# Randomization tests of causal effects under interference between units

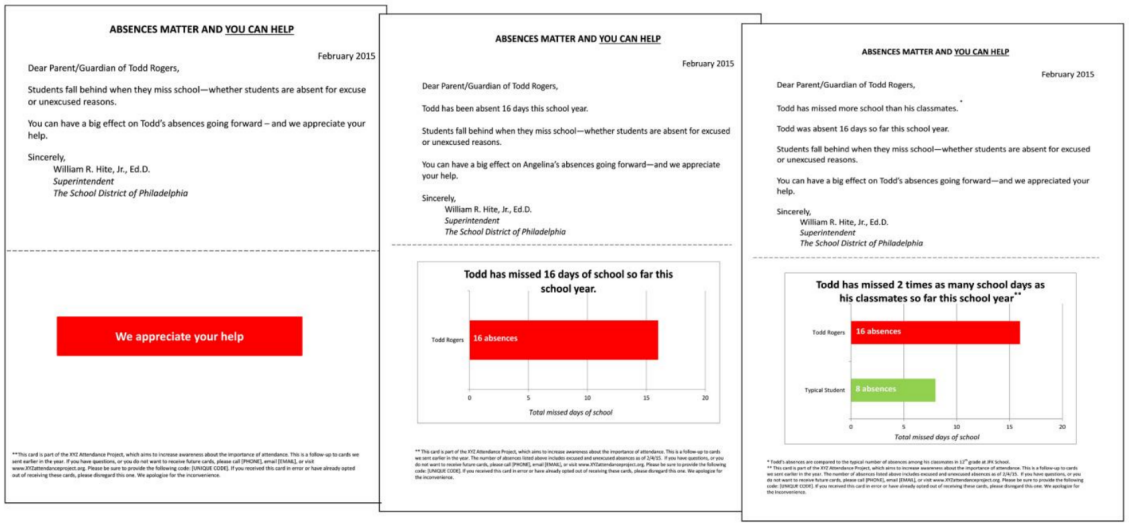
Panagiotis (Panos) Toulis
panos.toulis@chicagobooth.edu

Econometrics and Statistics
University of Chicago, Booth School of Business

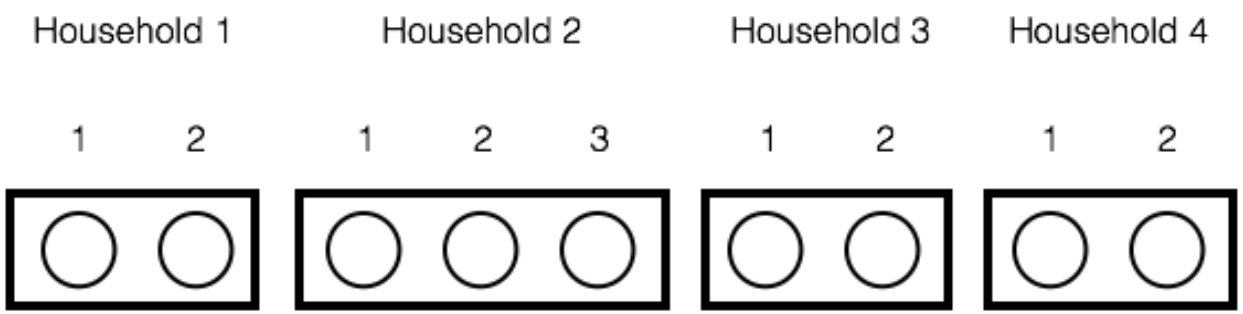
Joint work with Guillaume Basse (Harvard), Avi Feller (UC Berkeley)

# Motivation: reducing absenteeism at school

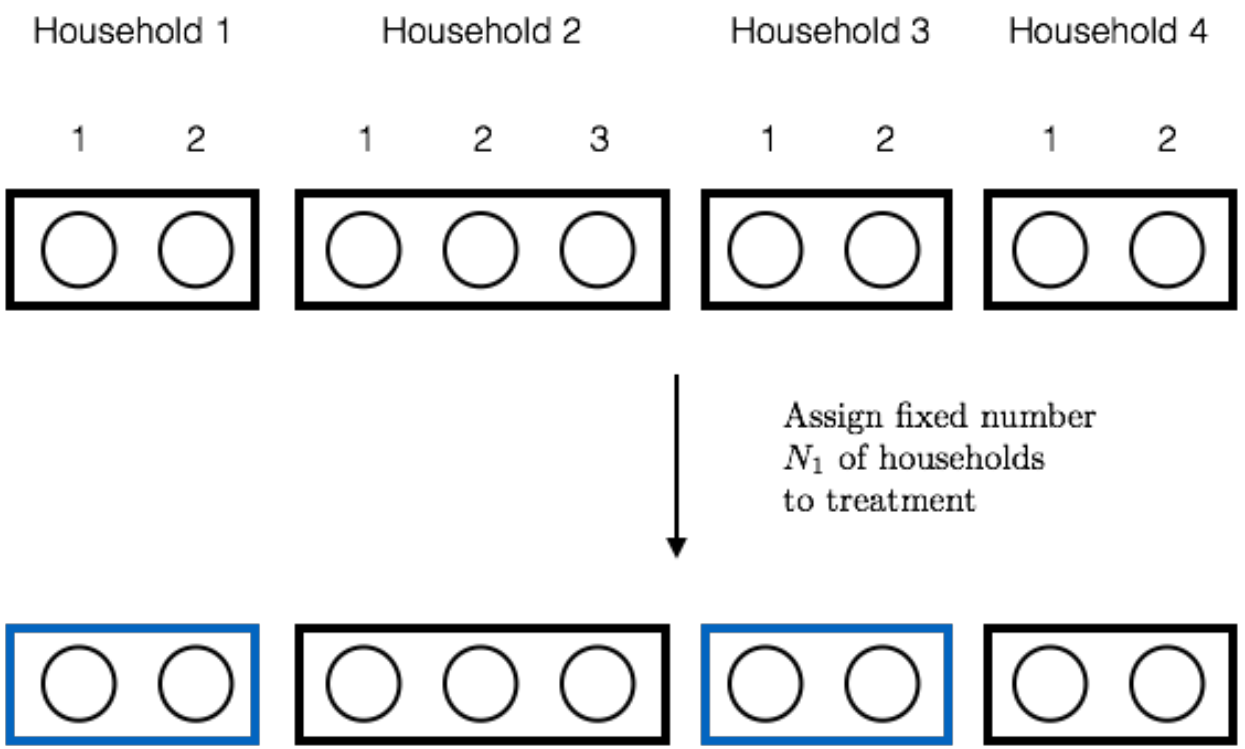
- Roger and Feller (2018) ran a two-stage randomized experiment, aiming to engage parents of students who were frequently absent.
- Data indicated strong *primary* effect for targeted student.
- Also interested in *spillovers* to siblings of the targeted student.



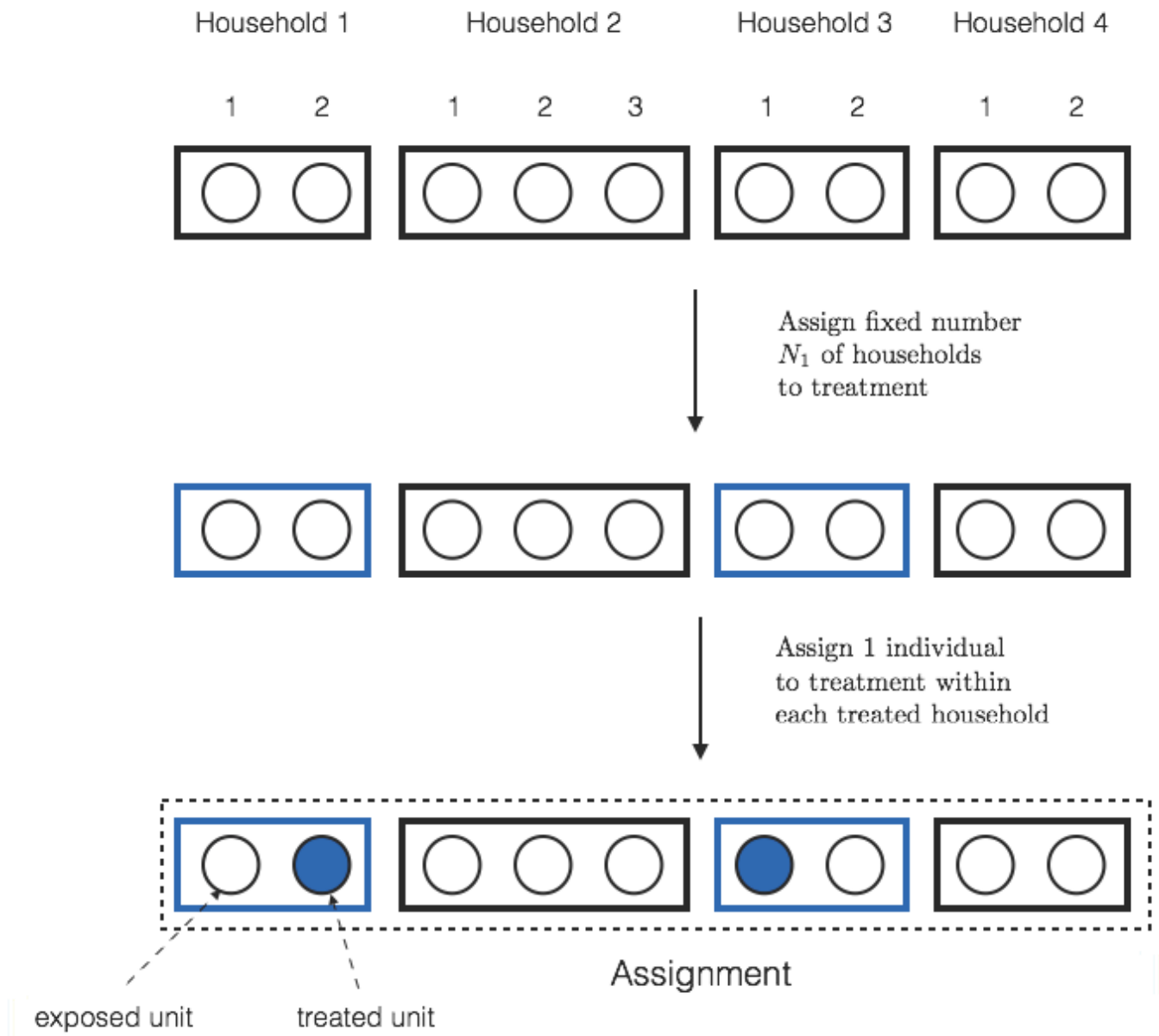
# Absenteeism design



# Absenteeism design



# Absenteeism design



# Outline

Goal: *Test for causal effects on treated and exposed units.*

1. Classical causal inference with no interference.
2. Challenges when interference is present.
3. Randomization tests with interference.
4. Application to absenteeism (and beyond).

# Classical causal inference — no interference

There are $N$ units. Unit $i$ is assigned to treatment $Z_i \in \{0, 1\}$.

$Z = (Z_1, \ldots, Z_N) \in \{0, 1\}^N$ is the full treatment vector.

$pr(Z) \in [0, 1)$ is the experimental design.

**No treatment interference** assumption (Cox, 1958):
$Y_i(1)$ is the potential outcome of $i$ when $Z_i = 1$ (treatment);
$Y_i(0)$ is its potential outcome when $Z_i = 0$ (control).

