^{(n)} \) for \( j \in J_1 \).

3. Equivalent to

   \[
   \hat{\theta}_{n,j} = \Delta_{n,j} - \frac{1}{|J_0|} \sum_{k \in J_0} \Delta_{n,k},
   \]

   where

   \[
   \Delta_{n,j} = \frac{1}{|T_1|} \sum_{t \in T_1} Y_{j,t} - \frac{1}{|T_0|} \sum_{t \in T_0} Y_{j,t}.
   \]

4. Under weak assumptions,

   \[
   S_n(X^{(n)}) \equiv \sqrt{T}(\hat{\theta}_{n,j} - \theta_0 : j \in J_1) \rightsquigarrow N(0, \Sigma),
   \]

   where \( \Sigma \) is diagonal.

5. Proceed as before. **Note:** \(|J_1|\) is not a choice.
Our Method: Comments

- Straightforward to modify for indiv.-level data and/or covariates.

- Est. $\hat{\theta}_{n,j}$ are not independent.

- Req. $\min\{|T_0|, |T_1|\} \to \infty$ could be relaxed ...

  ... but only under stronger ass. on $\epsilon_{j,t}$

  ... implementation of test does not change!
Our Method: Comments

- Straightforward to modify for indiv.-level data and/or covariates.

- Est. $\hat{\theta}_{n,j}$ are not independent.

- Req. $\min\{|T_0|, |T_1|\} \to \infty$ could be relaxed ...

  ... but only under stronger ass. on $\epsilon_{j,t}$

  ... implementation of test does not change!

- Req. $|J_0| \to \infty$ could be relaxed...

  ... form disjoint pairs of units from $J_0$ and $J_1$. 

♣ ★
Concluding Remarks

- Developed theory of rand. tests when symmetry only holds approx.

- The method is widely applicable.
  - Time Series Regression.
  - Differences-in-Differences.
  - Clustered Regression.

- Revisited the empirical application in Angrist & Lavy (2009).

- Several advantages over existing methods (robustness and size control).

- Easily implemented in standard packages, like STATA.
References


Simulations - Dif in Dif

We simulate data as

\[ Y_{j,t} = \theta D_{j,t} + \beta Z_{j,t} + \epsilon_{j,t}, \]
\[ \epsilon_{j,t} = \rho \epsilon_{j,t-1} + \nu_{1,j,t}, \]
\[ Z_{j,t} = \gamma D_{j,t} + \nu_{2,j,t}, \quad \nu_{2,j,t} \sim N(0,1), \]
\[ D_{j,t} = I\{j \in J_1, \ t \geq t_j^*\} . \]

Base design: \(|J_0| + |J_1| = 100, \ q = |J_1| = 8, \ T = 10, \ t_j^* = \min\{2j, \ T\}, \)
\[ \rho = 0.5, \text{ and } \nu_{1,j,t} \sim N(0,1). \]
**Simulations - Dif in Dif**

We simulate data as

\[ Y_{j,t} = \theta D_{j,t} + \beta Z_{j,t} + \epsilon_{j,t}, \]

\[ \epsilon_{j,t} = \rho \epsilon_{j,t-1} + \nu_{1,j,t}, \]

\[ Z_{j,t} = \gamma D_{j,t} + \nu_{2,j,t}, \quad \nu_{2,j,t} \sim N(0, 1), \]

\[ D_{j,t} = I\{j \in J_1, \ t \geq t^*_j\}. \]

**Base design:** \(|J_0| + |J_1| = 100, \ q = |J_1| = 8, \ T = 10, \ t^*_j = \min\{2j, \ T\}, \)

\(\rho = 0.5, \) and \(\nu_{1,j,t} \sim N(0, 1).\)

