Exact tests for two-stage randomization in the presence of interference

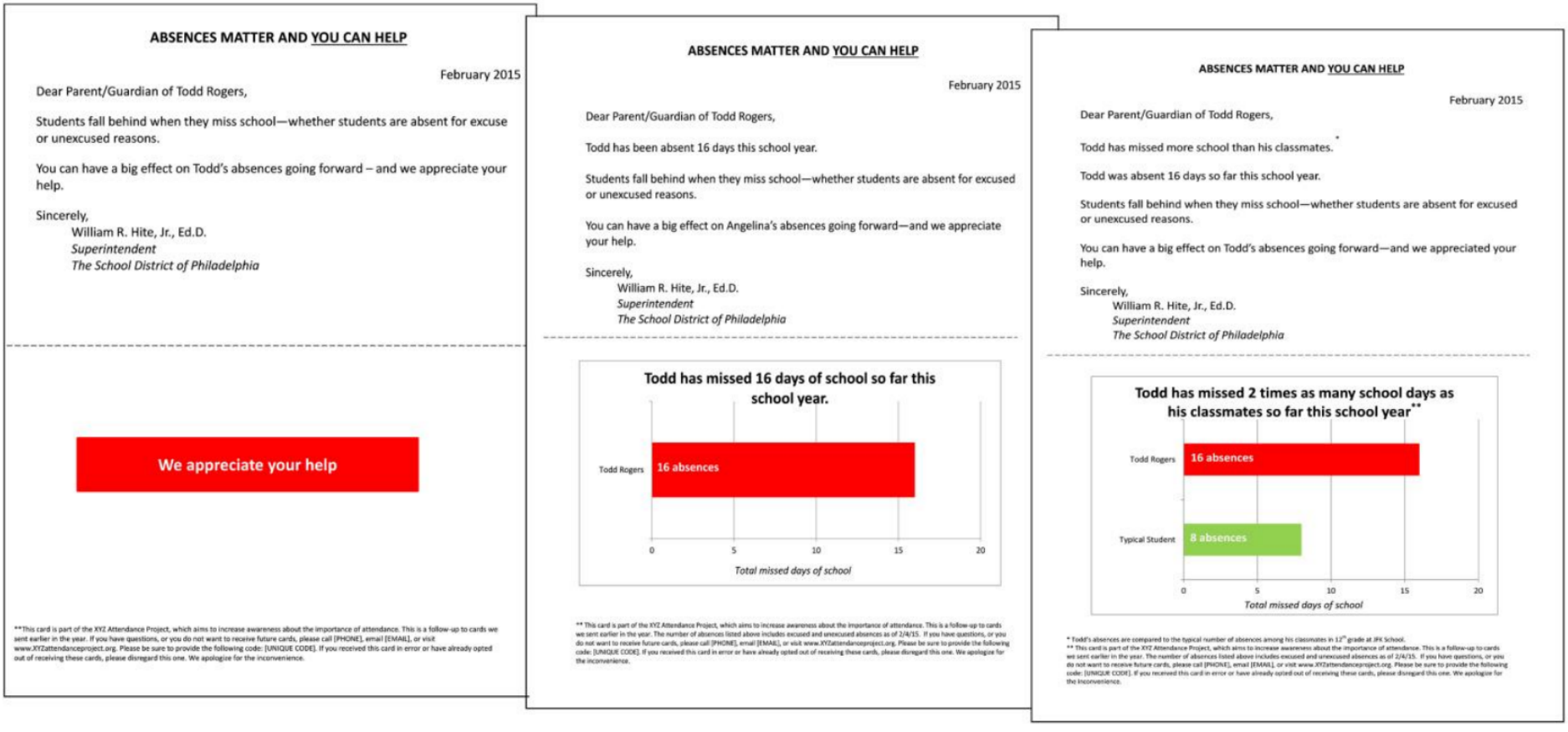
Panagiotis (Panos) Toulis
panos.toulis@chicagobooth.edu

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 (2016) ran a two-stage randomized experiment, aiming to engage parents of a student who was frequently absent from school.
- Data indicated strong primary effect for targeted student.
- Also some positive spillovers to siblings of the targeted student.



ABSENCES MATTER AND YOU CAN HELP

February 2015

Dear Parent/Guardian of Todd Rogers,

Students fall behind when they miss school—whether students are absent for excuse or unexcused reasons.

You can have a big effect on Todd's absences going forward – and we appreciate your help.

Sincerely,
William R. Hite, Jr., Ed.D.
Superintendent
The School District of Philadelphia

We appreciate your help

ABSENCES MATTER AND YOU CAN HELP

February 2015

Dear Parent/Guardian of Todd Rogers,

Todd has been absent 16 days this school year.

Students fall behind when they miss school—whether students are absent for excused or unexcused reasons.

You can have a big effect on Angelina's absences going forward—and we appreciate your help.

Sincerely,
William R. Hite, Jr., Ed.D.
Superintendent
The School District of Philadelphia

Todd has missed 16 days of school so far this school year.

Todd Rogers 16 absences

0 5 10 15 20

Total missed days of school

ABSENCES MATTER AND YOU CAN HELP

February 2015

Dear Parent/Guardian of Todd Rogers,

Todd has missed more school than his classmates.*

Todd was absent 16 days so far this school year.

Students fall behind when they miss school—whether students are absent for excused or unexcused reasons.

You can have a big effect on Todd's absences going forward—and we appreciated your help.

Sincerely,
William R. Hite, Jr., Ed.D.
Superintendent
The School District of Philadelphia

Todd has missed 2 times as many school days as his classmates so far this school year**