Causal effect of treatment on $i$ may be defined as

$$Y_i(1) - Y_i(0).$$

---

$*$ Only one potential outcome can ever be observed $\Rightarrow$ *fundamental problem of causal inference.*

# Two approaches to causal inference

Model-based inference; e.g.,

$$Y_i(Z_i) = \beta_0 + \beta_1 Z_i + \ldots + \varepsilon_i.$$

In experimental studies, $E(\hat{\beta}_1) = E(Y_i(1) - Y_i(0))$ — average treatment effect (ATE).

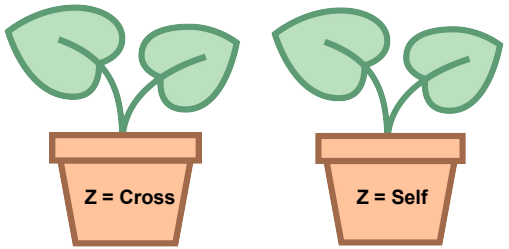
But standard errors **depend heavily** on the model.

In observational studies, most methods still rely on regression while trying to emulate experimental assignment of treatment (e.g., instrumental variables, propensity score matching).

In randomization inference, $Y$ is fixed, only $Z$ is random. Standard errors reflect *actual* variation from treatment assignment.

# Illustration

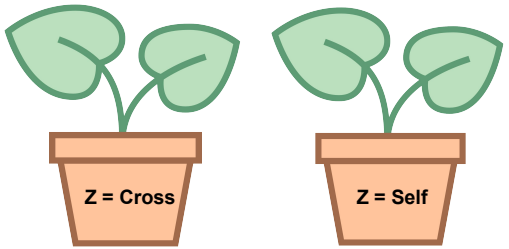
$Z_i \in \{\text{self-fertilized, cross-fertilized}\}$; $Y_i = \text{height in cm.}$




| unit $i$ | treat. $Z_i$ | obs. outcome $Y_i$ | potential outcomes $Y_i(0)$ | $Y_i(1)$ |
|---|---|---|---|---|
| 1 | 0 | 15 | 15 | ? |
| 2 | 1 | 20 | ? | 20 |
| 3 | 1 | $20+\epsilon$ | ? | $20+\epsilon$ |
| 4 | 0 | $15-\epsilon$ | $15-\epsilon$ | ? |

# Illustration

$Z_i \in \{\text{self-fertilized, cross-fertilized}\}$; $Y_i = \text{height in cm.}$




| unit $i$ | treat. $Z_i$ | obs. outcome $Y_i$ | potential outcomes $Y_i(0)$ | $Y_i(1)$ |
|---|---|---|---|---|
| 1 | 0 | 15 | 15 | ? |
| 2 | 1 | 20 | ? | 20 |
| 3 | 1 | $20+\epsilon$ | ? | $20+\epsilon$ |
| 4 | 0 | $15-\epsilon$ | $15-\epsilon$ | ? |

- Could estimate the causal effect through linear regression of $Y \sim Z$.
- Point estimate: $5+\epsilon$ cm (diff. in means).
- Standard error is $O(\epsilon)$. Possibly misleading as it depends on how well a line fits the data.

# Randomization inference (Fisher, 1935)

| unit | treat. | obs. outcome | potential outcomes | |
|---|---|---|---|---|
| $i$ | $Z_i$ | $Y_i$ | $Y_i(0)$ | $Y_i(1)$ |
| 1 | 0 | 15 | 15 | ? |
| 2 | 1 | 20 | ? | 20 |
| 3 | 1 | $20+\epsilon$ | ? | $20+\epsilon$ |
| 4 | 0 | $15-\epsilon$ | $15-\epsilon$ | ? |

- Fisher considered the null hypothesis of no treatment effect:

$$H_0 : Y_i(0) = Y_i(1).$$

# Randomization inference (Fisher, 1935)

| unit | treat. | obs. outcome | potential outcomes | |
|---|---|---|---|---|
| $i$ | $Z_i$ | $Y_i$ | $Y_i(0)$ | $Y_i(1)$ |
| 1 | 0 | 15 | 15 | 15 |
| 2 | 1 | 20 | 25 | 20 |
| 3 | 1 | $20+\epsilon$ | $20+\epsilon$ | $20+\epsilon$ |
| 4 | 0 | $15-\epsilon$ | $15-\epsilon$ | $15-\epsilon$ |

- Fisher considered the null hypothesis of no treatment effect:

$$H_0 : Y_i(0) = Y_i(1).$$

- Under $H_0$ the missing data can be filled in. Now, we can produce estimates of causal effect under *counterfactual* randomized assignments.

# Randomization inference (Fisher, 1935)

| unit | treat. | obs. outcome | potential outcomes | |
|---|---|---|---|---|
| $i$ | $Z'_i$ | $Y'_i$ | $Y_i(0)$ | $Y_i(1)$ |
| 1 | 1 | 15 | 15 | 15 |
| 2 | 0 | 20 | 25 | 20 |
| 3 | 1 | $20+\epsilon$ | $20+\epsilon$ | $20+\epsilon$ |
| 4 | 0 | $15-\epsilon$ | $15-\epsilon$ | $15-\epsilon$ |

e.g., say $Z' = (1, 0, 1, 0)$ as counterfactual assignment. Counterfactual outcomes $Y'$ can be calculated under $H_0$. In fact, $Y' = Y$. Regress $Y' \sim Z'$, repeat..

# Randomization inference (Fisher, 1935)

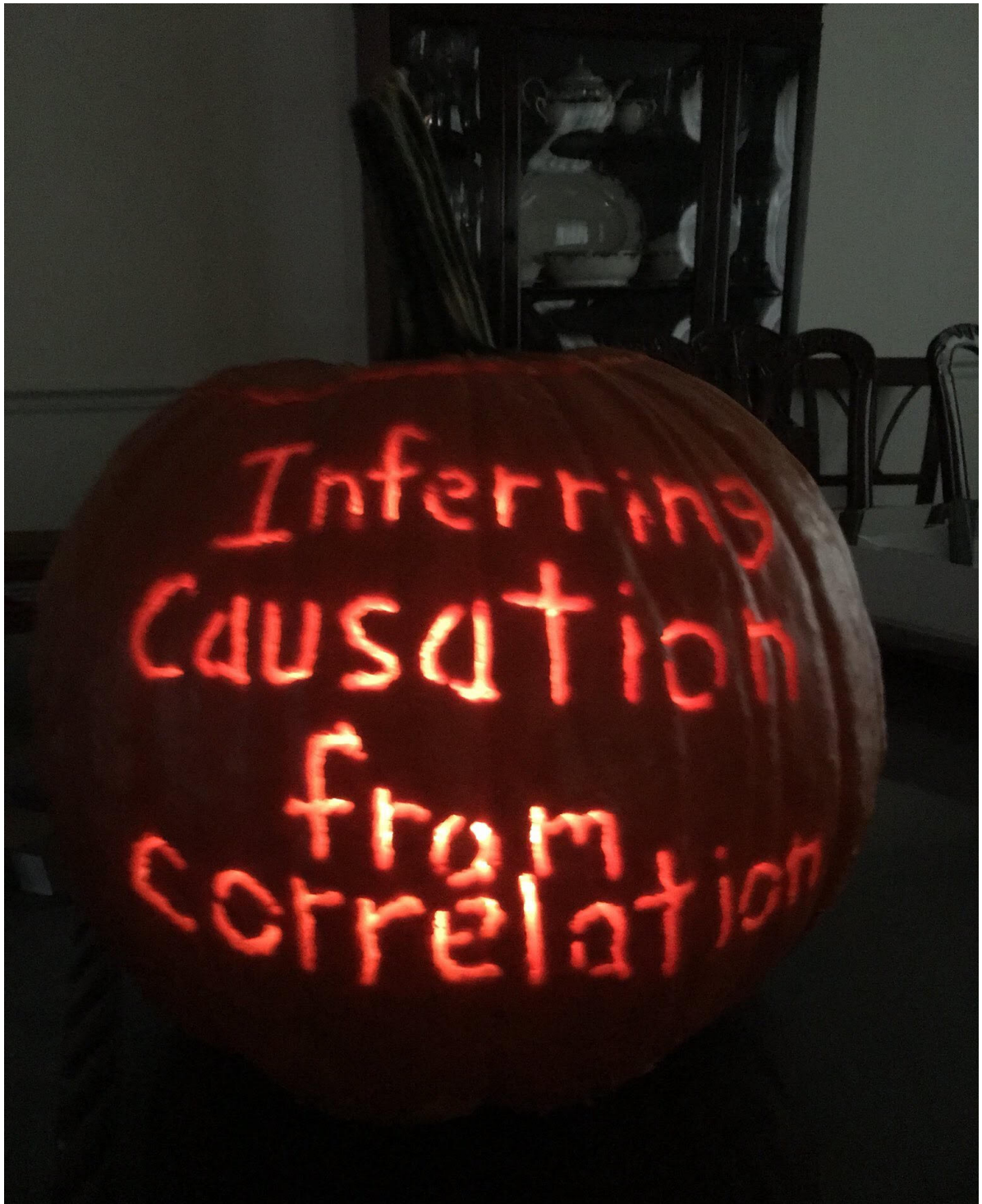
| unit | treat. | obs. outcome | potential outcomes | |
|---|---|---|---|---|
| $i$ | $Z_i'$ | $Y_i'$ | $Y_i(0)$ | $Y_i(1)$ |
| 1 | 1 | 15 | 15 | 15 |
| 2 | 0 | 20 | 25 | 20 |
| 3 | 1 | $20+\epsilon$ | $20+\epsilon$ | $20+\epsilon$ |
| 4 | 0 | $15-\epsilon$ | $15-\epsilon$ | $15-\epsilon$ |

e.g., say $Z' = (1, 0, 1, 0)$ as counterfactual assignment. Counterfactual outcomes $Y'$ can be calculated under $H_0$. In fact, $Y' = Y$. Regress $Y' \sim Z'$, repeat..

*Randomization distribution* of regression estimate:

| -(5+e) | -e | 0 | +e | (5+e) |
|---|---|---|---|---|
| 16.7% | 16.7% | 33.3% | 16.7% | 16.7% |

So, $-(5+\epsilon)$ is equally likely as $5+\epsilon$. Standard error = const. $+ O(\epsilon) > O(\epsilon)$.


Inferring
causation
from
correlation

# Fisher exact test

1. Pick a test statistic $T$ (e.g., diff. in means).
2. Observed value $T^{\text{obs}} = T(Z, Y)$.
3. For $r = 1, 2, \ldots, M$:
   (i) Sample $Z' \sim pr(Z')$ according to design.
   (ii) Store $T_r = T(Z', Y)$.

Summarize: e.g., $p$-value $= E\left(\mathbb{I}\{T_r \geq T^{\text{obs}}\}\right)$.

# Fisher exact test

1. Pick a test statistic $T$ (e.g., diff. in means).
2. Observed value $T^{\text{obs}} = T(Z, Y)$.
3. For $r = 1, 2, \ldots, M$:
   (i) Sample $Z' \sim pr(Z')$ according to design.
   (ii) Store $T_r = T(Z', Y)$.

Summarize: e.g., $p$-value $= E\left(\mathbb{I}\{T_r \geq T^{\text{obs}}\}\right)$.


Notes:

- **Finite-sample validity:** we need $T(Z', Y') \stackrel{d}{=} T(Z, Y)$.
  For any fixed $Y$,
  $$T(Z', Y') \stackrel{H_0}{=} T(Z', Y) \stackrel{d}{=} T(Z, Y).$$
- No model on $Y$ is necessary. No asymptotic/normal/other approximations.
- **Robust** to any order-preserving transformations of $Y$.
- Randomness comes only from design (which we control).