**8 Additional designs:**

- (b) \(|J_0| + |J_1| = 50, \) (c) \(q = 12, \) (d) \(t^*_j = T/2\) for all \(j \in J_1,\)
- (e) \(\rho = 0.95, \) (f) \(T = 3,\)
- (f: conditional heteroskedastic)
  \(\nu_{1,j,t} \sim N(0, 4)\) for \(j \in J_1\) and \(\nu_{1,j,t} \sim N(0, 1)\) for \(j \in J_0,\)
- (g: heterogeneity in treatment units)
  \(\nu_{1,j,t} \sim N(0, 4)\) for \(j \leq 4\) and \(\nu_{1,j,t} \sim N(0, 1)\) for \(j > 4,\)
- (h: Asym. & Het in \(Z\))
  \(\nu_{1,j,t} \sim \text{Gamma}(0.25, 1),\) and \(\nu_{2,j,t}\) as \(\nu_{1,j,t}\) in (g).
\[ Y_{j,t} = \theta D_{j,t} + \beta Z_{j,t} + \epsilon_{j,t}, \]
\[ \epsilon_{j,t} = \rho \epsilon_{j,t-1} + \nu_{1,j,t}, \]
\[ Z_{j,t} = \gamma D_{j,t} + \nu_{2,j,t}, \quad \nu_{2,j,t} \sim N(0, 1). \]

<table>
<thead>
<tr>
<th>Spec.</th>
<th>Rejection probabilities under ( \theta = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rand</td>
</tr>
<tr>
<td>(a) Base</td>
<td>5.58</td>
</tr>
<tr>
<td>(b) ( n = 50 )</td>
<td>6.39</td>
</tr>
<tr>
<td>(c) ( q = 12 )</td>
<td>6.26</td>
</tr>
<tr>
<td>(d) ( t^* = T/2 )</td>
<td>5.56</td>
</tr>
<tr>
<td>(e) ( \rho = 0.95 )</td>
<td>6.06</td>
</tr>
<tr>
<td>(f) ( T = 3 )</td>
<td>5.41</td>
</tr>
<tr>
<td>(g) Cond. Heterosk.</td>
<td>4.78</td>
</tr>
<tr>
<td>(h) Het. in ( J_1 )</td>
<td>5.52</td>
</tr>
<tr>
<td>(i) Asym. &amp; Het in ( Z )</td>
<td>5.60</td>
</tr>
</tbody>
</table>

**Table:** Rejection rate (in %). Parameter values for the base design are \( |J_0| + |J_1| = 100, \ T = 10, \ q = 8, \ \rho = 0.5, \) and \( \alpha = 5\% \). Results based on 10,000 Monte Carlo replications.
Simulations - Dif in Dif: power

\[ Y_{j,t} = \theta D_{j,t} + \beta Z_{j,t} + \epsilon_{j,t}, \]
\[ \epsilon_{j,t} = \rho \epsilon_{j,t-1} + \nu_{1,j,t}, \]
\[ Z_{j,t} = \gamma D_{j,t} + \nu_{2,j,t}, \quad \nu_{2,j,t} \sim N(0, 1). \]

<table>
<thead>
<tr>
<th>Spec.</th>
<th>Rejection probabilities under ( \theta = 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rand</td>
</tr>
<tr>
<td>(a) Base</td>
<td>66.49</td>
</tr>
<tr>
<td>(b) ( n = 50 )</td>
<td>64.69</td>
</tr>
<tr>
<td>(c) ( q = 12 )</td>
<td>85.37</td>
</tr>
<tr>
<td>(d) ( t^* = T/2 )</td>
<td>69.44</td>
</tr>
<tr>
<td>(e) ( \rho = 0.95 )</td>
<td>32.29</td>
</tr>
<tr>
<td>(f) ( T = 3 )</td>
<td>59.06</td>
</tr>
<tr>
<td>(g) Cond. Heterosk.</td>
<td>9.54</td>
</tr>
<tr>
<td>(h) Het. in ( J_1 )</td>
<td>20.11</td>
</tr>
<tr>
<td>(i) Asym. &amp; Het in ( Z )</td>
<td>60.21</td>
</tr>
</tbody>
</table>