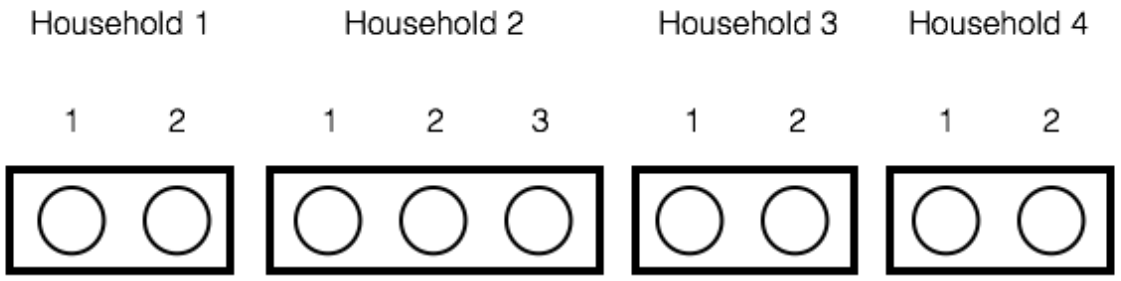
Todd Rogers 16 absences

Typical Student 8 absences

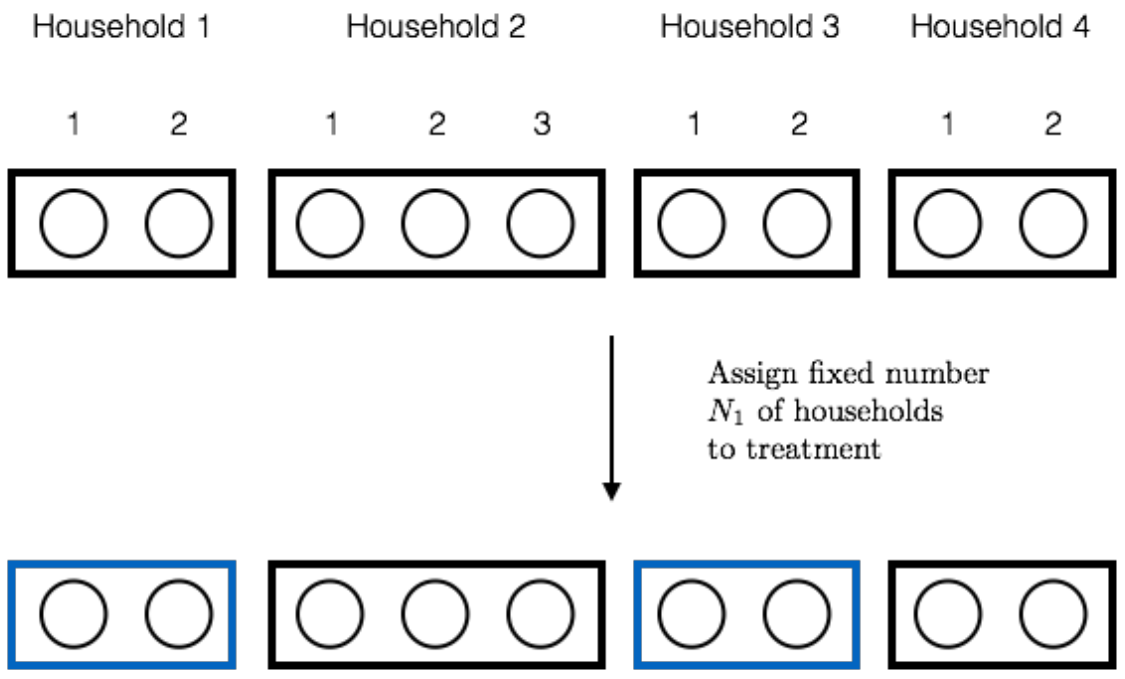
0 5 10 15 20

Total missed days of school

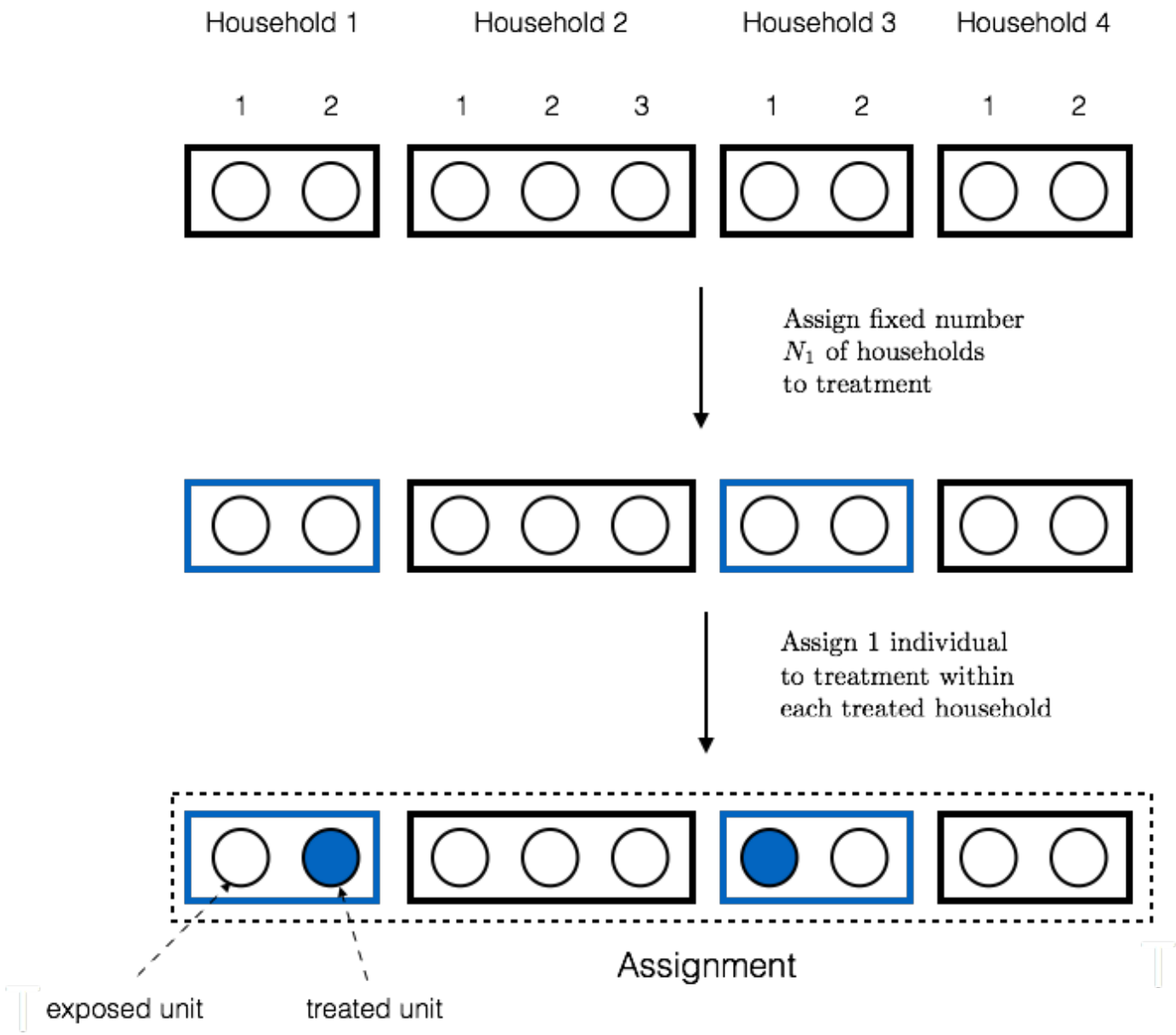
Two-stage randomization



Two-stage randomization



Two-stage randomization



Talk summary

Contribution: construct powerful randomization tests of primary and spillover effects of such interventions by flexible *conditioning mechanisms*.

Talk summary

Contribution: construct powerful randomization tests of primary and spillover effects of such interventions by flexible *conditioning mechanisms*.

Methodology: The concept of conditioning mechanism extends classical conditional randomization tests to cases where

- treatment levels are interdependent and cannot be freely permuted;
- and where the conditioning events have to overlap for more power.

Talk summary

Contribution: construct powerful randomization tests of primary and spillover effects of such interventions by flexible *conditioning mechanisms*.

Methodology: The concept of conditioning mechanism extends classical conditional randomization tests to cases where

- treatment levels are interdependent and cannot be freely permuted;
- and where the conditioning events have to overlap for more power.

Application: For the two-stage design in our application we can derive conditioning mechanisms that can be described as classical permutation tests, but with subtle twists.

Related work

Two-stage randomization: Economics (Crépon et al., 2013), Education (Somers et al., 2010), Political Science (Sinclair et al. 2012), Public Health (Hudgens and Holloran 2008).

Estimation with Interference: Sobel (2006), Hudgens and Halloran (2008), Toulis and Kao (2013), Rigdon and Hudgens (2015), Kang and Imbens (2016), Aronow and Samii (2017), Basse and Feller (2017).

Testing: Aronow (2012), Rosenbaum (2007), Bowers et. al. (2013), (Athey et.al., 2016).

Notation

- Unit = student; Household = collection of students (siblings).
- Indexing: i = unit, j = household.
- $R_{ij} = 1$ is unit i is in j household; 0 otherwise.
- **Treatment:** $Z_i \in \{0, 1\}$ = treatment of unit i; $Z = (Z_i)$.
- **Design:** $p(Z)$ = unif. over set where we treat half of households, and at most one unit per household.
- **Outcomes:** $Y_i(Z)$ = outcome of unit i under assignment Z.

Notation

- Unit = student; Household = collection of students (siblings).
- Indexing: i = unit, j = household.
- $R_{ij} = 1$ is unit i is in j household; 0 otherwise.
- **Treatment:** $Z_i \in \{0, 1\}$ = treatment of unit i; $Z = (Z_i)$.
- **Design:** $p(Z)$ = unif. over set where we treat half of households, and at most one unit per household.
- **Outcomes:** $Y_i(Z)$ = outcome of unit i under assignment Z.

♠ Derived notation:

$W_j = \sum_i Z_i R_{ij}$ = treatment of household j (either 0 or 1).

$[i] = \sum_j j R_{ij}$ = household where unit i resides.

Interference

- We usually assume only two outcomes for unit i, namely $Y_i(0), Y_i(1)$.
- But this relies on **no interference** assumption (Cox, 1958).

Interference

- We usually assume only two outcomes for unit i, namely $Y_i(0), Y_i(1)$.
- But this relies on **no interference** assumption (Cox, 1958).
- Frequently, the outcome of a unit may depend on treatment of others!
  - e.g., I watched the movie because my friend saw the movie ad and told me about it;
  - student gets message about absenteeism affecting educational outcomes of siblings.
- When interference is present, outcome is better denoted by $Y_i(Z)$.

Interference

- We usually assume only two outcomes for unit i, namely $Y_i(0), Y_i(1)$.
- But this relies on **no interference** assumption (Cox, 1958).
- Frequently, the outcome of a unit may depend on treatment of others!
  - e.g., I watched the movie because my friend saw the movie ad and told me about it;
  - student gets message about absenteeism affecting educational outcomes of siblings.
- When interference is present, outcome is better denoted by $Y_i(Z)$.

No interference assumption (SUTVA, Rubin, 1974)

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

$\Rightarrow$ implies only two potential outcomes for unit i, namely $Y_i(0), Y_i(1)$.

Our interference assumptions

- In our application, we assume that the outcome of a unit may depend only on its treatment and the treatment of the household; formally,

$$Y_i(Z) = Y_i(Z') \text{ if } Z_i = Z_i' \text{ and } \mathbf{W}_{[\mathbf{i}]} = \mathbf{W}'_{[\mathbf{i}]}.$$

- Consequently,

$$Y_i(Z) \equiv \begin{cases} Y_i(1,1) = & \text{treated unit in treated household.} \\ Y_i(0,1) = & \text{control unit in treated household.} \\ Y_i(0,0) = & \text{control unit in control household.} \end{cases}$$

Causal effects

- **Causal primary effect** for unit i can be defined as

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

- **Causal spillover effect** can be defined as:

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

- But how to test for such causal effects...?

Causal effects

- **Causal primary effect** for unit i can be defined as

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

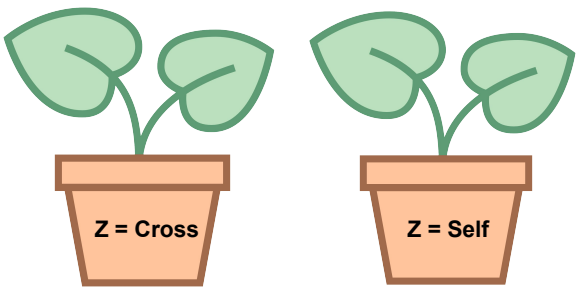
- **Causal spillover effect** can be defined as:

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

- But how to test for such causal effects...?
- A very powerful idea that makes no modeling assumptions and uses only the available information from the design is that of **randomization tests**.

Classical randomization test for causal effects

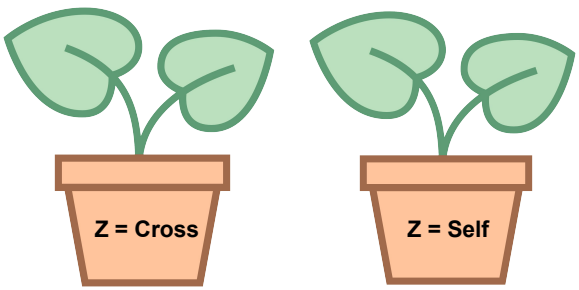
- Working with Darwin's data RA Fisher wanted to compare the effects of self-fertilization with cross-fertilization on plant height – the problem is that there are missing data.



unit	treatment	
	cross-fertilized	self-fertilized
1	15	?
2	?	20

Classical randomization test for causal effects

- Working with Darwin's data RA Fisher wanted to compare the effects of self-fertilization with cross-fertilization on plant height – the problem is that there are missing data.



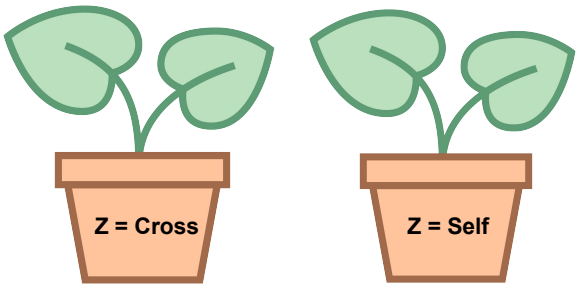
unit	treatment	
	cross-fertilized	self-fertilized
1	15	?
2	?	20

- Fisher noted that under the null hypothesis of no difference between treatments, we **can fill in the missing data**:

$$\text{Height}_i(\text{self-fertilization}) = \text{Height}_i(\text{cross-fertilization}).$$

But how to test?