# Fisher exact test

1. Pick a test statistic $T$ (e.g., diff. in means).
2. Observed value $T^{\text{obs}} = T(Z, Y)$.
3. For $r = 1, 2, \ldots, M$:
   (i) Sample $Z' \sim pr(Z')$ according to design.
   (ii) Store $T_r = T(Z', Y)$.

Summarize: e.g., $p$-value $= E\left(\mathbb{I}\{T_r \geq T^{\text{obs}}\}\right)$.


Notes:

- **Finite-sample validity:** we need $T(Z', Y') \stackrel{d}{=} T(Z, Y)$.
  For any fixed $Y$,
  $$T(Z', Y') \stackrel{H_0}{=} T(Z', Y) \stackrel{d}{=} T(Z, Y).$$
- No model on $Y$ is necessary. No asymptotic/normal/other approximations.
- **Robust** to any order-preserving transformations of $Y$.
- Randomness comes only from design (which we control).

Next: Fisher exact tests with interference?

# Interference

There is *interference* when the outcome of a unit may be affected by the treatment assignment of others.

Unlikely that interference is not there: (social) networks between people, firms, schools, etc., enable interference.

Peer effects, contagion, spillovers.., are special cases.

# Causal inference under interference

New notation: $Y_i(Z)$ is the potential outcome of unit $i$ under full treatment assignment vector $Z$.

There are $2^N$ possible potential outcomes since $Z \in \{0,1\}^N$. Causal inference is impossible without *stability* assumptions; e.g., "Stable treatment unit value assumption" (Rubin, 1974)

$$Y_i(Z) = Y_i(Z') \text{ if } Z_i = Z_i';$$

$\Rightarrow$ implies no interference.

Randomization inference is tricky: $H_0 : Y_i(...) = Y_i(...)$ ?

Although unit receives treatment $Z_i$, it is *exposed* to a more complex treatment version. Need to define what this exposure is.

# Exposures

We don't believe that every different $Z$ implies a different exposure for a particular unit. More reasonable that there are equivalence classes.

# Exposures

We don't believe that every different $Z$ implies a different exposure for a particular unit. More reasonable that there are equivalence classes.

An *exposure mapping* describes such equivalence through a function $h_i(Z) : \{0,1\}^N \to \mathcal{E}$, where $\mathcal{E}$ is a finite set of possible exposures;

e.g., if a network $G$ is present we could define $h_i(Z) = (Z_i, G_i^\top Z)$, where $G_i^\top Z = \#\ i$'s treated neighbors under $Z$ (Toulis & Kao, 2013).

# Exposures in absenteeism study

Assumption 1: No interference between households.

Assumption 2: In a treated household, the identity of the treated unit does not matter (Manski's "anonymous interactions").

# Exposures in absenteeism study

Assumption 1: No interference between households.

Assumption 2: In a treated household, the identity of the treated unit does not matter (Manski's "anonymous interactions").

Assumptions 1 + 2 imply the exposure mapping:

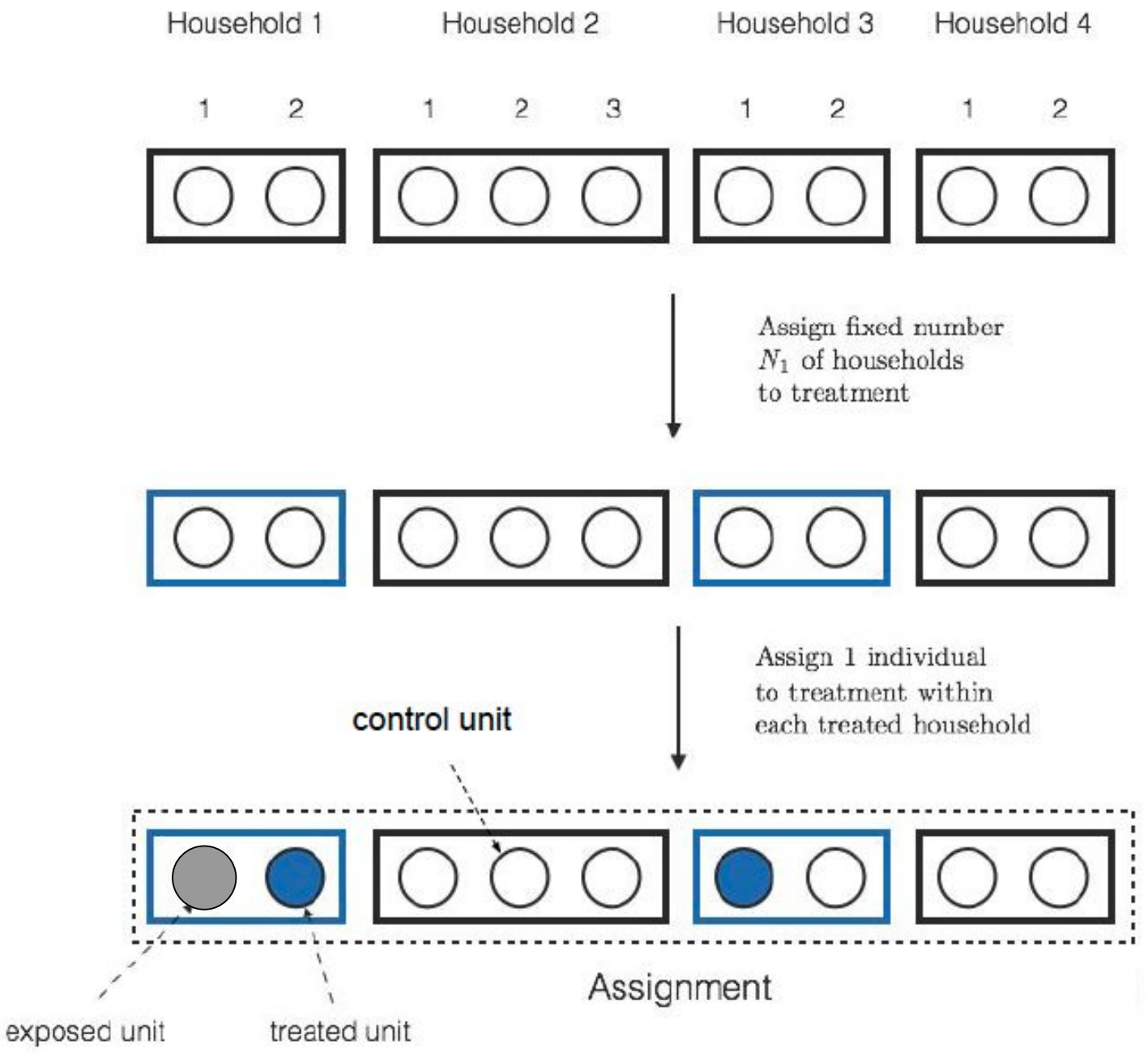
$$h_i(Z) = (H_{j_i}, Z_i), \text{ where } R_{ij_i} = 1^*.$$

$*$ $R_{ij}$ = residence index for unit $i$ in household $j$; and $H_j = \sum_i Z_i R_{ij}$ = household treatment.

**Three** potential outcomes:

- $Y_i(0,0) \equiv Y_i(\texttt{control})$ = control unit in control household;
- $Y_i(1,0) \equiv Y_i(\texttt{exposed})$ = control unit in treated household;
- $Y_i(1,1) \equiv Y_i(\texttt{treated})$ = treated unit in treated household.

# Exposures in absenteeism study



# Hypotheses for randomization tests

These exposure mappings now enable definition and testing of causal effects in the absenteeism study.

# Hypotheses for randomization tests

These exposure mappings now enable definition and testing of causal effects in the absenteeism study.

**No effects whatsoever:** (i.e., *control = exposed = treated)*

$$Y_i(0,0) = Y_i(1,0) = Y_i(1,1).$$

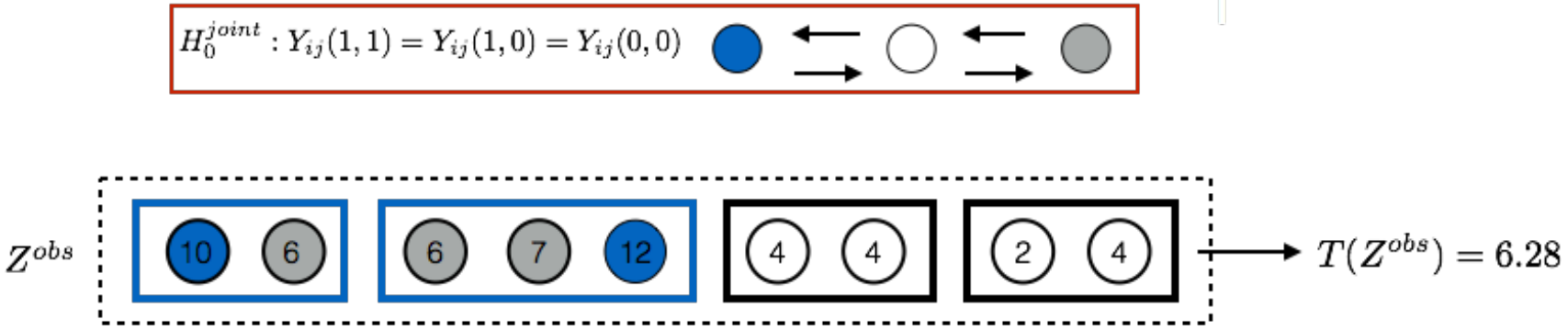
**No primary effect:** (i.e., *control = treated)*