**TABLE:** Unadjusted power (in %). Parameter values for the base design are \( |J_0| + |J_1| = 100, \ T = 10, \ q = 8, \ \rho = 0.5, \) and \( \alpha = 5\%. \) Results based on 10,000 Monte Carlo replications.
1. Review of Randomization Tests
   A) Symmetric Location Example

2. Randomization Tests under Approximate Symmetry
   A) Asymptotic Results

3. Applications
   A) Time Series Regression
   B) Difference in Differences


5. Conclusions
**Empirical Application - Angrist & Lavy, 2009**

- **AL09 study effect of cash award on Bagrut:**
  - Bagrut: high school matriculation certificate awarded after a sequence of subject tests (Israel).
  - Greater weight given to exams in later years.
  - Formal prerequisite for university admission.

- **Achievement Award demonstration**
  - Project that provided cash award for low-achieving high school students.
Empirical Application - Angrist & Lavy, 2009

- **AL09 study effect of cash award on Bagrut:**
  - **Bagrut:** high school matriculation certificate awarded after a sequence of subject tests (Israel).
  - Greater weight given to exams in later years.
  - Formal prerequisite for university admission.

- **Achievement Award demonstration**
  - Project that provided cash award for low-achieving high school students.

- **Angrist and Lavy (2009):**
  - Find an increase in Bagrut rates in treated schools (mostly female students).
  - Driven largely by “marginal” female students.
Empirical Application - Angrist & Lavy, 2009 (cont)

Data: 40 high schools with the lowest 1999 Bagrut rates. All students in treated schools eligible for cash award.
Empirical Application - Angrist & Lavy, 2009 (cont)

Form pairs of schools that are similar ex-ante
20 Pairs: 1999 Bagrut rates used for matching
Empirical Application - Angrist & Lavy, 2009 (cont)

Treatment was assigned randomly in 2001 within pairs

Clusters: either a pair or two pairs
Empirical Application - Angrist & Lavy, 2009 (cont)

Let $i$ index students and $j$ index schools

$$E[Y_{ij} | .] = \Lambda[D_j \theta + Z_j \gamma + \sum_{k=1}^{3} d_{ki} \delta_k + W_{ij} \beta] ,$$

▶ Angrist and Lavy (2009):


▶ Clustered s.e. at school level (using Bias Reduced Linearization)
  - Do not use DoF correction to critical value.
Empirical Application - Angrist & Lavy, 2009 (cont)

Let $i$ index students and $j$ index schools

$$E[Y_{ij}|\cdot] = \Lambda[D_j \theta + Z_j \gamma + \sum_{k \leq 3} d_{ki} \delta_k + W_{ij} \beta],$$

- Angrist and Lavy (2009):
  - OLS and Logit: $Z_j$: school covariates. $d_{ki}$: lagged scores quartiles.
    $W_{ij}$: student covariates. $D_j$: treatment indicator.
  - Clustered s.e. at school level (using Bias Reduced Linearization)
    - Do not use DoF correction to critical value.

- Our Approach:
  1. Estimate $\theta$ for each cluster (1 or 2 pairs of schools)
     - cluster $\equiv$ school → does not identify $\theta$
     - cluster $\equiv$ pair → does not allow for $Z_j$ in some clusters
  2. Point estimator: average of $\hat{\theta}_{n,j}$
  3. Apply Theorem 2 with $G = \{-1, 1\}^{19}$ and $T = |t_{stat}|$
  4. Construct confidence intervals by inverting our test
**Empirical Application** - Angrist & Lavy, 2009 (cont)

Let $i$ index students and $j$ index schools

$$E[Y_{ij}|.] = \Lambda[D_j \theta + Z_j \gamma + \sum_{k \leq 3} d_{ki} \delta_k + W_{ij} \beta],$$

- **Angrist and Lavy (2009):**
  - Clustered s.e. at school level (using Bias Reduced Linearization)
    - Do not use DoF correction to critical value.