- Working with Darwin's data RA Fisher compared the effects of self-fertilization with cross-fertilization on plant height.



unit	treatment	
	cross-fertilized	self-fertilized
1	15	15
2	20	20

- Fisher noted that under the null hypothesis of no difference between treatments, we can fill in the missing data:

$$\text{Height}_i(\text{self-fertilization}) = \text{Height}_i(\text{cross-fertilization}).$$

- The null hypothesis is **sharp** because it allows imputation of missing outcomes, and thus *randomization-based inference*.

Randomization-based inference through Fisher test

1. Pick a test statistic $T(Z|y)$ that is a reasonable estimate of the causal effect of interest ($y =$ observed data) – could actually rely on a model!
   - e.g., difference in means between treated and control units:

$$T(Z|y) = \frac{1}{\sum_i Z_i} \sum_i Z_i y_i - \frac{1}{\sum_i 1 - Z_i} \sum_i (1 - Z_i) y_i$$

Randomization-based inference through Fisher test

1. Pick a test statistic $T(Z|y)$ that is a reasonable estimate of the causal effect of interest ($y =$ observed data) – could actually rely on a model!
   - e.g., difference in means between treated and control units:

$$T(Z|y) = \frac{1}{\sum_i Z_i} \sum_i Z_i y_i - \frac{1}{\sum_i 1 - Z_i} \sum_i (1 - Z_i) y_i$$

2. Calculate observed value $T^{\text{obs}} = T(Z^{\text{obs}}|y)$.

Randomization-based inference through Fisher test

1. Pick a test statistic $T(Z|y)$ that is a reasonable estimate of the causal effect of interest ($y =$ observed data) – could actually rely on a model!
   - e.g., difference in means between treated and control units:

$$T(Z|y) = \frac{1}{\sum_i Z_i} \sum_i Z_i y_i - \frac{1}{\sum_i 1 - Z_i} \sum_i (1 - Z_i) y_i$$

2. Calculate observed value $T^{\text{obs}} = T(Z^{\text{obs}}|y)$.
3. For $m = 1, 2, \ldots$:
   (i) Sample $Z \sim p(Z)$ according to design.
   (ii) Store $T_m = T(Z|y)$.

Randomization-based inference through Fisher test

1. Pick a test statistic $T(Z|y)$ that is a reasonable estimate of the causal effect of interest ($y =$ observed data) – could actually rely on a model!
   - e.g., difference in means between treated and control units:

$$T(Z|y) = \frac{1}{\sum_i Z_i} \sum_i Z_i y_i - \frac{1}{\sum_i 1 - Z_i} \sum_i (1 - Z_i) y_i$$

2. Calculate observed value $T^{\text{obs}} = T(Z^{\text{obs}}|y)$.
3. For $m = 1, 2, \ldots$:
   (i) Sample $Z \sim p(Z)$ according to design.
   (ii) Store $T_m = T(Z|y)$.
4. Calculate $\texttt{p-value} = E\left(\mathbb{I}\{T_m \geq T^{\text{obs}}\}\right)$.

Randomization-based inference through Fisher test

1. Pick a test statistic $T(Z|y)$ that is a reasonable estimate of the causal effect of interest ($y =$ observed data) – could actually rely on a model!
   - e.g., difference in means between treated and control units:

$$T(Z|y) = \frac{1}{\sum_i Z_i} \sum_i Z_i y_i - \frac{1}{\sum_i 1 - Z_i} \sum_i (1 - Z_i) y_i$$

2. Calculate observed value $T^{\text{obs}} = T(Z^{\text{obs}}|y)$.
3. For $m = 1, 2, \ldots$:
   (i) Sample $Z \sim p(Z)$ according to design.
   (ii) Store $T_m = T(Z|y)$.
4. Calculate $\texttt{p-value} = E\left(\mathbb{I}\{T_m \geq T^{\text{obs}}\}\right)$.

Notes:

- **No model assumption** on outcomes (actually outcomes y are assumed fixed).
- Randomness comes only from design (which we control!).

Randomization-based inference through Fisher test

1. Pick a test statistic $T(Z|y)$ that is a reasonable estimate of the causal effect of interest (y = observed data) – could actually rely on a model!
   - e.g., difference in means between treated and control units:

$$T(Z|y) = \frac{1}{\sum_i Z_i} \sum_i Z_i y_i - \frac{1}{\sum_i 1 - Z_i} \sum_i (1 - Z_i) y_i$$

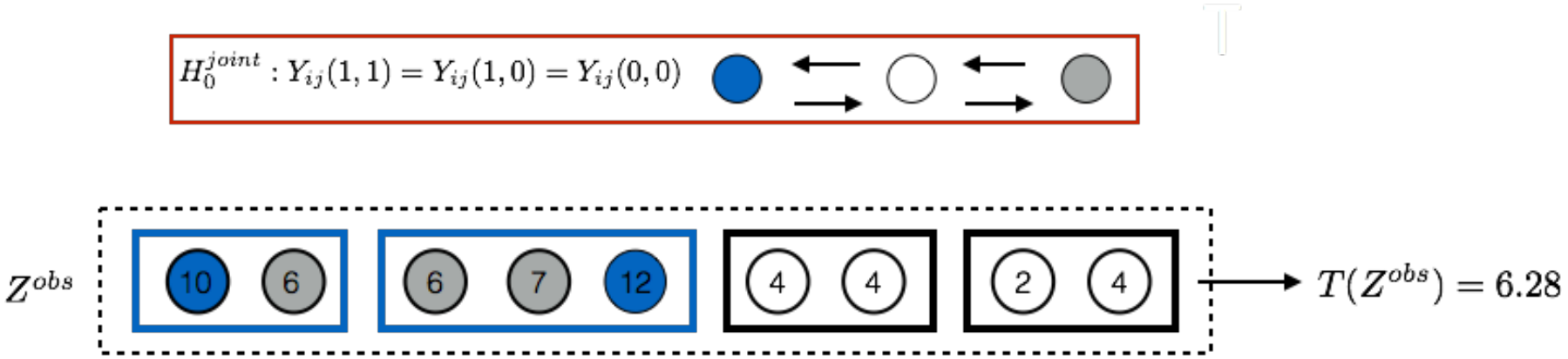
2. Calculate observed value $T^{\text{obs}} = T(Z^{\text{obs}}|y)$.
3. For $m = 1, 2, \ldots$:
   (i) Sample $Z \sim p(Z)$ according to design.
   (ii) Store $T_m = T(Z|y)$.
4. Calculate $\texttt{p-value} = E\left(\mathbb{I}\{T_m \geq T^{\text{obs}}\}\right)$.

Notes:

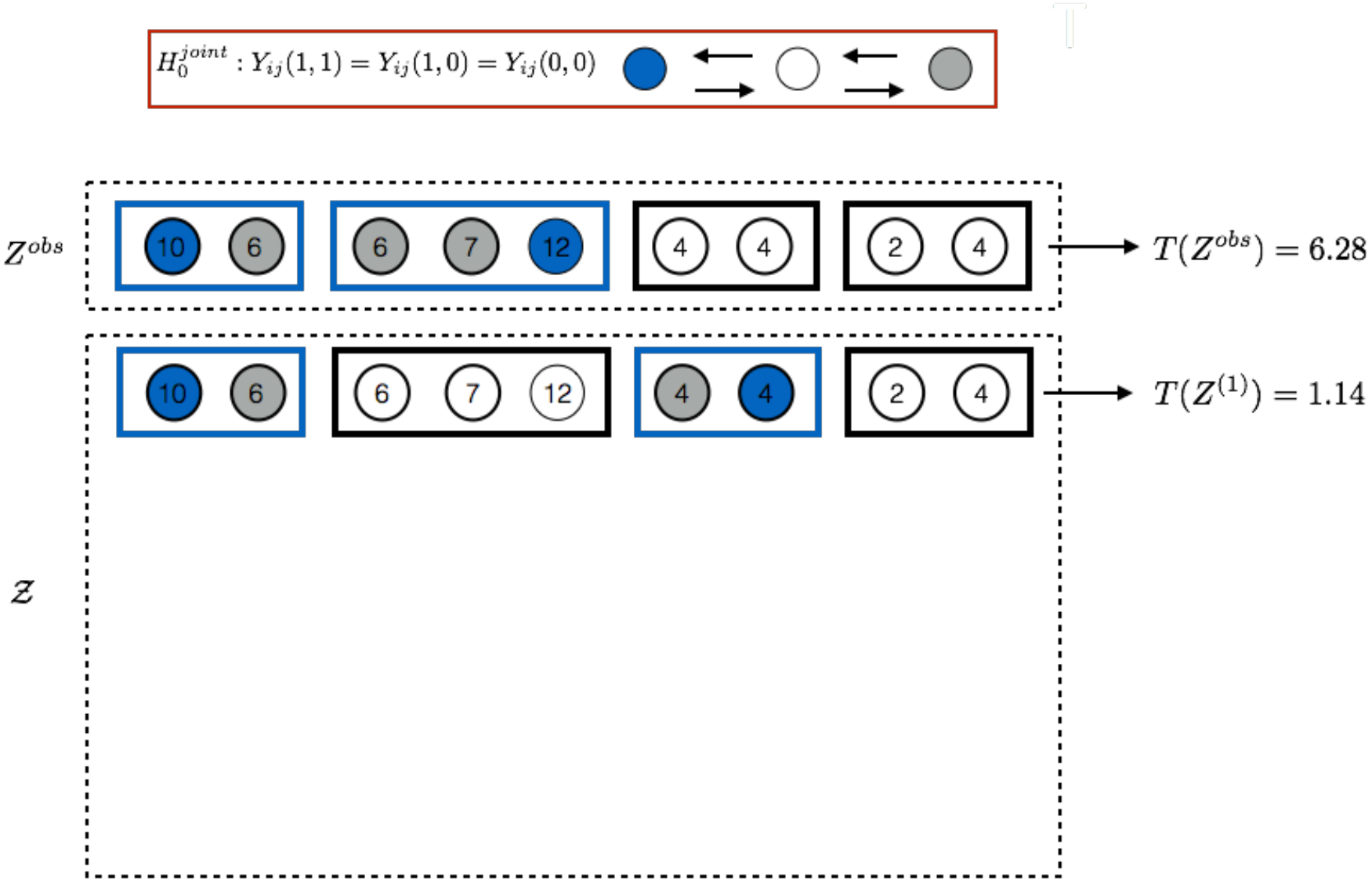
- **No model assumption** on outcomes (actually outcomes y are assumed fixed).
- Randomness comes only from design (which we control!).
- Step 3(ii) is **only possible** because hypothesis is sharp, since