$$Y_i(0,0) = Y_i(1,1).$$

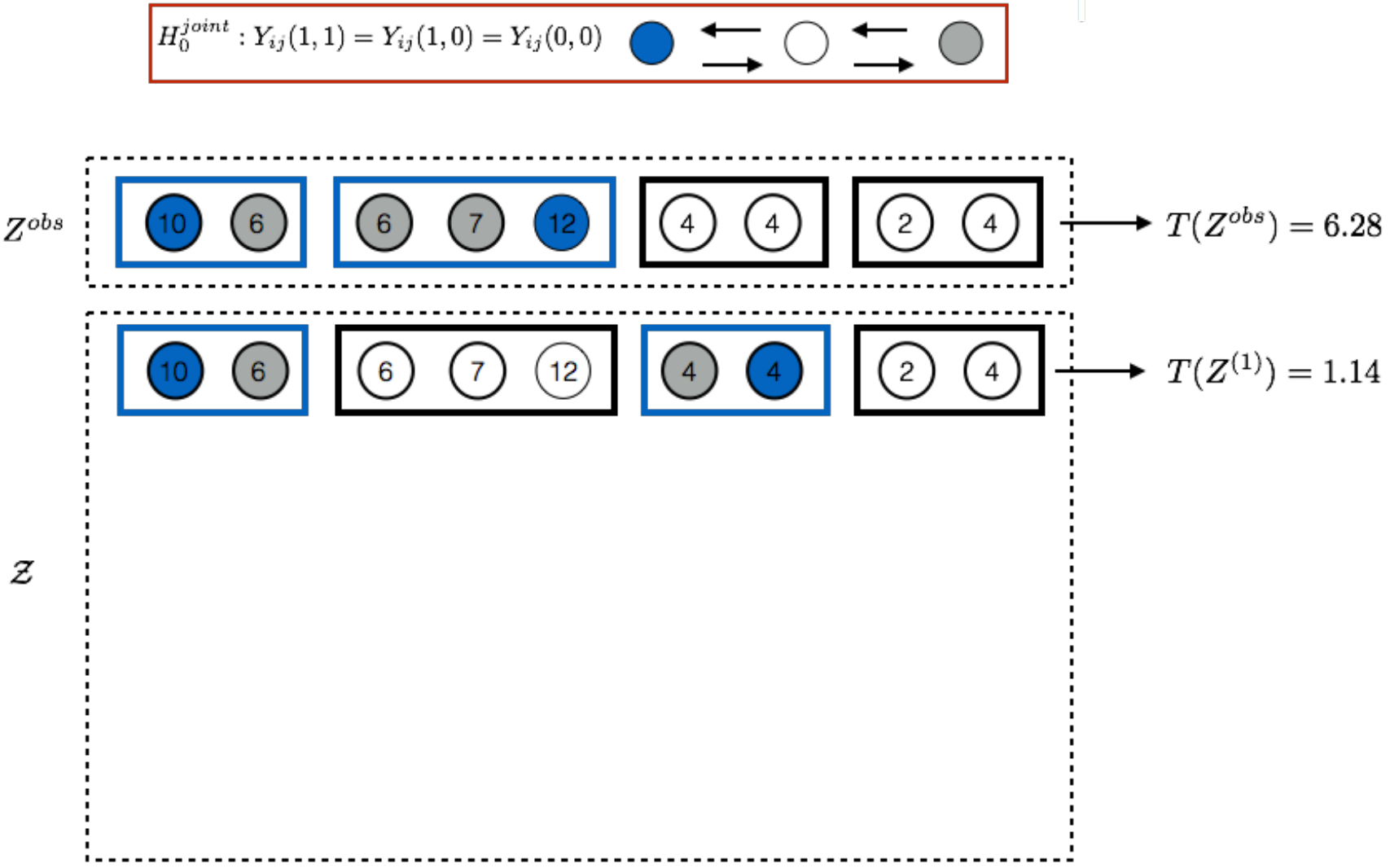
**No spillover effect:** (i.e., *control = exposed)*

$$Y_i(0,0) = Y_i(1,0).$$

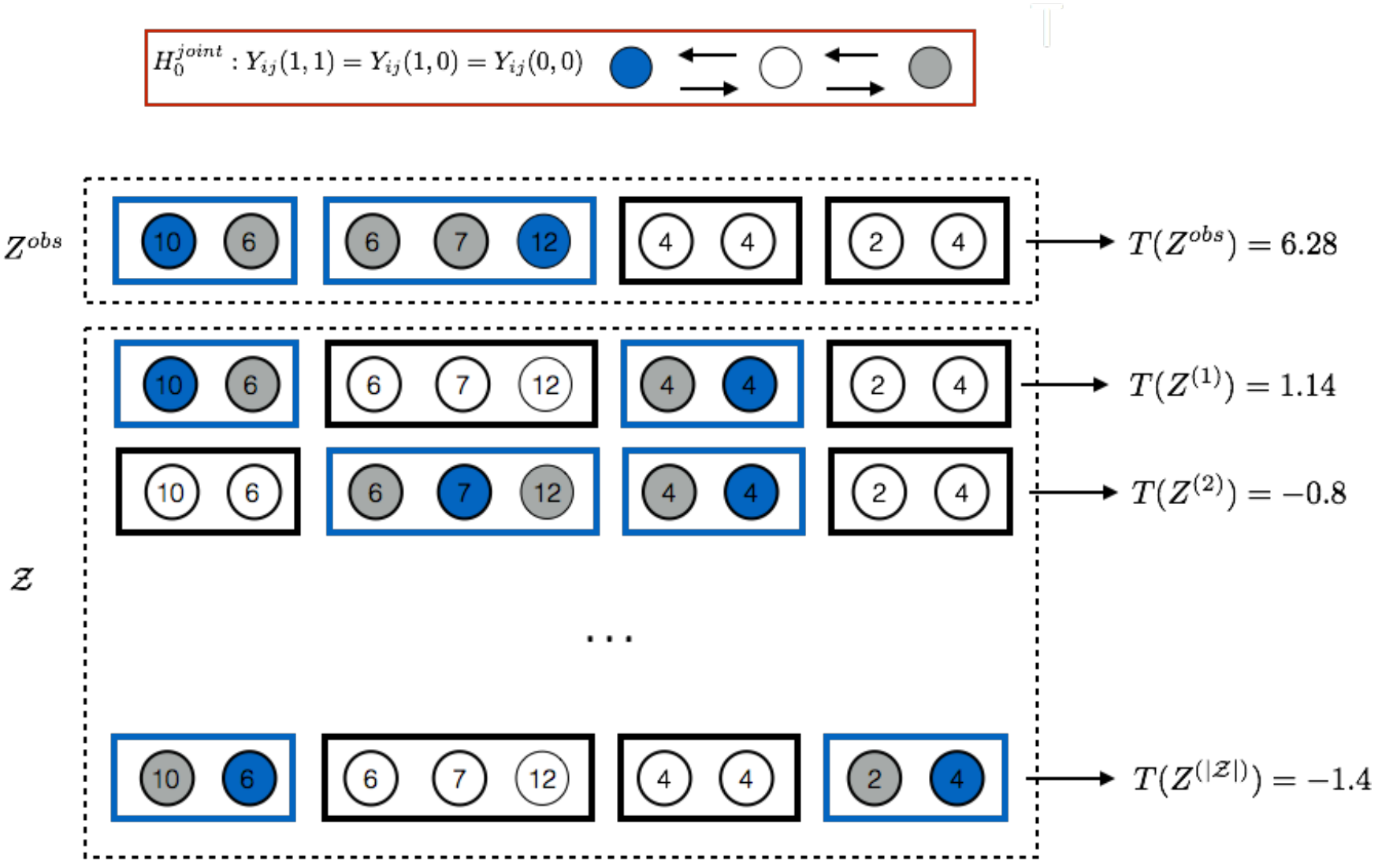
# Fisher test for “no effect whatsoever”



# Fisher test for “no effect whatsoever”

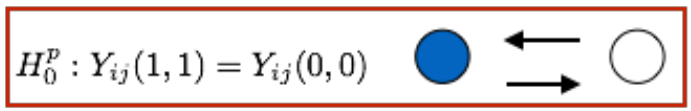


# Fisher test for “no effect whatsoever”

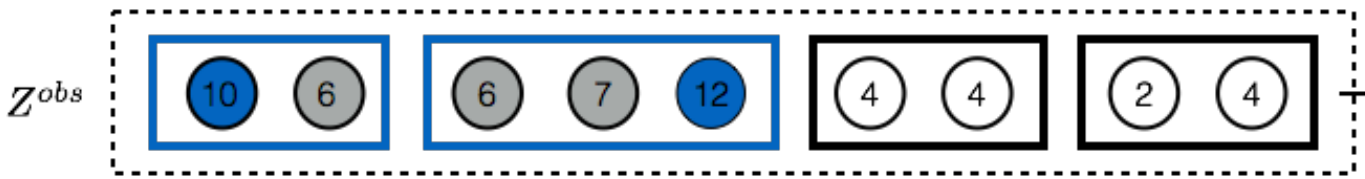



∗ Works without problem but null is strong.