- **Our Approach:**
  1. Estimate $\theta$ for each cluster (1 or 2 pairs of schools)
    - cluster $\equiv$ school $\rightarrow$ does not identify $\theta$
    - cluster $\equiv$ pair $\rightarrow$ does not allow for $Z_j$ in some clusters
  2. Point estimator: average of $\hat{\theta}_{n,j}$
  3. Apply Theorem 2 with $G = \{-1, 1\}^{19}$ and $T = |t-stat|$
  4. Construct confidence intervals by inverting our test
Let $i$ index students and $j$ index schools

\[
E[Y_{ij} \mid \cdot] = \Lambda[D_j \theta + Z_j \gamma + \sum_{k \leq 3} d_{ki} \delta_k + W_{ij} \beta],
\]

### Treatment Effect: Boys & Girls

<table>
<thead>
<tr>
<th></th>
<th>Randomization Test</th>
<th>Angrist and Lavy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OLS</td>
<td>Logit</td>
</tr>
<tr>
<td>Sch. cov. only</td>
<td>0.049</td>
<td>-0.017</td>
</tr>
<tr>
<td>90%</td>
<td>[-0.078, 0.164]</td>
<td>[-0.147, 0.093]</td>
</tr>
<tr>
<td>95%</td>
<td>[-0.109, 0.182]</td>
<td>[-0.180, 0.105]</td>
</tr>
<tr>
<td>Lagged score, micro cov.</td>
<td>0.075</td>
<td>0.022</td>
</tr>
<tr>
<td>90%</td>
<td>[-0.034, 0.178]</td>
<td>[-0.058, 0.102]</td>
</tr>
<tr>
<td>95%</td>
<td>[-0.059, 0.198]</td>
<td>[-0.077, 0.117]</td>
</tr>
</tbody>
</table>

**TABLE:** Year 2001. Confidence interval for AL09 computed using their reported point estimates, the BRL standard errors, and a conventional standard normal critical value. Randomization test with $q = 11$, where the pairs of schools were clustered as follows: {1,3}, {2,4}, {5,8}, {7}, {9,10}, {11}, {12,13}, {14,15}, {16,17}, {18,20}, {19}. 

---

**Results: all students**
RESULTS: GIRLS ONLY

Let $i$ index students and $j$ index schools.

\[ E[Y_{ij} | \cdot] = \Lambda[D_j \theta + Z_j \gamma + \sum_{k \leq 3} d_{ki} \delta_k + W_{ij} \beta], \]

<table>
<thead>
<tr>
<th>Sch. cov. only</th>
<th>OLS</th>
<th>Logit</th>
<th>OLS</th>
<th>Logit</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>0.036</td>
<td>0.037</td>
<td>0.105</td>
<td>0.093</td>
</tr>
<tr>
<td>95%</td>
<td>[-0.132, 0.195]</td>
<td>[-0.099, 0.165]</td>
<td>[0.005, 0.205]</td>
<td>[0.006, 0.179]</td>
</tr>
<tr>
<td>90%</td>
<td>[-0.182, 0.234]</td>
<td>[-0.144, 0.183]</td>
<td>[-0.014, 0.224]</td>
<td>[-0.010, 0.197]</td>
</tr>
<tr>
<td>95%</td>
<td>0.090</td>
<td>0.058</td>
<td>0.105</td>
<td>0.097</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lagged score, micro cov.</th>
<th>OLS</th>
<th>Logit</th>
<th>OLS</th>
<th>Logit</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>[-0.049, 0.226]</td>
<td>[-0.020, 0.140]</td>
<td>[0.027, 0.182]</td>
<td>[0.021, 0.172]</td>
</tr>
<tr>
<td>95%</td>
<td>[-0.099, 0.256]</td>
<td>[-0.047, 0.157]</td>
<td>[0.012, 0.197]</td>
<td>[0.006, 0.187]</td>
</tr>
</tbody>
</table>

**TABLE:** Year 2001. Confidence interval for AL09 computed using their reported point estimates, the BRL standard errors, and a conventional standard normal critical value. Randomization test with $q = 9$, where the pairs of schools were clustered as follows: {1,3}, {16,4}, {5,7}, {2,12}, {10,11}, {8,19}, {13}, {14,15}, {18,20}. 
Results: Girls on Top of Cohort

AL09: largest effect on “marginal” female students.