$$T(Z|Y(Z)) = T(Z|y), \text{ for all } Z.$$

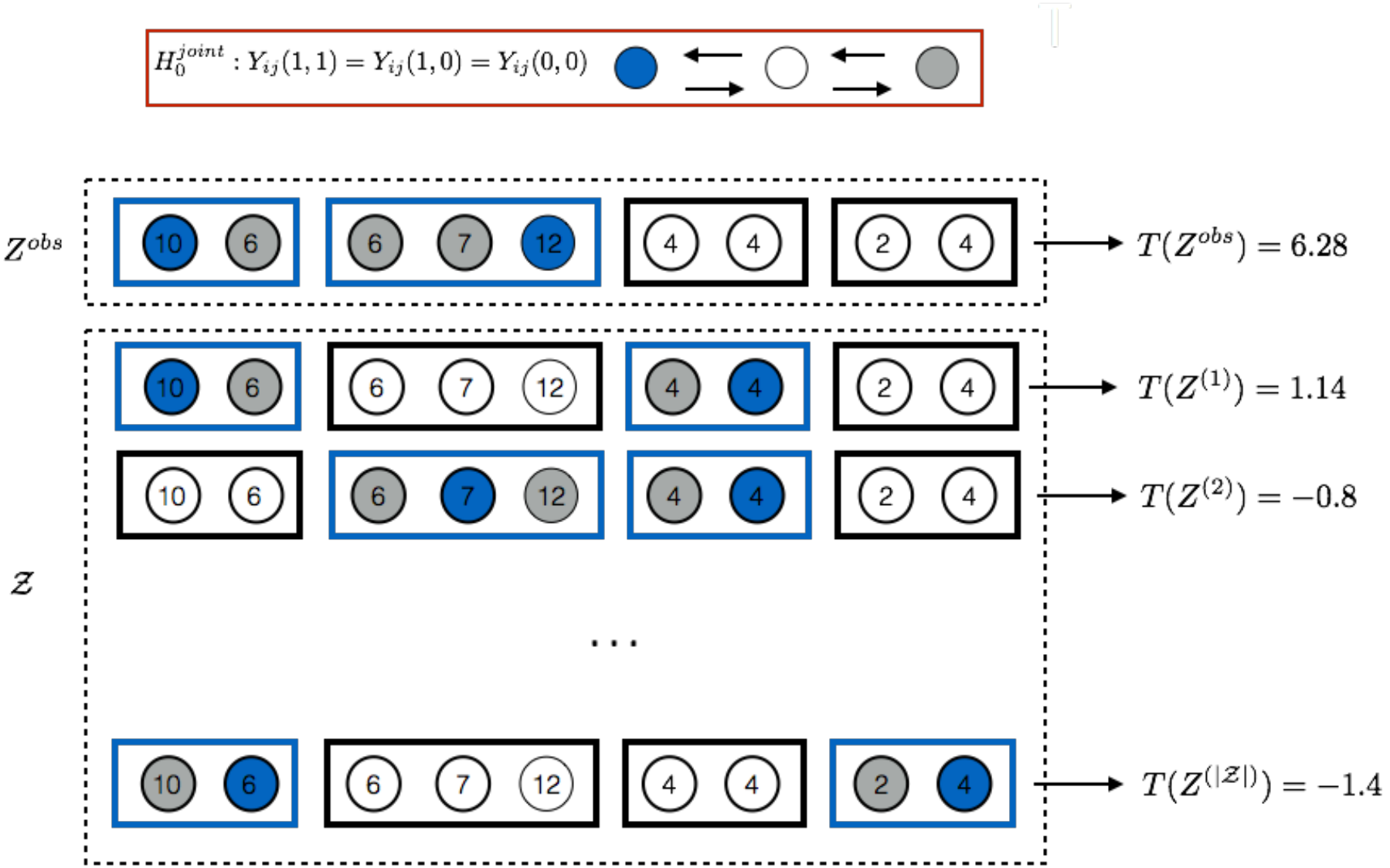
Fisher test in our setting: simple case



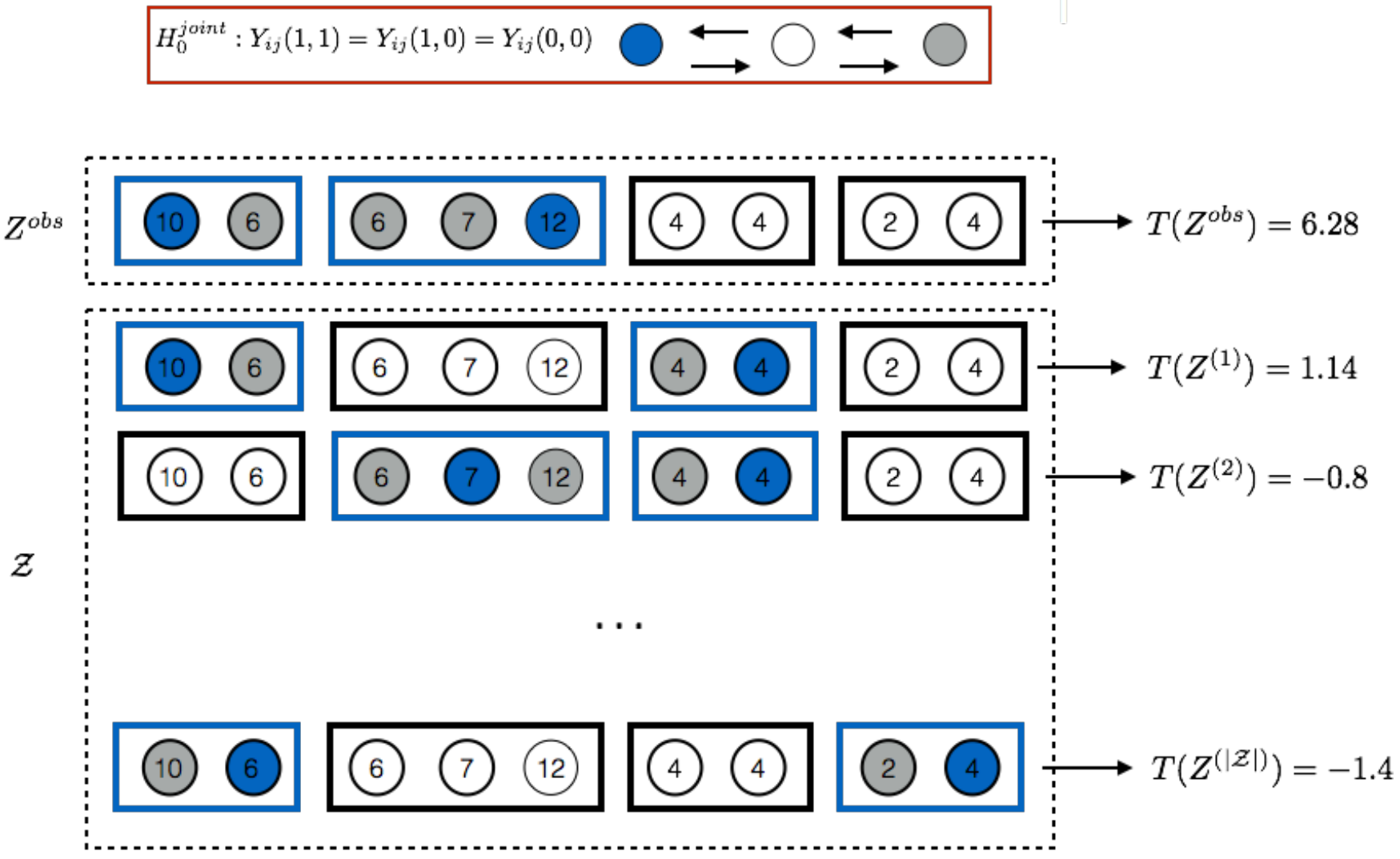
Fisher test in our setting: simple case



Fisher test in our setting: simple case

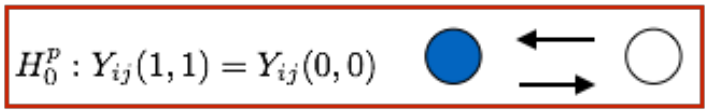


Fisher test in our setting: simple case

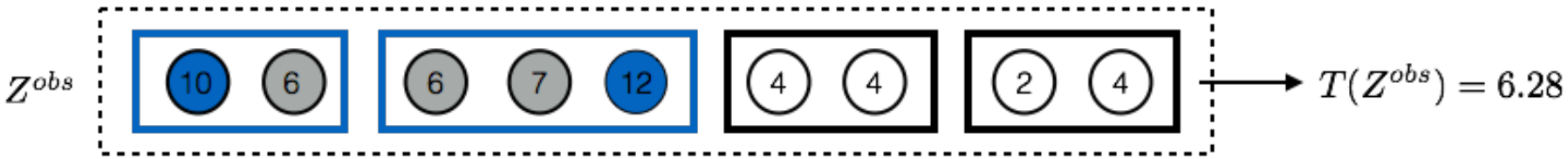


$$p_{val} = \sum_{Z \in \mathcal{Z}} p(Z) \mathbb{1}(T^{obs} > T(Z)) \qquad \text{where} \qquad T(Z) = \overline{Y}_T^{obs} - \overline{Y}_C^{obs}$$