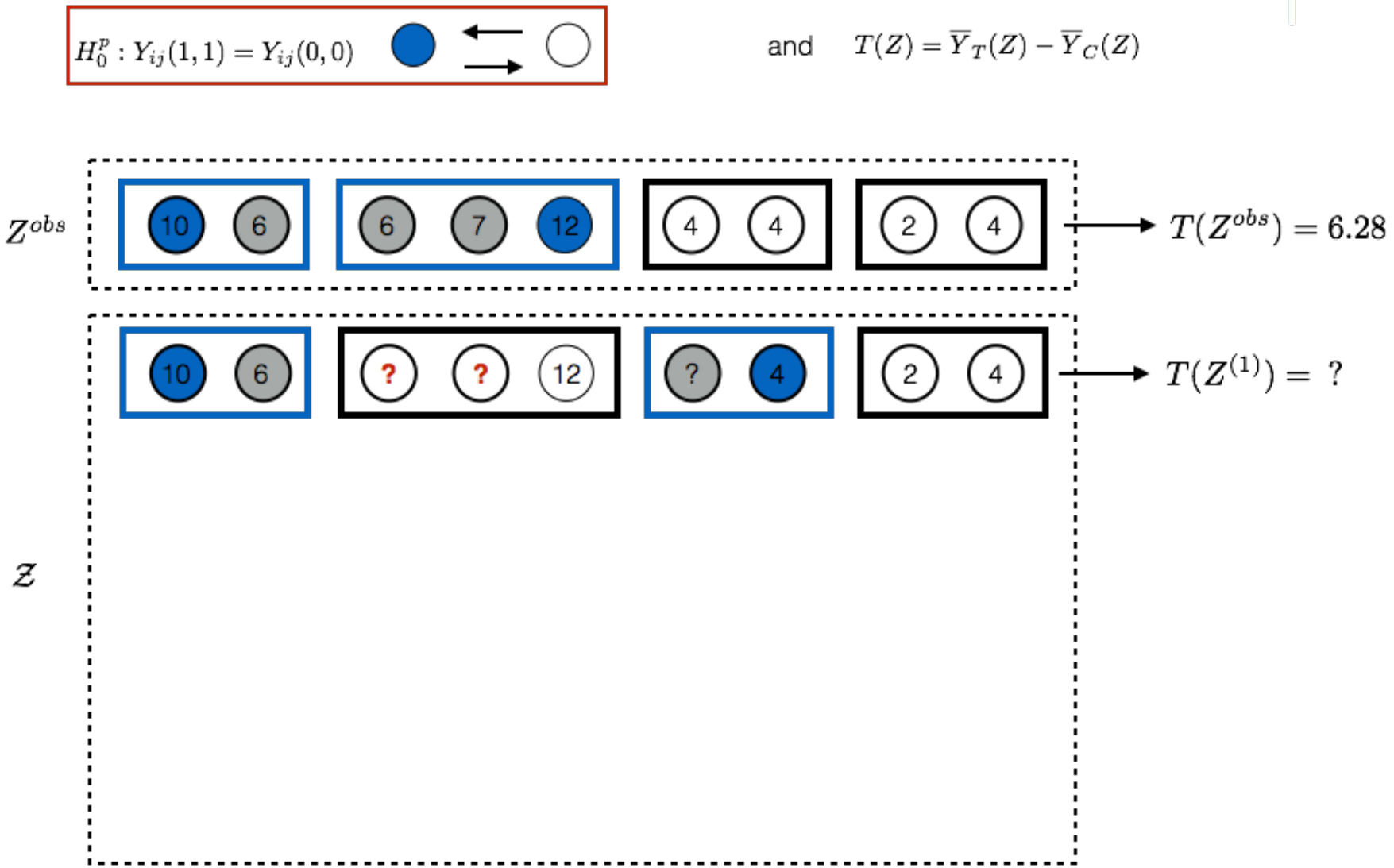
# Fisher test for primary effect



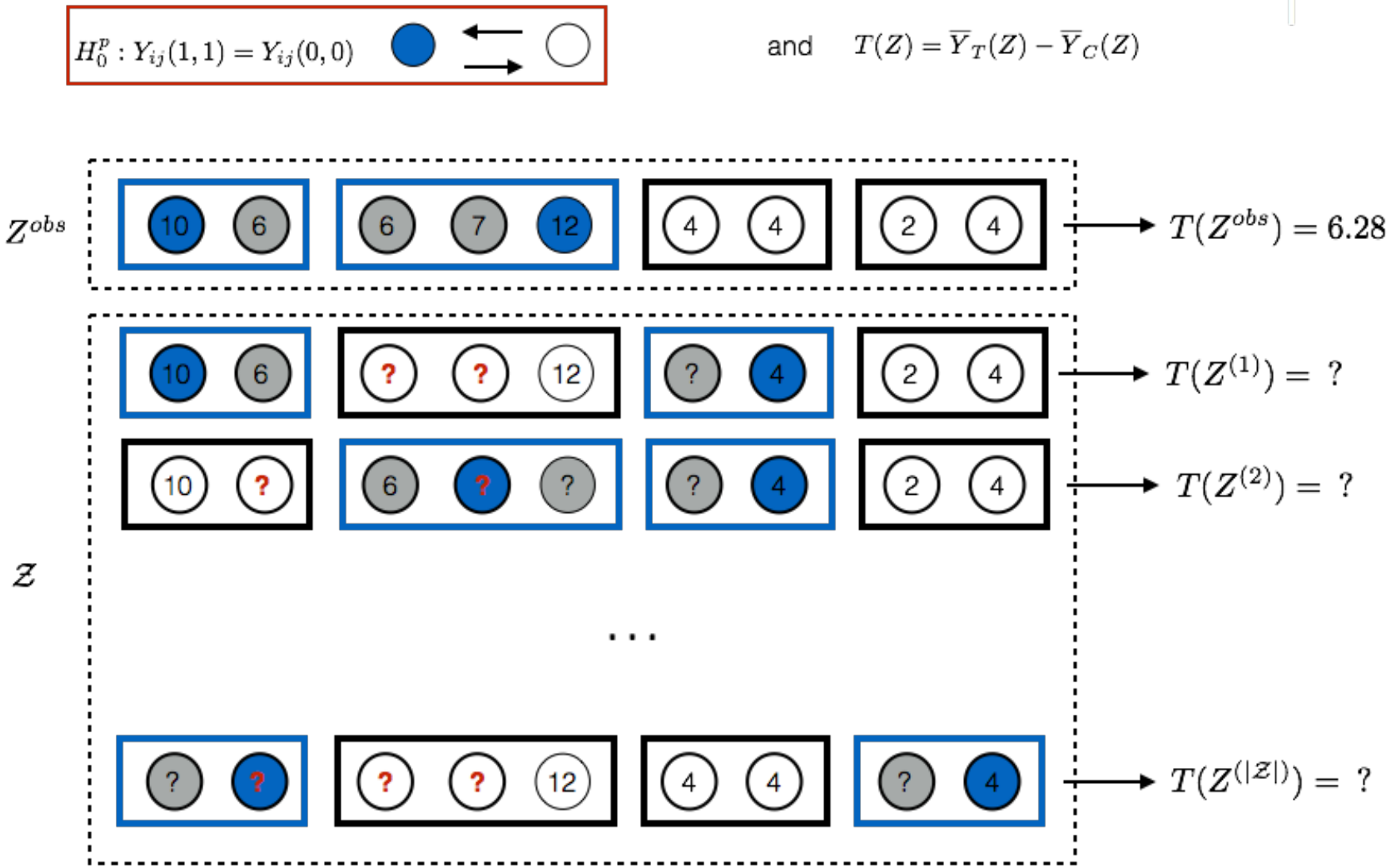

and $T(Z) = \overline{Y}_T(Z) - \overline{Y}_C(Z)$



# Fisher test for primary effect



# Fisher test for primary effect




pval = ?

# Problem with interference

Under interference, there are *multiple* potential exposures but null hypothesis is about only *two* of them (usually).

Thus, null hypotheses are no longer **sharp** $\Rightarrow$ cannot impute potential outcomes (i.e., can't do $Y' = Y$ any more.)

# Problem with interference

Under interference, there are *multiple* potential exposures but null hypothesis is about only *two* of them (usually).

Thus, null hypotheses are no longer **sharp** $\Rightarrow$ cannot impute potential outcomes (i.e., can't do $Y' = Y$ any more.)

**Consequence:** If imputation is not possible either drop units or assignments.

# Problem with interference

Under interference, there are *multiple* potential exposures but null hypothesis is about only *two* of them (usually).

Thus, null hypotheses are no longer **sharp** $\Rightarrow$ cannot impute potential outcomes (i.e., can't do $Y' = Y$ any more.)

**Consequence:** If imputation is not possible either drop units or assignments.

The idea is to perform the test **conditional** on subsets of units and assignments such that $H_0$ is **conditionally sharp** (Aronow, 2012; Athey et. al., 2017).

# Conditional randomization tests under interference

1. Sample a (sub)set of units $F \subset \{1, \ldots, N\}$ independently of $Z$. Known as *focal units*.
2. Define test statistic $T$ only on units in $F$.
3. Compute set of assignments for which $H_0$ is sharp:

$$\mathcal{Z}_F = \left\{ Z' : Y_i(Z') \overset{H_0}{=} Y_i(Z), \text{ for every } i \in F \right\}.$$

4. Run Fisher test sampling *uniformly from* $\mathcal{Z}_F$.

# Conditional randomization tests under interference

1. Sample a (sub)set of units $F \subset \{1, \ldots, N\}$ independently of $Z$. Known as *focal units*.
2. Define test statistic $T$ only on units in $F$.
3. Compute set of assignments for which $H_0$ is sharp:

$$\mathcal{Z}_F = \left\{ Z' : Y_i(Z') \overset{H_0}{=} Y_i(Z), \text{ for every } i \in F \right\}.$$

4. Run Fisher test sampling *uniformly from* $\mathcal{Z}_F$.

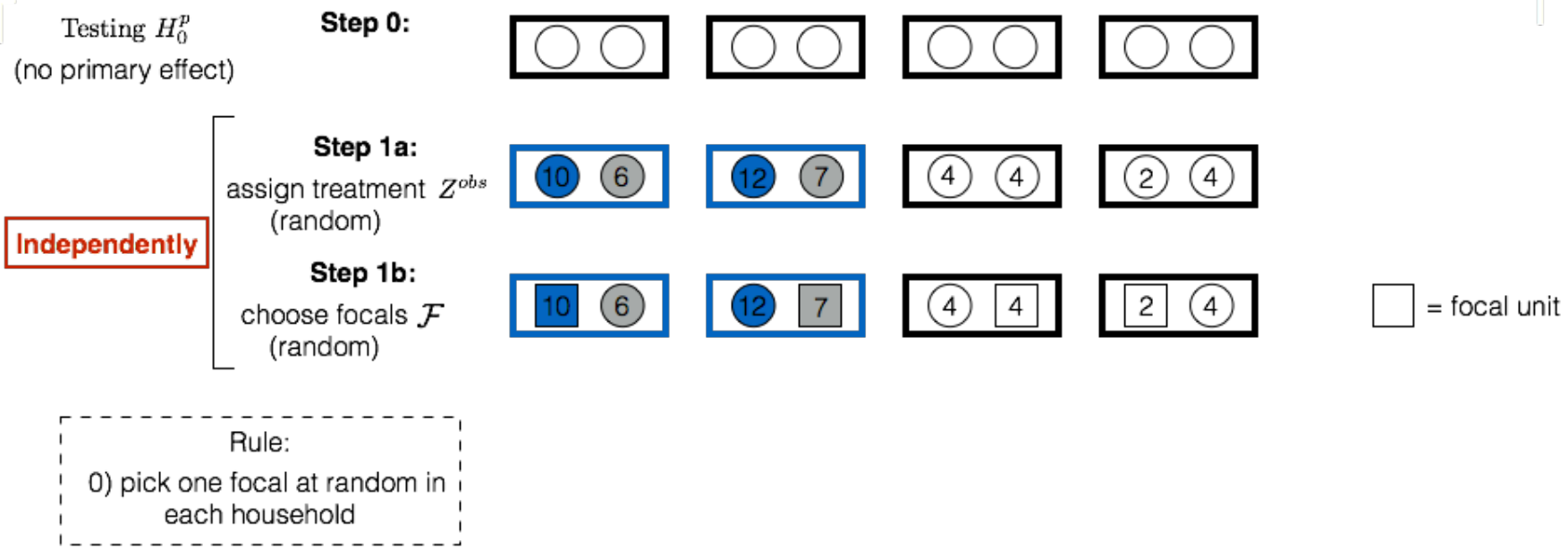
---

$*$ Test is straightforward (conceptually) and valid.

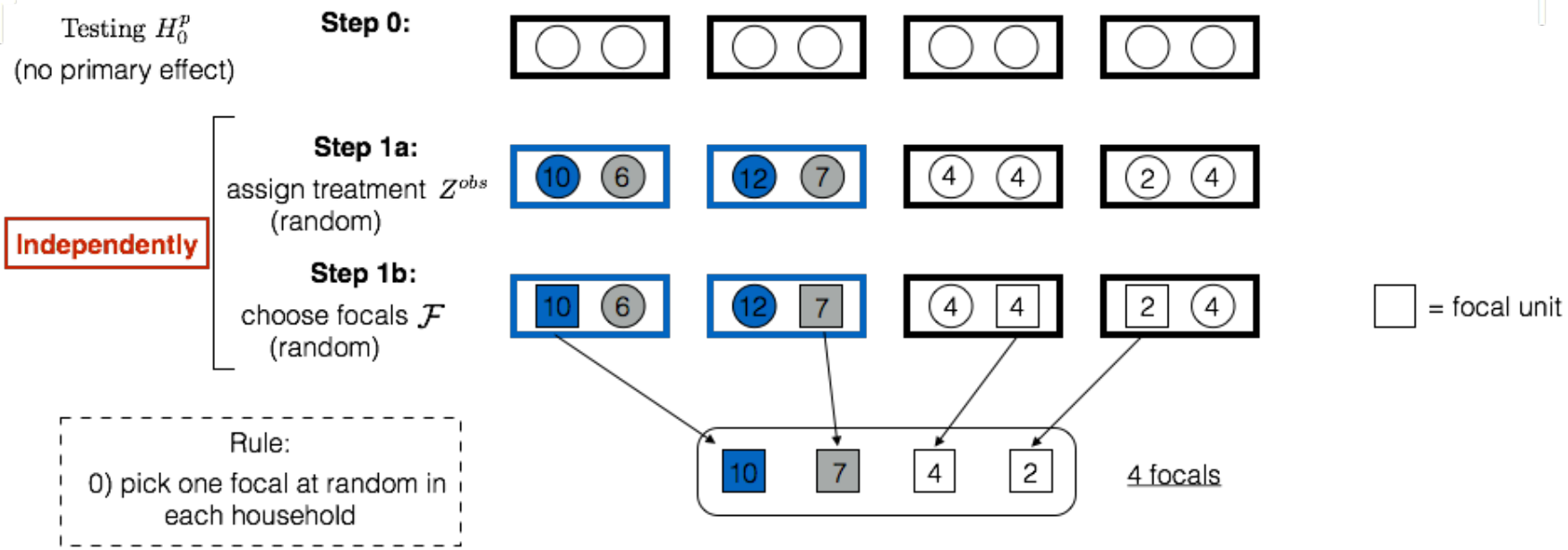
$*$ We need $T(Z', Y') \overset{d}{=} T(Z, Y) \mid F, \mathcal{Z}_F$. It holds:

$$T(Z', Y') \overset{(2)}{=} T(Z'_F, Y'_F) \overset{(3)}{=} T(Z'_F, Y_F) \overset{d}{=} T(Z_F, Y_F) = T(Z, Y).$$