- Use scores on tests prior to Jan. 2001.
- Use pred. prob. from Logit.

<table>
<thead>
<tr>
<th></th>
<th>Treatment Effect: Girls on top half of cohort</th>
<th>Sch. cov. only</th>
<th>Lagged s. or p. prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Randomization Test</td>
<td>90%</td>
<td>90%</td>
</tr>
<tr>
<td></td>
<td>by lagged score by pred. probability</td>
<td>[-0.077 , 0.259]</td>
<td>[-0.064 , 0.252]</td>
</tr>
<tr>
<td></td>
<td>by lagged score by pred. probability</td>
<td>0.089</td>
<td>0.091</td>
</tr>
<tr>
<td></td>
<td>Angrist and Lavy</td>
<td>0.081</td>
<td>0.076</td>
</tr>
<tr>
<td></td>
<td>by lagged score by pred. probability</td>
<td>0.206</td>
<td>0.213</td>
</tr>
<tr>
<td></td>
<td>by pred. probability</td>
<td></td>
<td>0.076</td>
</tr>
<tr>
<td></td>
<td>by pred. probability</td>
<td></td>
<td>0.207</td>
</tr>
</tbody>
</table>

90% [-0.077 , 0.259] [-0.064 , 0.252] [0.076 , 0.335] [0.067 , 0.320]
95% [-0.129 , 0.289] [-0.156 , 0.295] [0.051 , 0.360] [0.043 , 0.344]

Table: Results for Table 4 in AL: Top Girls only. Year 2001. Clusters as in previous table.
Proof of Theorem 1

Let \( P \in P_0 \) be given.
Proof of Theorem 1

Let \( P \in \mathbf{P}_0 \) be given. 

Note

\[
E_P \left[ \sum_{g \in G} \phi(gX) \right] = ME_P[\phi(X)].
\]
Proof of Theorem 1

Let $P \in P_0$ be given.

Note

$$E_P \left[ \sum_{g \in G} \phi(gX) \right] = ME_P[\phi(X)].$$

Since $T^{(m)}(x) = T^{(m)}(gx)$ for all $g \in G$ and $1 \leq j \leq M$, we have

$$\sum_{g \in G} \phi(gx) = M^+(x) + M^0(x)a(x) = M\alpha.$$  

Hence,

$$E_P \left[ \sum_{g \in G} \phi(gX) \right] = M\alpha.$$ 

The result follows. ■
**NORMAL EXAMPLE – POWER**

\[ X_j \sim N(0, 1) \text{ for } j \leq q/2 \text{ and } X_j \sim N(0, a^2) \text{ for } j > q/2 \]

t-test vs. rand. test using \( T = |t - \text{stat}| \)

**Figure:** Rejection rates. t-test and randomization test. Parameter values: \( q = 8 \) (left panel) and \( q = 16 \) (right panel), \( \alpha = 0.05 \), and 100,000 MC. ♢
Proof of Theorem 2

1. Let \( \{P_n \in P_{n,0} : n \geq 1\} \) be given.

By Almost Sure Rep. Thm., choose \( \tilde{S}_n, \tilde{S}, \) and \( U \sim U(0,1) \) s.t.

\[ \tilde{S}_n \to \tilde{S} \text{ w.p.1.} \]

with

\[ \tilde{S}_n \overset{d}{=} S_n, \tilde{S} \overset{d}{=} S, \text{ and } U \perp (\tilde{S}_n, \tilde{S}) \]