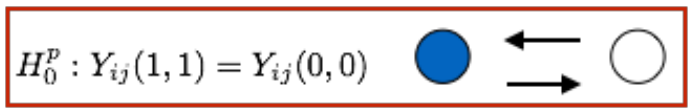
The problem with interference: testing primary effect



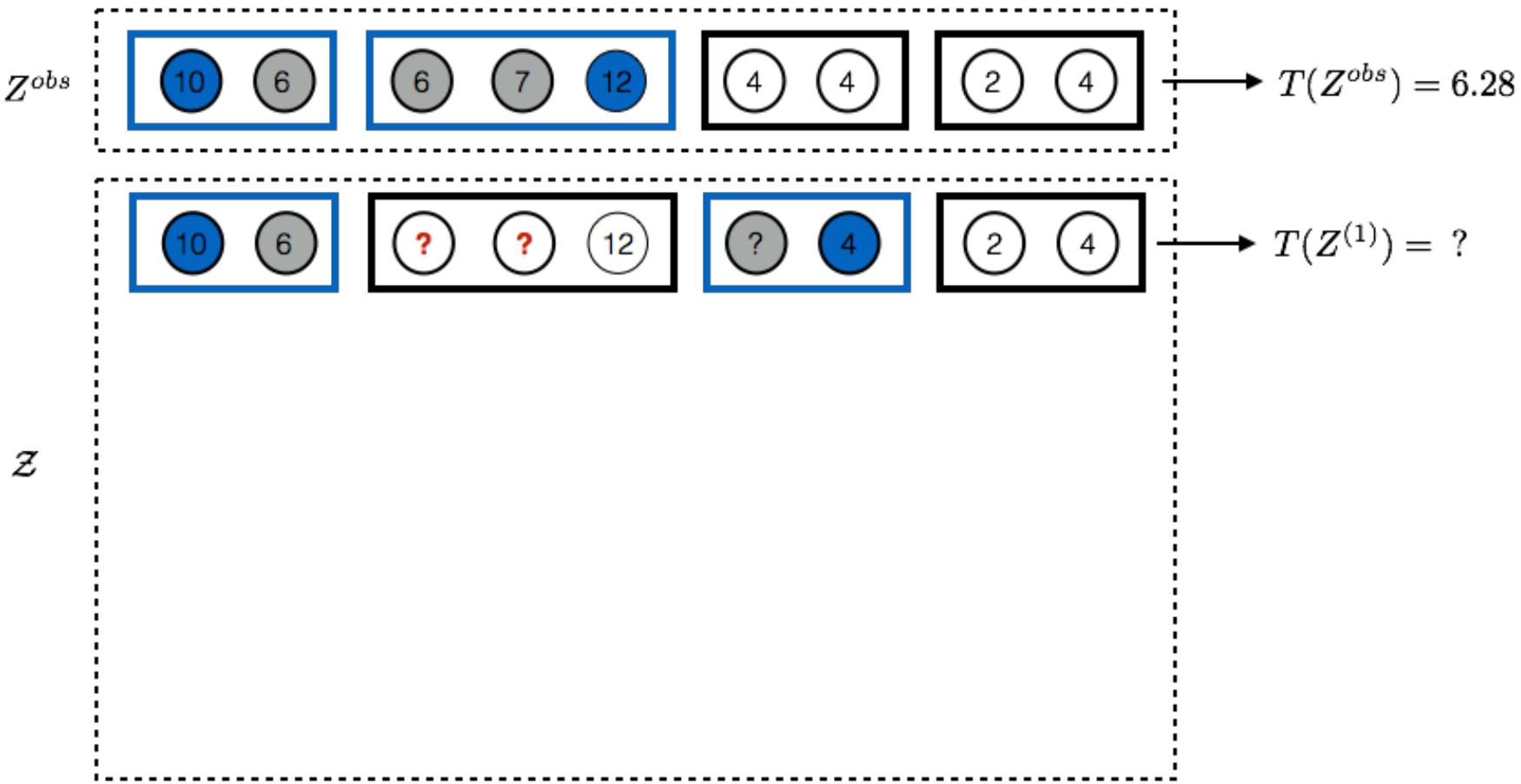
and $T(Z) = \overline{Y}_T(Z) - \overline{Y}_C(Z)$



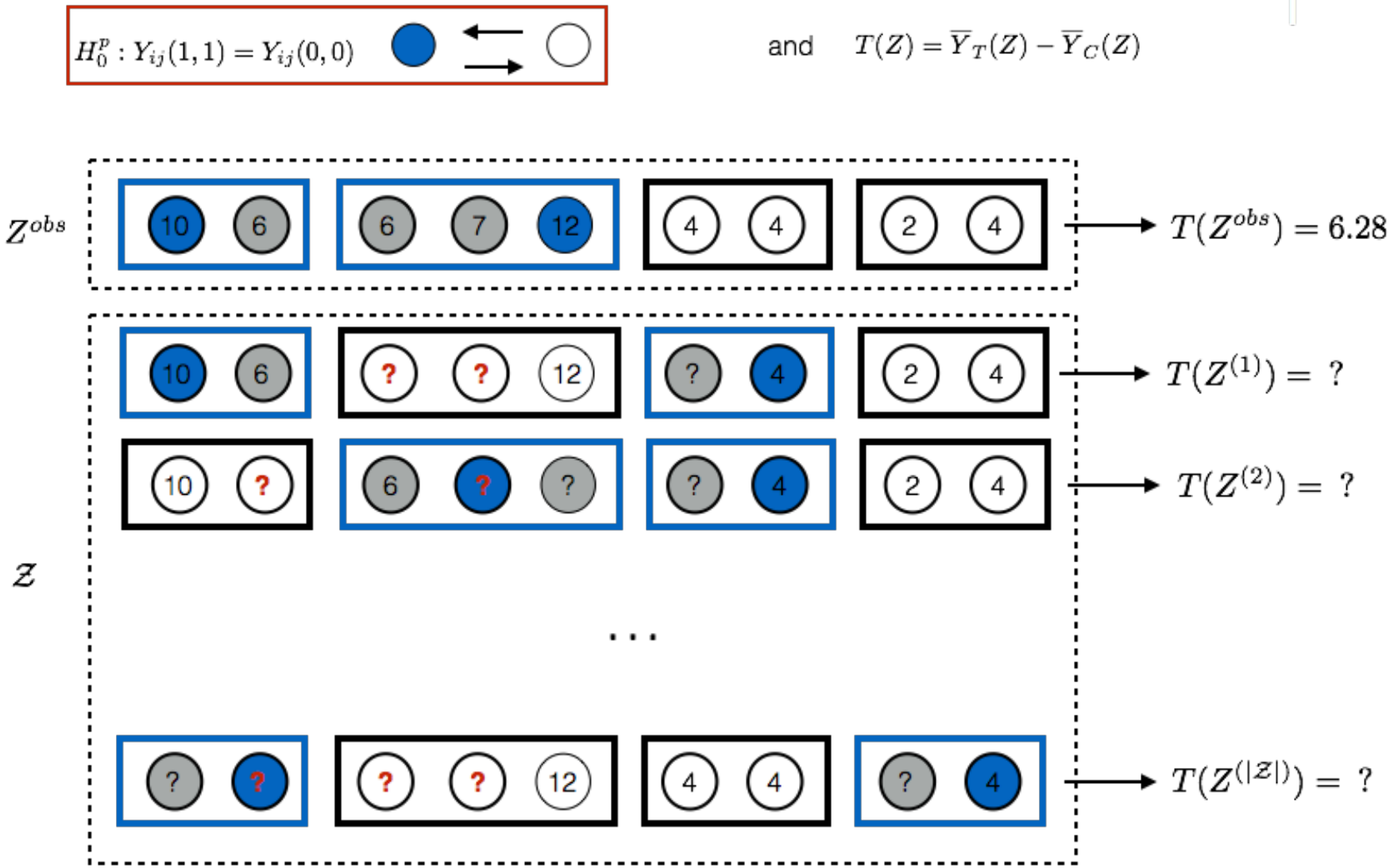
The problem with interference: testing primary effect



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



The problem with interference: testing primary effect



$\text{pval} = ?$

Problem with interference

- Null hypothesis (of no primary effect) is **not sharp** $\Rightarrow$ cannot fill in all potential outcomes!
  - Here, we have three levels of treatment: treated, exposed, and control; but null hypothesis of primary (or spillovers) claims the equality of only two of them, and says nothing about the third.

Problem with interference

- Null hypothesis (of no primary effect) is **not sharp** $\Rightarrow$ cannot fill in all potential outcomes!
  - Here, we have three levels of treatment: treated, exposed, and control; but null hypothesis of primary (or spillovers) claims the equality of only two of them, and says nothing about the third.
- Treatment levels depend on each other – unrestricted permutation in the randomization test is not possible.
  - e.g., cannot have a control unit and an exposed unit both in the same household, by assumption.