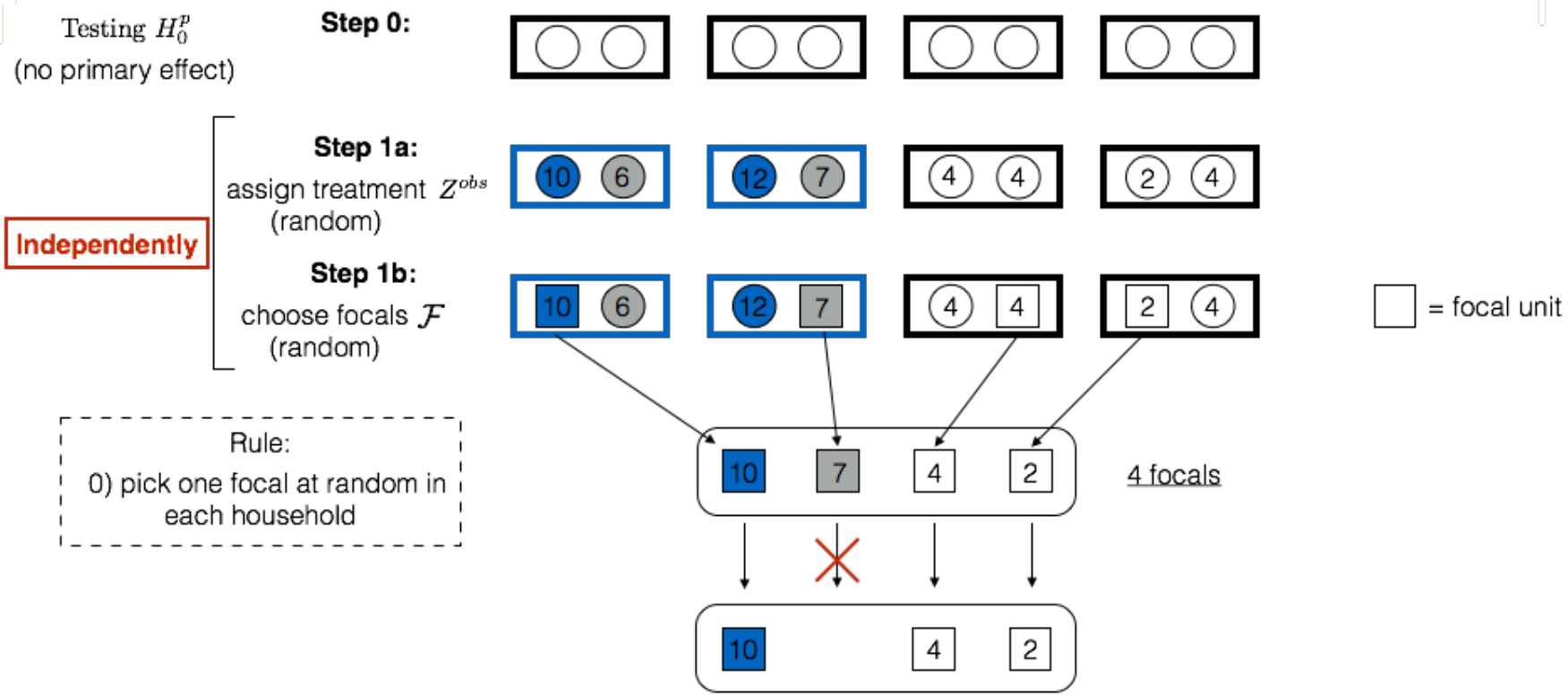
# A valid test (Aronow, 2012; Athey et. al., 2017)



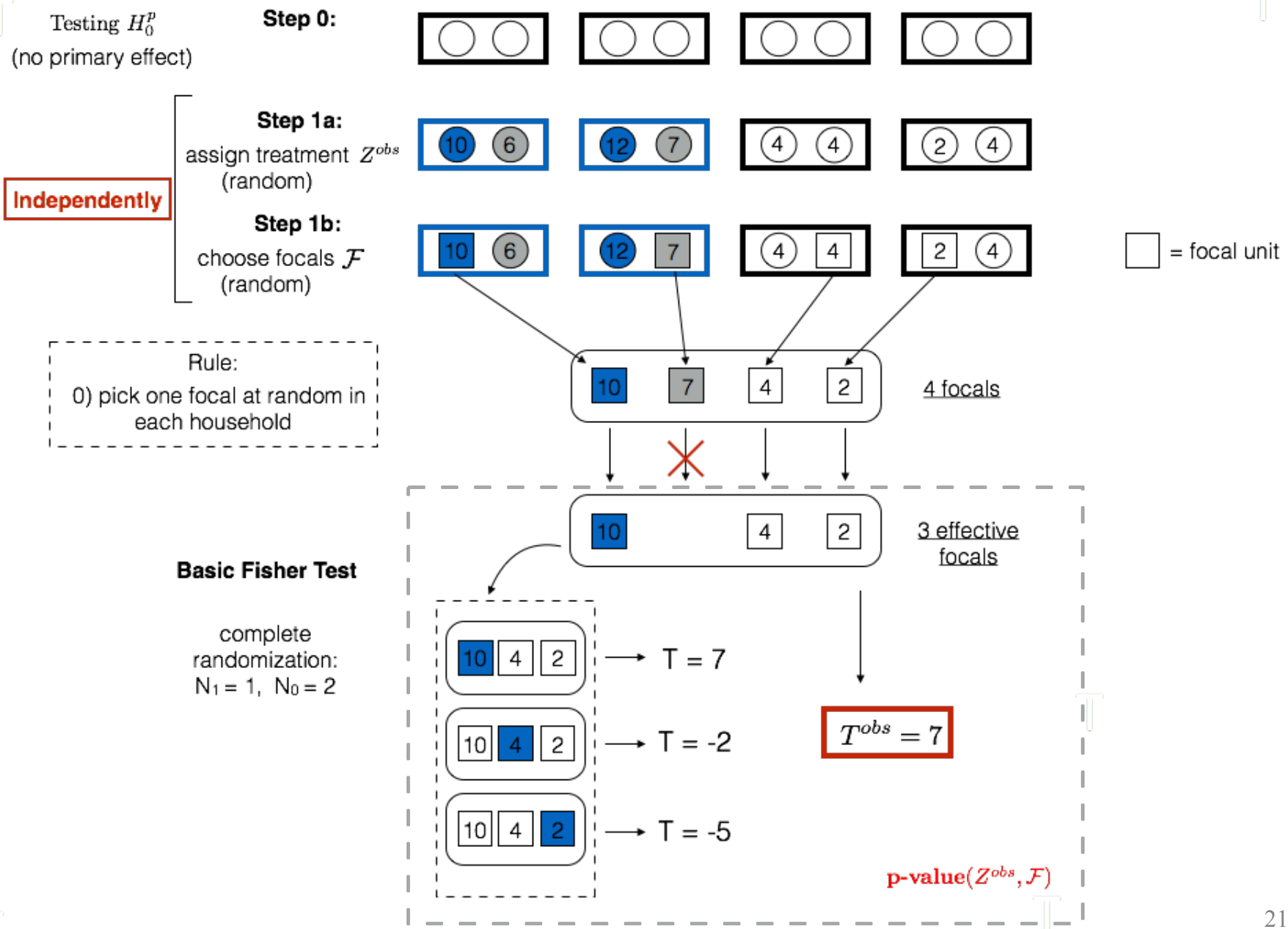
# A valid test (Aronow, 2012; Athey et. al., 2017)



# A valid test (Aronow, 2012; Athey et. al., 2017)



# A valid test (Aronow, 2012; Athey et. al., 2017)



# Problems with current methods

**Power.** Focal units are selected independently of observed $Z$. So, it can happen that $H_0$ is not sharp for some of them.

These focal units are effectively dropped from the study. Implies *loss of power.*

# Problems with current methods

**Power.** Focal units are selected independently of observed $Z$. So, it can happen that $H_0$ is not sharp for some of them.

These focal units are effectively dropped from the study. Implies *loss of power.*

**Computation.** Furthermore, $\mathcal{Z}_F$ needs to be calculated explicitly. The resulting test is not a permutation test, in general, and so it is *not computationally feasible.*

e.g., when households have unequal sizes, the test of (Aronow 2012, Athey et.al., 2017) is not a permutation test.

# Permutation tests

*Why permutation tests are hard to achieve under interference?*

Permutation tests rely on *symmetries* in experimental design of treatment assignment (e.g., complete randomization across households, within households, etc.)

Under interference, a permutation test requires symmetries on the "exposure level". However, the design offers symmetries on the " treatment level".

In other words, in an experiment we randomize $Z_i$ but we would like to have randomized $h_i(Z)$.

# Permutation tests

*Why permutation tests are hard to achieve under interference?*

Permutation tests rely on *symmetries* in experimental design of treatment assignment (e.g., complete randomization across households, within households, etc.)

Under interference, a permutation test requires symmetries on the "exposure level". However, the design offers symmetries on the " treatment level".

In other words, in an experiment we randomize $Z_i$ but we would like to have randomized $h_i(Z)$.

Goal: How to achieve powerful permutation tests?

# Conditioning mechanisms

Formalize conditioning of randomization test through *conditioning event* $C$:

$C = (F, \mathcal{Z})$

$F =$ focal units in the test, and

$\mathcal{Z} =$ assignments in the test.

Then, conditional distribution of $Z$ given $C$ is:

$$P(Z|C) = \underbrace{P(C|Z)}_{\text{conditioning mechanism}} \times \underbrace{pr(Z)}_{\text{experiment design}}$$

# Conditioning mechanisms

Formalize conditioning of randomization test through *conditioning event* $C$:

$C = (F, \mathcal{Z})$

$F =$ focal units in the test, and

$\mathcal{Z} =$ assignments in the test.

Then, conditional distribution of $Z$ given $C$ is:

$$P(Z|C) = \underbrace{P(C|Z)}_{\text{conditioning mechanism}} \times \underbrace{pr(Z)}_{\text{experiment design}}$$

- Analyst chooses $P(C|Z)$, the conditioning mechanism. It should correct asymmetries in the design, so that a permutation test is possible.

- Current methods use the following simple mechanism:

$$P(C|Z) = P(F) \times \text{degen}(\mathcal{Z} = \mathcal{Z}_F).$$

- More complex CM can lead to important improvements.

# Permutation tests for absenteeism study

Test for primary effect $H_0^p$:

1. In a treated household, choose the exposed unit as focal.
2. In a control household, choose any one unit as focal at random.
3. Perform a test by *permuting exposures* on the focal units.

Conditioning mechanism is $P(C|Z) = P(F|Z)P(\mathcal{Z}|F, Z)$ where:

$*$ $P(F|Z)$ is described in Step (1) and is simple. Conditioning on observed $Z$ allows having to drop focal units (we select only "blue" and "white" units).