(on a common prob. space with prob. measure \( P \)).
Proof of Theorem 2

1. Let \( \{P_n \in P_{n,0} : n \geq 1\} \) be given.

By Almost Sure Rep. Thm., choose \( \tilde{S}_n, \tilde{S}, \) and \( U \sim U(0,1) \) s.t.

\[ \tilde{S}_n \to \tilde{S} \text{ w.p.1.} \]

with

\[ \tilde{S}_n \overset{d}{=} S_n, \tilde{S} \overset{d}{=} S, \text{ and } U \perp (\tilde{S}_n, \tilde{S}) \]

(on a common prob. space with prob. measure \( P \)).

2. Define

\[
\bar{\phi}(\tilde{S}_n, U) \equiv \begin{cases} 
1 & T(\tilde{S}_n) > T^{(k)}(\tilde{S}_n) \text{ or } T(\tilde{S}_n) = T^{(k)}(\tilde{S}_n) \text{ and } U < a(\tilde{S}_n) \\
0 & T(\tilde{S}_n) < T^{(k)}(\tilde{S}_n) 
\end{cases}
\]
Proof of Theorem 2 (cont.)

By construction,

\[ E_{P_n}[\phi(S_n)] = E_P[\tilde{\phi}(\tilde{S}_n, U)] . \]

Also, from thm. 1

\[ E_P[\tilde{\phi}(\tilde{S}, U)] = \alpha . \]

It is enough to show

\[ E_P[\tilde{\phi}(\tilde{S}_n, U)] \to E_P[\tilde{\phi}(\tilde{S}, U)] . \]
Proof of Theorem 2 (cont.)

3. Wish to show

\[ E_P[\tilde{\phi}(\tilde{S}_n, U)] \rightarrow E_P[\tilde{\phi}(\tilde{S}, U)] . \]

Let \( E_n \) be event where orderings of

\[ \{ T(g\tilde{S}) : g \in G \} \text{ and } \{ T(g\tilde{S}_n) : g \in G \} \]

are the same.

Under our assumptions,

\[ I\{E_n\} \rightarrow 1 \text{ w.p.1.} \]
Proof of Theorem 2 (cont.)

3. Wish to show

\[ E_P[\tilde{\phi}(\tilde{S}_n, U)] \rightarrow E_P[\tilde{\phi}(\tilde{S}, U)] . \]

Let \( E_n \) be event where orderings of

\[ \{ T(g\tilde{S}) : g \in G \} \quad \text{and} \quad \{ T(g\tilde{S}_n) : g \in G \} \]

are the same.

Under our assumptions,

\[ I\{E_n\} \rightarrow 1 \text{ w.p.1.} \]

On the event \( E_n \),

\[ \tilde{\phi}(\tilde{S}_n, U) = \tilde{\phi}(\tilde{S}, U) , \]

so

\[ E_P[\tilde{\phi}(\tilde{S}_n, U)I\{E_n\}] = E_P[\tilde{\phi}(\tilde{S}, U)I\{E_n\}] . \]
Proof of Theorem 2 (cont.)

3. Wish to show
\[ E_P[\mathcal{S}(\tilde{S}_n, U)] \rightarrow E_P[\mathcal{S}(\tilde{S}, U)] . \]

Let \( E_n \) be event where orderings of

\[ \{ T(g\tilde{S}) : g \in G \} \text{ and } \{ T(g\tilde{S}_n) : g \in G \} \]

are the same.

Under our assumptions,

\[ I\{E_n\} \rightarrow 1 \text{ w.p.1.} \]

On the event \( E_n \),

\[ \mathcal{S}(\tilde{S}_n, U) = \mathcal{S}(\tilde{S}, U) , \]

so

\[ E_P[\mathcal{S}(\tilde{S}_n, U)I\{E_n\}] = E_P[\mathcal{S}(\tilde{S}, U)I\{E_n\}] . \]

4. The desired result thus follows by Dom. Conv. \( \blacksquare \)