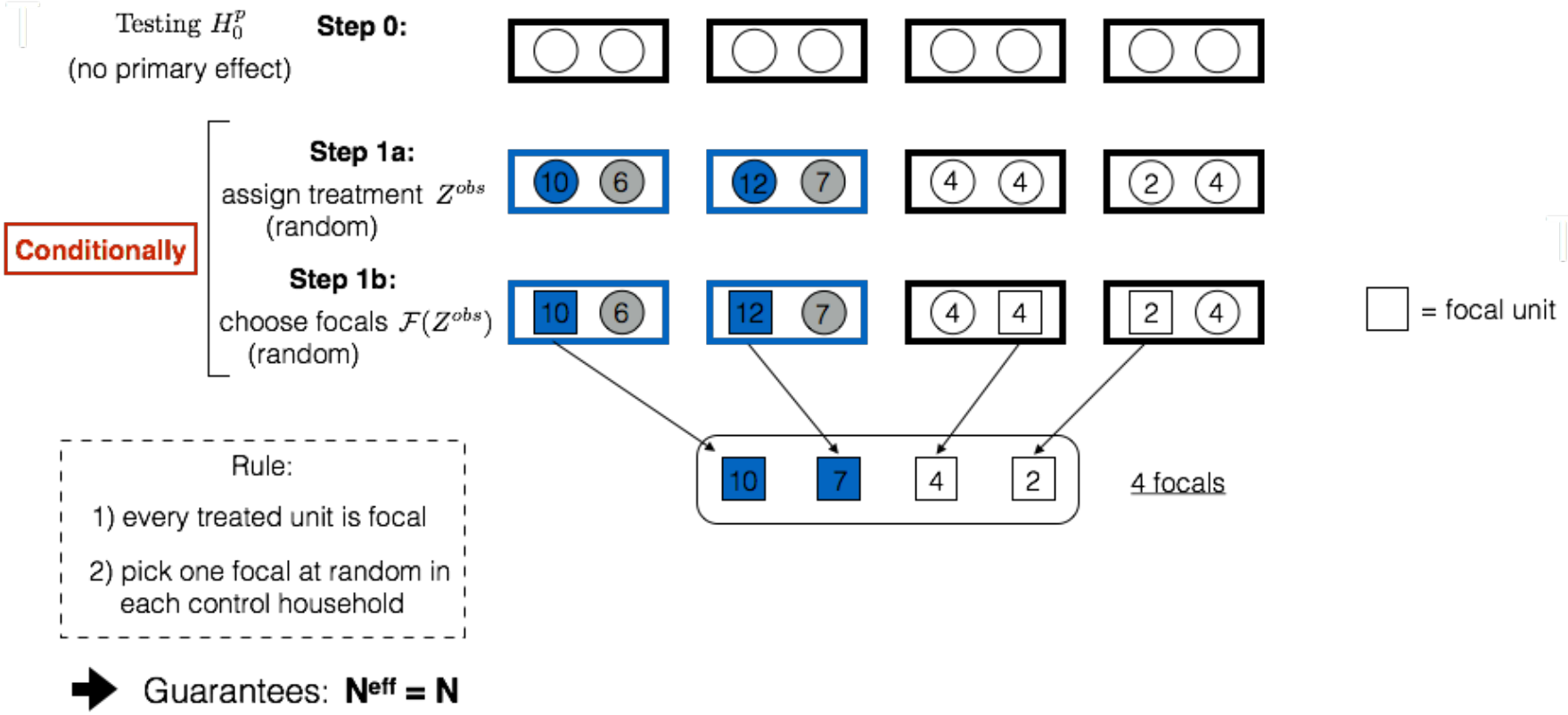
- Need to use **conditional testing**. In particular,
  - Work with a subset of units – called **focal units** by Athey et.al. (2016).
  - Resample within a subset of assignments.
  - Use a test statistic defined only on focal units.

Our methodology in practice: testing primary effect

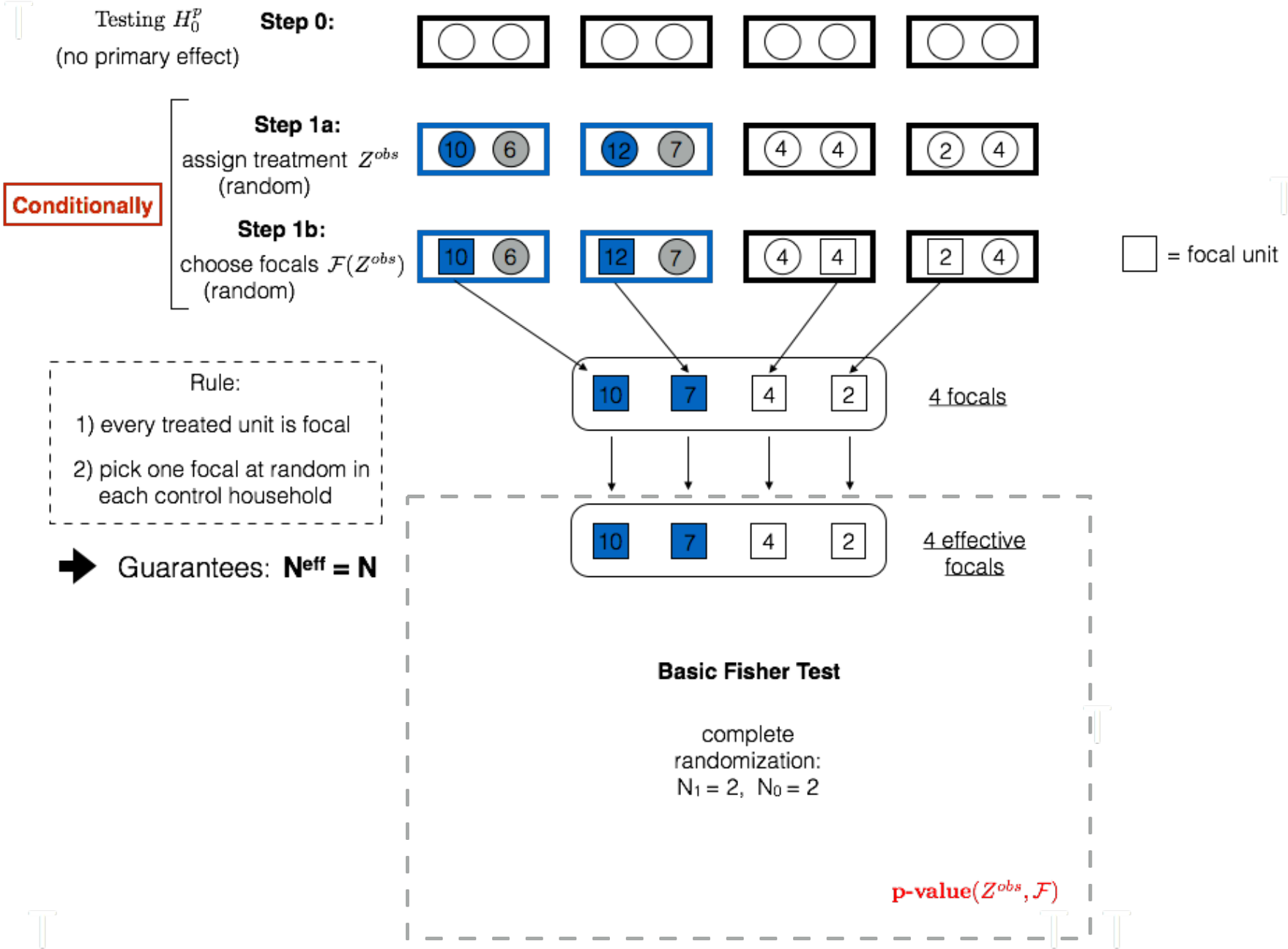
Testing H_0^p
(no primary effect)

Step 0:

Our methodology in practice: testing primary effect



Our methodology in practice: testing primary effect



Our methodological contribution

1. Sample $\mathcal{F}$ = set of units such that every unit in $\mathcal{F}$ is exposed to the treatment levels in H_0 under Z^{obs} – the focal set.

Our methodological contribution

1. Sample $\mathcal{F}$ = set of units such that every unit in $\mathcal{F}$ is exposed to the treatment levels in H_0 under Z^{obs} – the focal set.
2. Calculate $\mathcal{Z}$ = set of assignments for which we can impute missing outcomes under H_0 for all $i \in \mathcal{F}$.

Our methodological contribution

1. Sample $\mathcal{F}$ = set of units such that every unit in $\mathcal{F}$ is exposed to the treatment levels in H_0 under Z^{obs} – the focal set.
2. Calculate $\mathcal{Z}$ = set of assignments for which we can impute missing outcomes under H_0 for all $i \in \mathcal{F}$.
3. The pair $\mathcal{C} = (\mathcal{F}, \mathcal{Z})$ is the **conditioning event**.

Our methodological contribution

1. Sample $\mathcal{F}$ = set of units such that every unit in $\mathcal{F}$ is exposed to the treatment levels in H_0 under Z^{obs} – the focal set.
2. Calculate $\mathcal{Z}$ = set of assignments for which we can impute missing outcomes under H_0 for all $i \in \mathcal{F}$.
3. The pair $\mathcal{C} = (\mathcal{F}, \mathcal{Z})$ is the **conditioning event**.
4. Calculate $T^{\text{obs}} = T_{\mathcal{C}}(Z^{\text{obs}}|y)$, defined only on units in $\mathcal{F}$.
5. In the randomization test conditional on $\mathcal{C}$:
   (i) Sample $Z' \in \mathcal{Z}$ proportional to probability $p(Z')$ from design.
   (ii) Store $T_m = T_{\mathcal{C}}(Z'|y)$.
6. $\texttt{p-value} = E\left(\mathbb{I}\{T_m \geq T^{\text{obs}}\}\right)$.

Our methodological contribution

1. Sample $\mathcal{F}$ = set of units such that every unit in $\mathcal{F}$ is exposed to the treatment levels in H_0 under Z^{obs} – the focal set.
2. Calculate $\mathcal{Z}$ = set of assignments for which we can impute missing outcomes under H_0 for all $i \in \mathcal{F}$.
3. The pair $\mathcal{C} = (\mathcal{F}, \mathcal{Z})$ is the **conditioning event**.
4. Calculate $T^{\text{obs}} = T_{\mathcal{C}}(Z^{\text{obs}}|y)$, defined only on units in $\mathcal{F}$.
5. In the randomization test conditional on $\mathcal{C}$:
   (i) Sample $Z' \in \mathcal{Z}$ proportional to probability $p(Z')$ from design.
   (ii) Store $T_m = T_{\mathcal{C}}(Z'|y)$.
6. $\texttt{p-value} = E\left(\mathbb{I}\{T_m \geq T^{\text{obs}}\}\right)$.

♠ A conditioning mechanism M is formally defined by a set of conditioning events and a probability distribution $p(Z, \mathcal{C})$. The tuple (H_0, M, T) is a (generalized) conditional randomization test.

♠ In the paper, we show in detail some sufficient properties for M and T wrt to the null hypothesis, H_0, in order to have a valid conditional test.