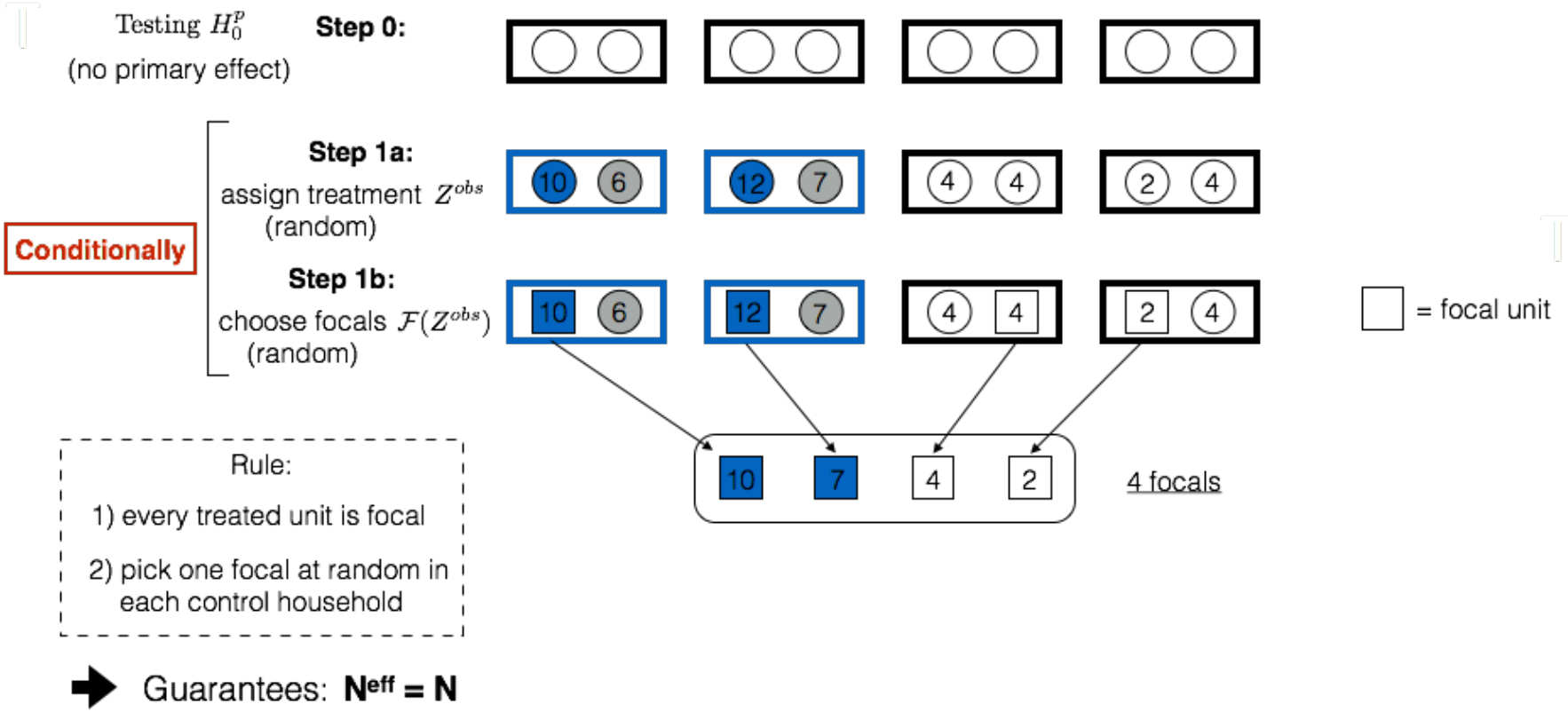
$*$ $P(\mathcal{Z}|F, Z)$ is a *permutation of exposures* of focal units, so there is no need to calculate $\mathcal{Z}$ explicitly.

# Permutation test for $H_0^p$

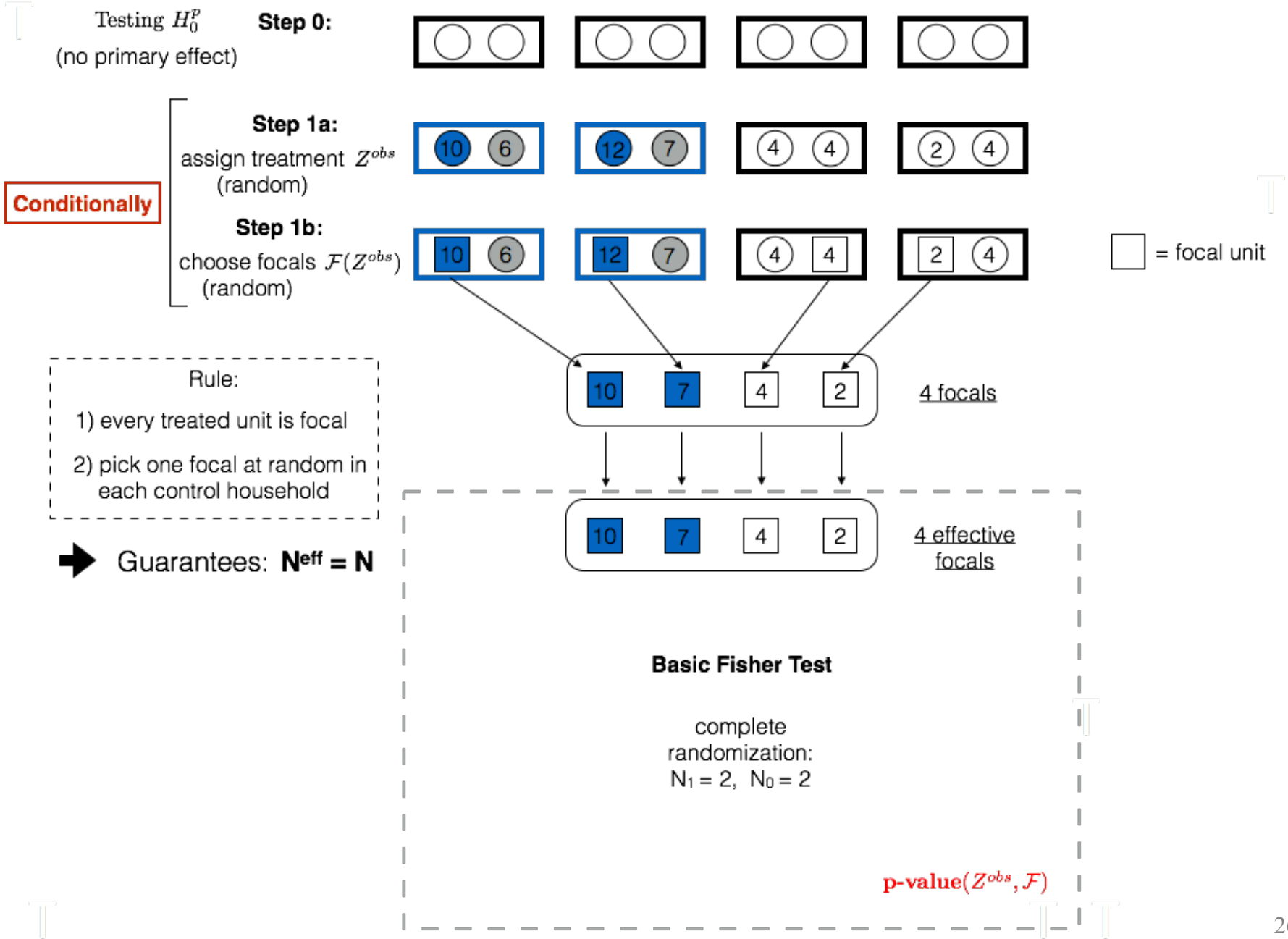
Testing $H_0^p$
(no primary effect)

**Step 0:**

# Permutation test for $H_0^p$



# Permutation test for $H_0^p$



# Why this works

**Theorem**

*Let $E$ be the exposure vector for focal units. Suppose that*

$$P(E) \propto 1 \text{ and } P(F|Z)P(Z|E) \propto 1.$$

*Then, $P(E|C) \propto 1$ and the conditional randomization test given $C$ can be implemented as a permutation test.*

# Why this works

**Theorem**

*Let $E$ be the exposure vector for focal units. Suppose that*

$$P(E) \propto 1 \text{ and } P(F|Z)P(Z|E) \propto 1.$$

*Then, $P(E|C) \propto 1$ and the conditional randomization test given $C$ can be implemented as a permutation test.*

$*$ In our test, $E \overset{a.e.}{=} H$ (where $H =$ household assignment) because we keep the focal from *every* household. Thus, $P(E) \propto 1$ because the design randomizes on the household level.

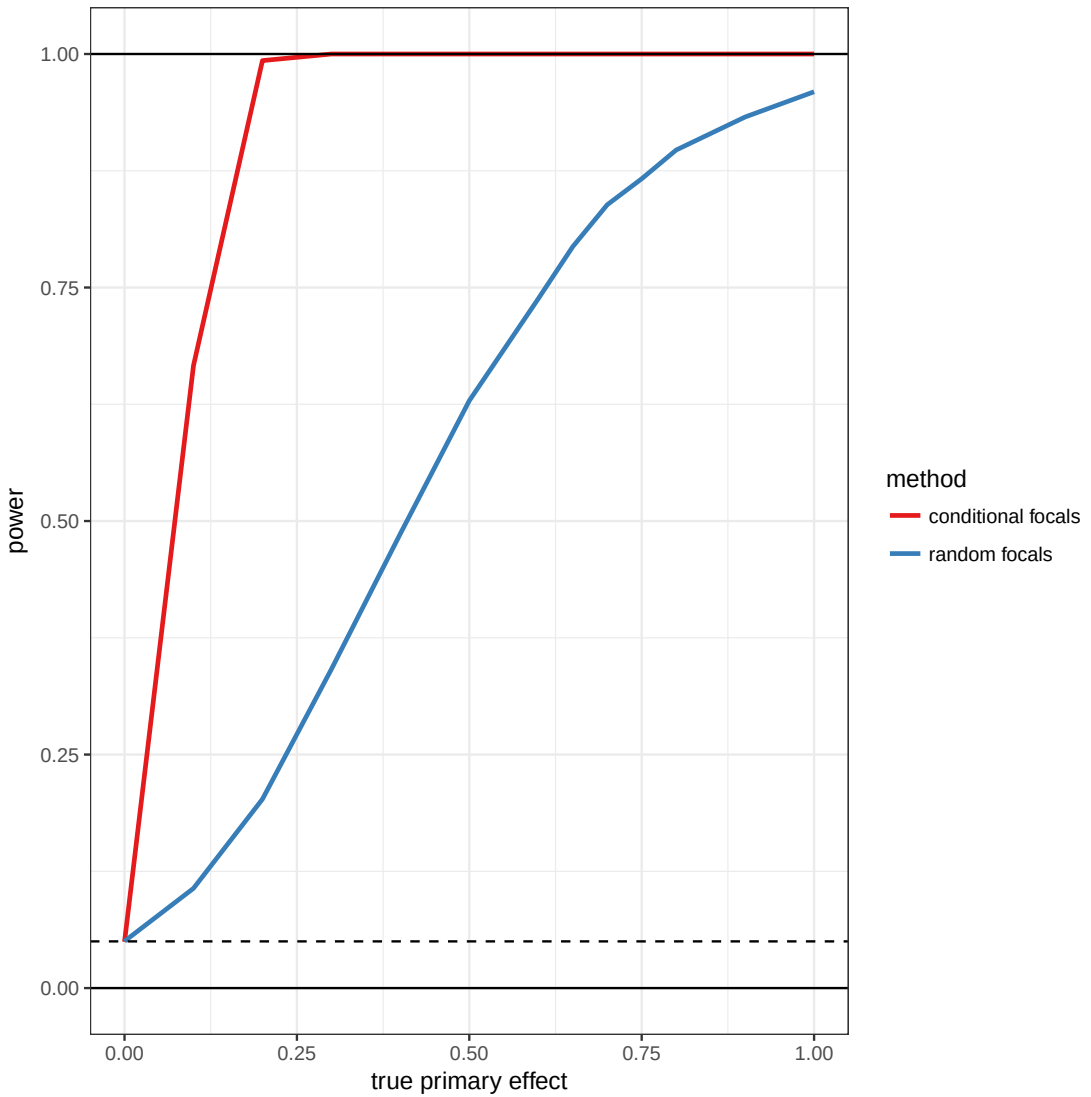
$*$ It holds, $P(Z|E) = P(Z|H) = \prod_{j:H_j=1} 1/n_j$. Implies asymmetry: treated units less likely to come from larger household.