Why it works

- Our test operates conditional on an event $\mathcal{C} = (\mathcal{F}, \mathcal{Z})$.
- To perform the test we simply need to adjust the resampling distribution:

$$p(Z|\mathcal{C}) \sim p(\mathcal{C}|Z) \cdot p(Z). \tag{1}$$

- $p(Z)$ is the design and may not be under our control.
- But $p(\mathcal{C}|Z)$ is defined by the conditioning mechanism.
  - Certain properties need to hold for $p(Z, \mathcal{C})$ and the test statistic to have a valid test (in the paper).

Why it works

- Our test operates conditional on an event $\mathcal{C} = (\mathcal{F}, \mathcal{Z})$.
- To perform the test we simply need to adjust the resampling distribution:

$$p(Z|\mathcal{C}) \sim p(\mathcal{C}|Z) \cdot p(Z). \tag{1}$$

- $p(Z)$ is the design and may not be under our control.
- But $p(\mathcal{C}|Z)$ is defined by the conditioning mechanism.
  - Certain properties need to hold for $p(Z, \mathcal{C})$ and the test statistic to have a valid test (in the paper).

♠ Pros and cons:

- Under (1) the test statistic $T_{\mathcal{C}}(Z|y)$ has the correct conditional distribution!
  - Flexibility in choosing $\mathcal{C}$ to improve on classical conditional randomization methods, and also achieve optimal power.

Why it works

- Our test operates conditional on an event $\mathcal{C} = (\mathcal{F}, \mathcal{Z})$.
- To perform the test we simply need to adjust the resampling distribution:

$$p(Z|\mathcal{C}) \sim p(\mathcal{C}|Z) \cdot p(Z). \tag{1}$$

- $p(Z)$ is the design and may not be under our control.
- But $p(\mathcal{C}|Z)$ is defined by the conditioning mechanism.
  - Certain properties need to hold for $p(Z, \mathcal{C})$ and the test statistic to have a valid test (in the paper).

♠ Pros and cons:

- Under (1) the test statistic $T_{\mathcal{C}}(Z|y)$ has the correct conditional distribution!
  - Flexibility in choosing $\mathcal{C}$ to improve on classical conditional randomization methods, and also achieve optimal power.
- Challenging to devise conditioning mechanisms in practice, and compute the conditional distribution $p(Z|\mathcal{C})$.

$\Rightarrow$ In two-stage randomization it all works out easily!

Problems with classical conditional tests

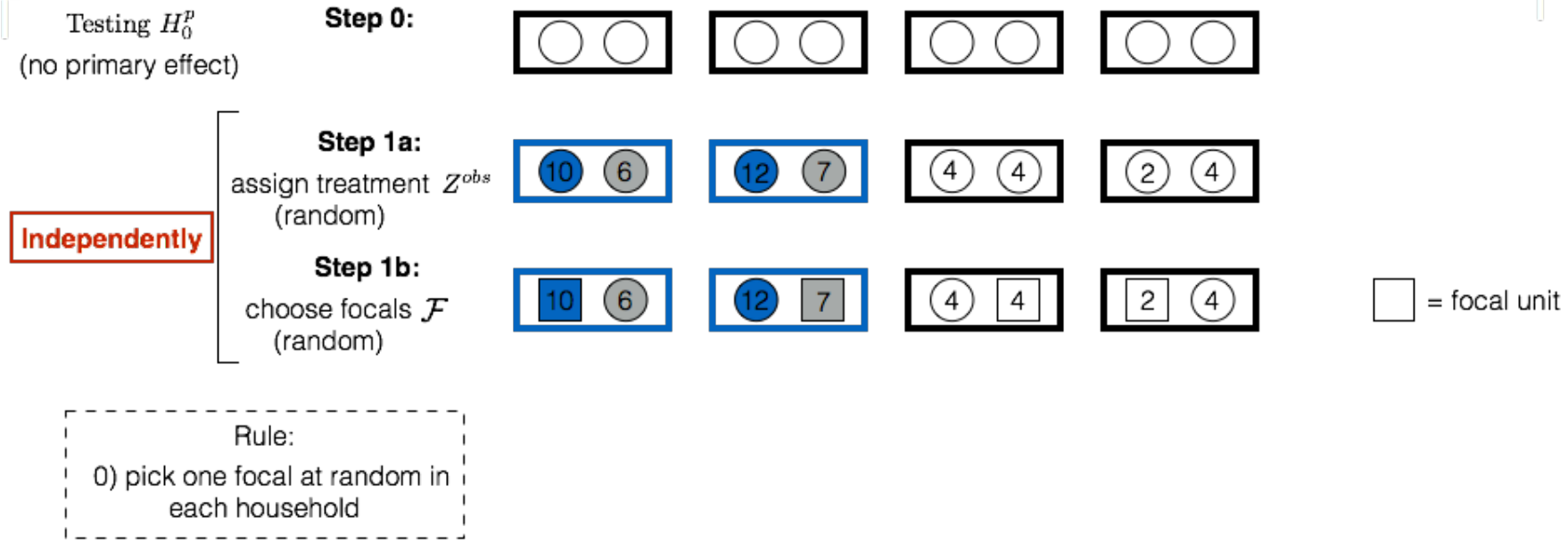
(Aronow, 2012); (Athey et.al., 2016)

Testing H_0^p
(no primary effect)

Step 0:

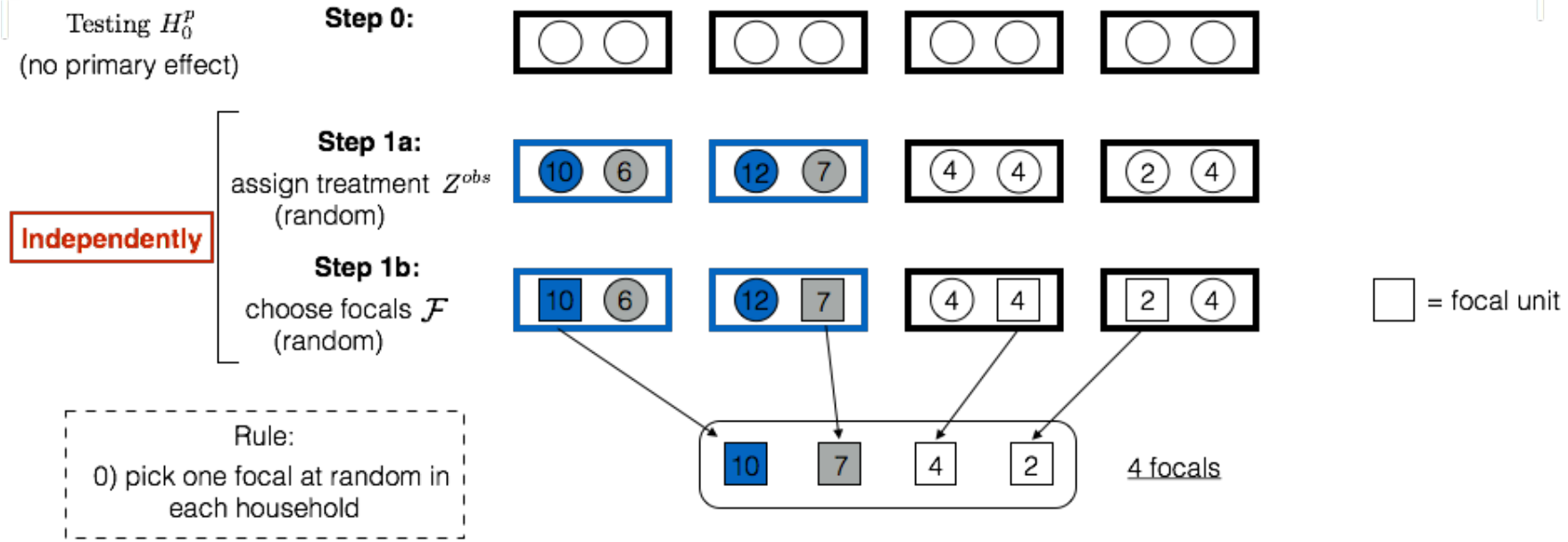
Problems with classical conditional tests

(Aronow, 2012); (Athey et.al., 2016)



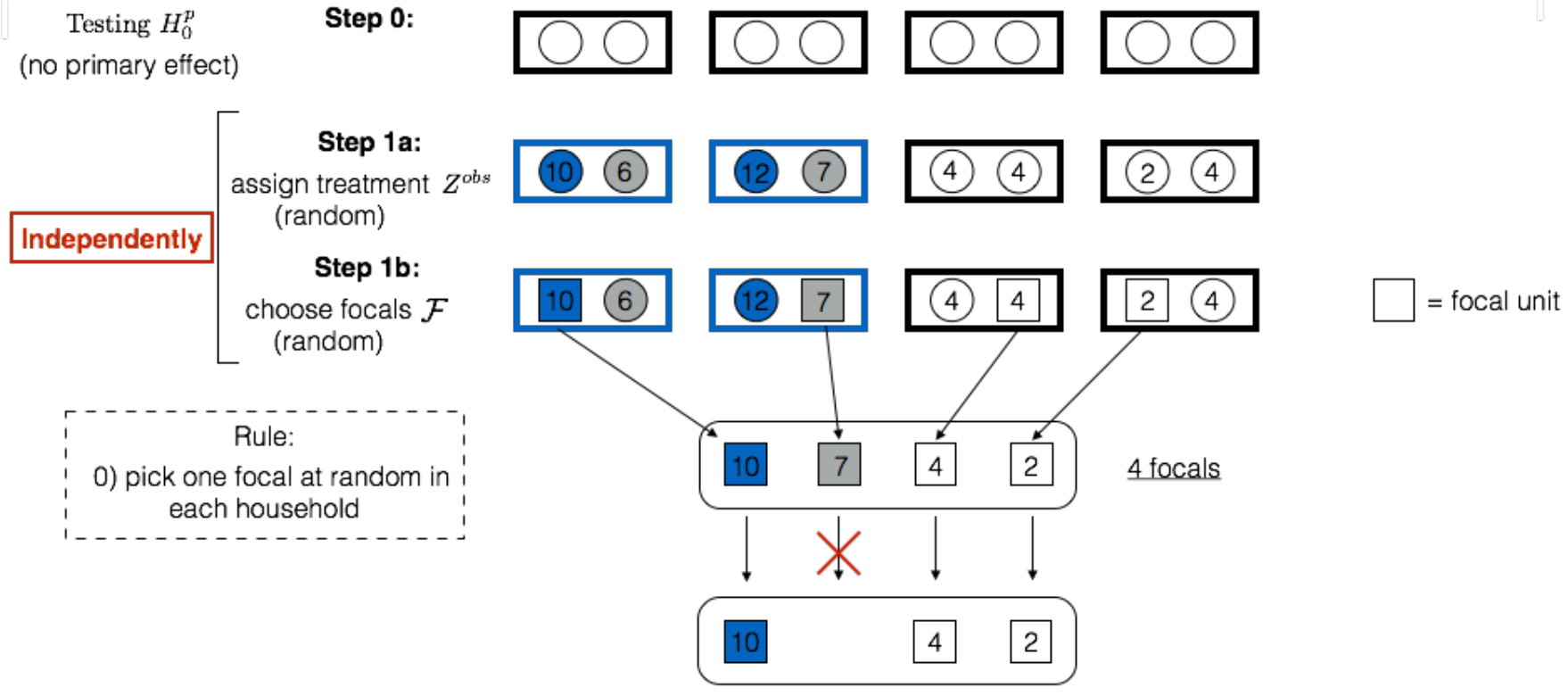
Problems with classical conditional tests