$*$ Correct asymmetry by defining $P(F|Z) = \prod_{j:H_j=0} 1/n_j$.

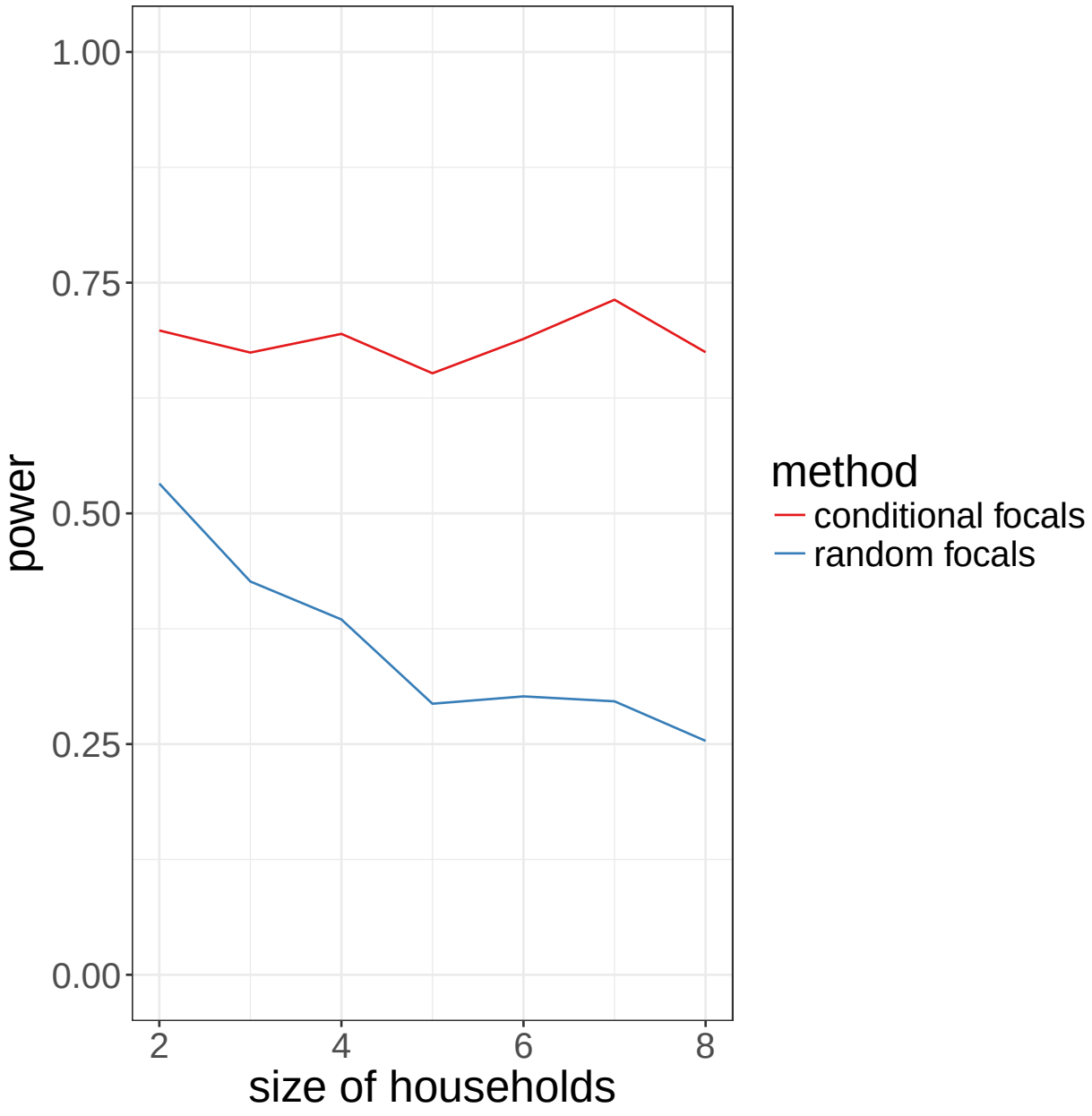
$*$ Then, $P(F|Z)P(Z|E) = \prod_j 1/n_j \propto 1$. Implies permutation test.

# Simulation – Power for test of no primary effect

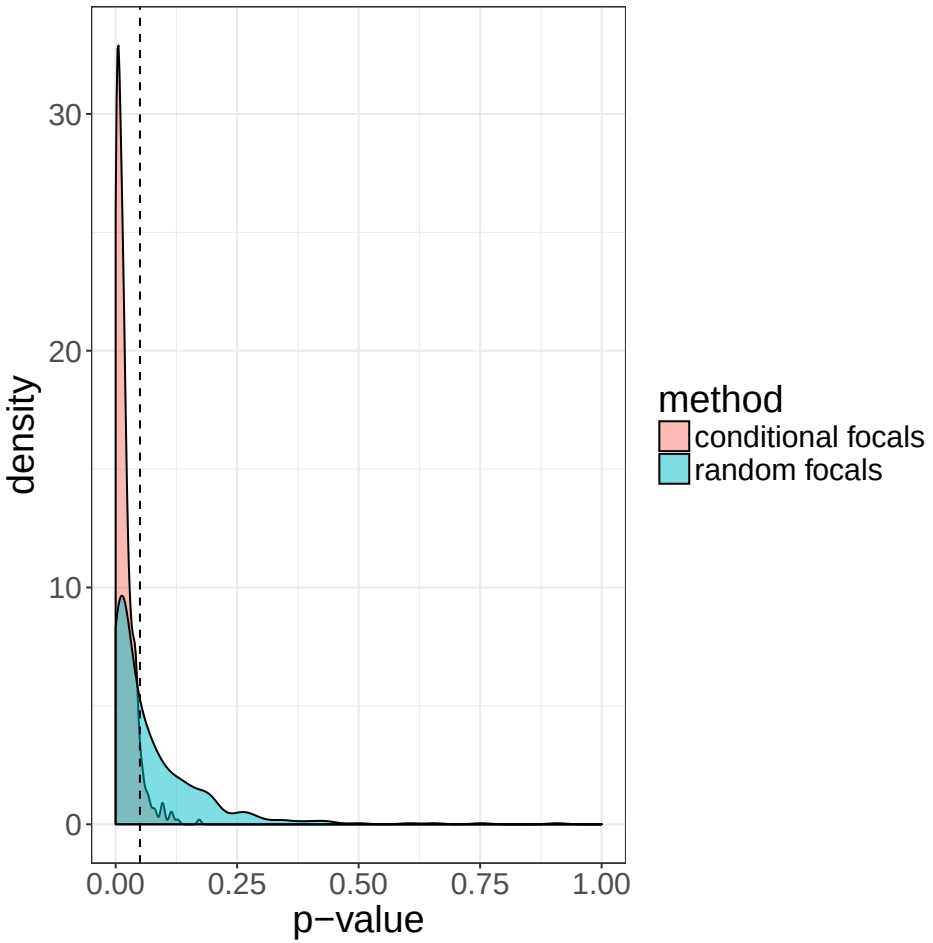
We set 500 households with 10 units each.
Outcome model: $Y_i(1,1) = Y_i(0,0) + \tau$, and $Y_i(0,0) \sim \mathcal{N}(0, \sigma^2)$.

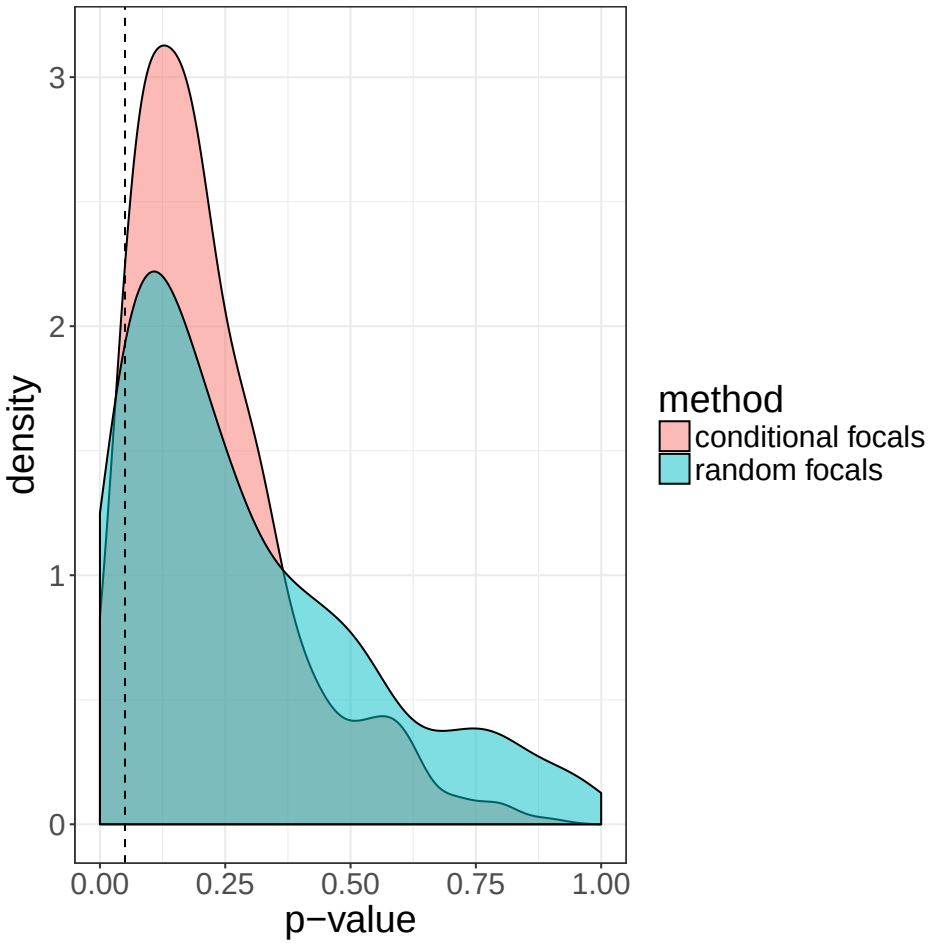


# Simulation – Power for test of no primary effect



# Absenteeism data – distribution of randomization $p$-values




Distribution of p-values over choices of focals for testing $H_0^p$ (left) and $H_0^s$ (right).

For primary effect test, conditional focal selection rejects 91% vs 65% for random focals.

# Conclusion

Randomization inference is appealing – makes minimal assumptions.
But hard when there is interference.

Current methods are impractical. They lose power and cannot be implemented as permutation tests, in general.

Conditioning mechanisms offer a principled guide to address both issues.

**Future work:**

- Aggregate testing across conditioning events for same $H_0$.
- Aggregate testing across multiple $H_0$.
- Extend to more complex interference (e.g., crime intervention studies).

# Thank you!

Basse, Feller, Toulis (2018). Randomization tests of causal effects with interference between units. *Biometrika* (*arxiv/1709.08036*)

Aronow (2012). A general method for detecting interference between units in randomized experiments. *Sociological Methods & Research 41 (1), 3-16.*

Athey, Eckles, and Imbens (2017). Exact p-values for network interference. *Journal of the American Statistical Association.*

# Simulation – Power for test of no spillover effect