(Aronow, 2012); (Athey et.al., 2016)



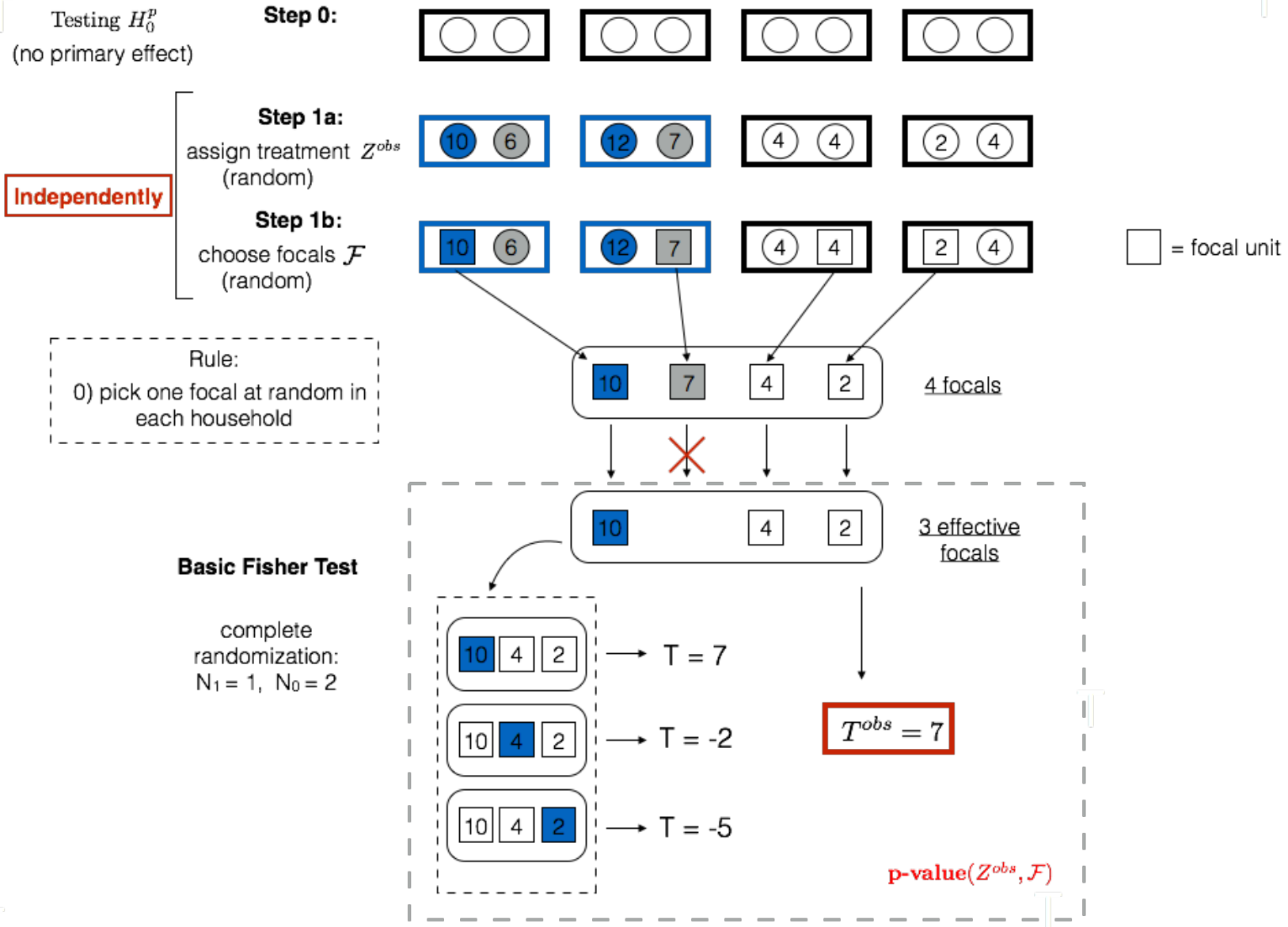
Problems with classical conditional tests

(Aronow, 2012); (Athey et.al., 2016)



Problems with classical conditional tests

(Aronow, 2012); (Athey et.al., 2016)



Problems with classical tests

- Classical conditional tests require that the conditioning events form a partition of the sample space.
- It is then necessary that the focal units are selected **independently** of the observed assignment Z^{obs}.

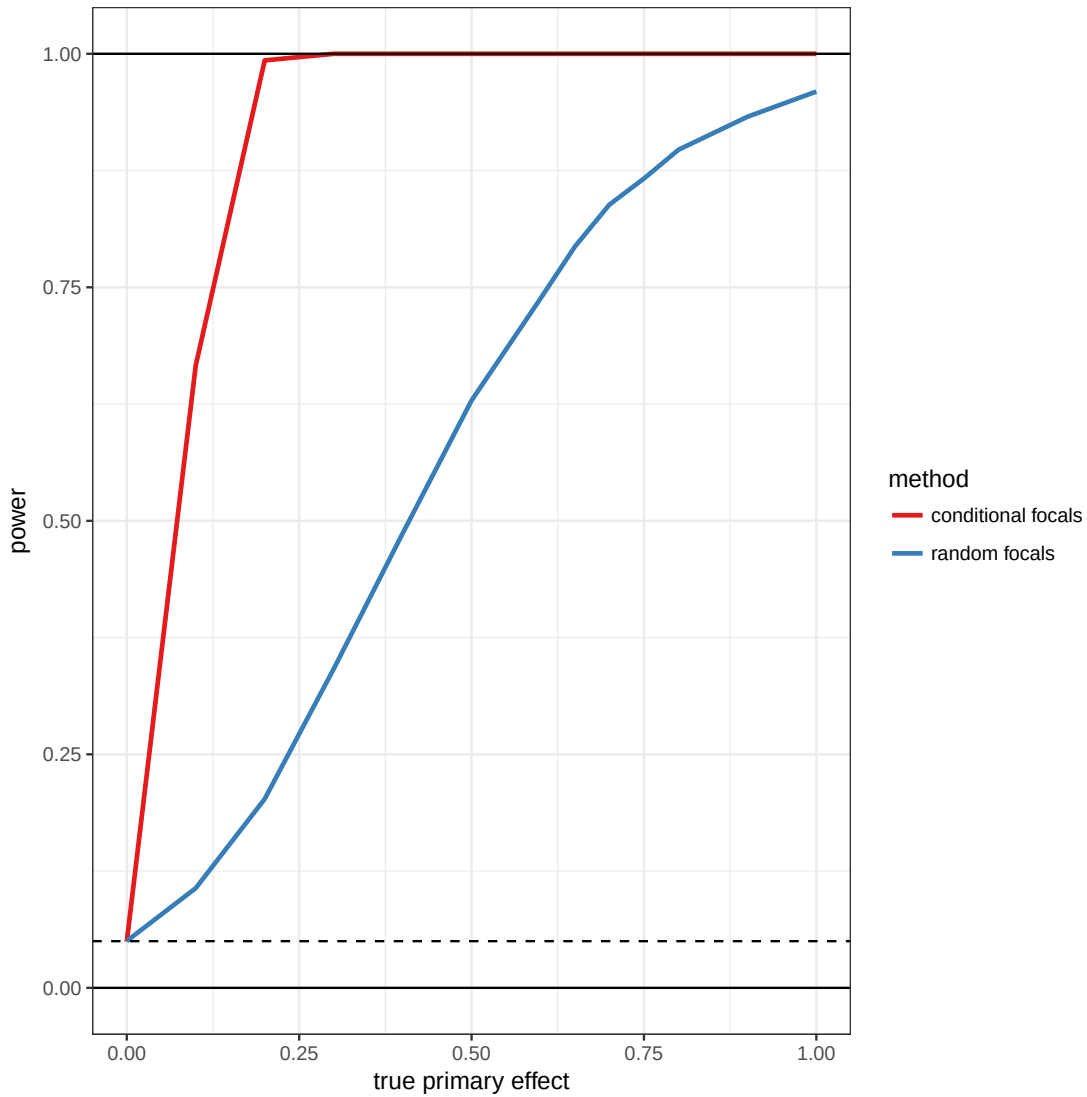
Problems with classical tests

- Classical conditional tests require that the conditioning events form a partition of the sample space.
- It is then necessary that the focal units are selected **independently** of the observed assignment Z^{obs}.
- **The problem:** focal units that are exposed to a treatment not considered in H_0 cannot be used in the test. This leads to loss of information.

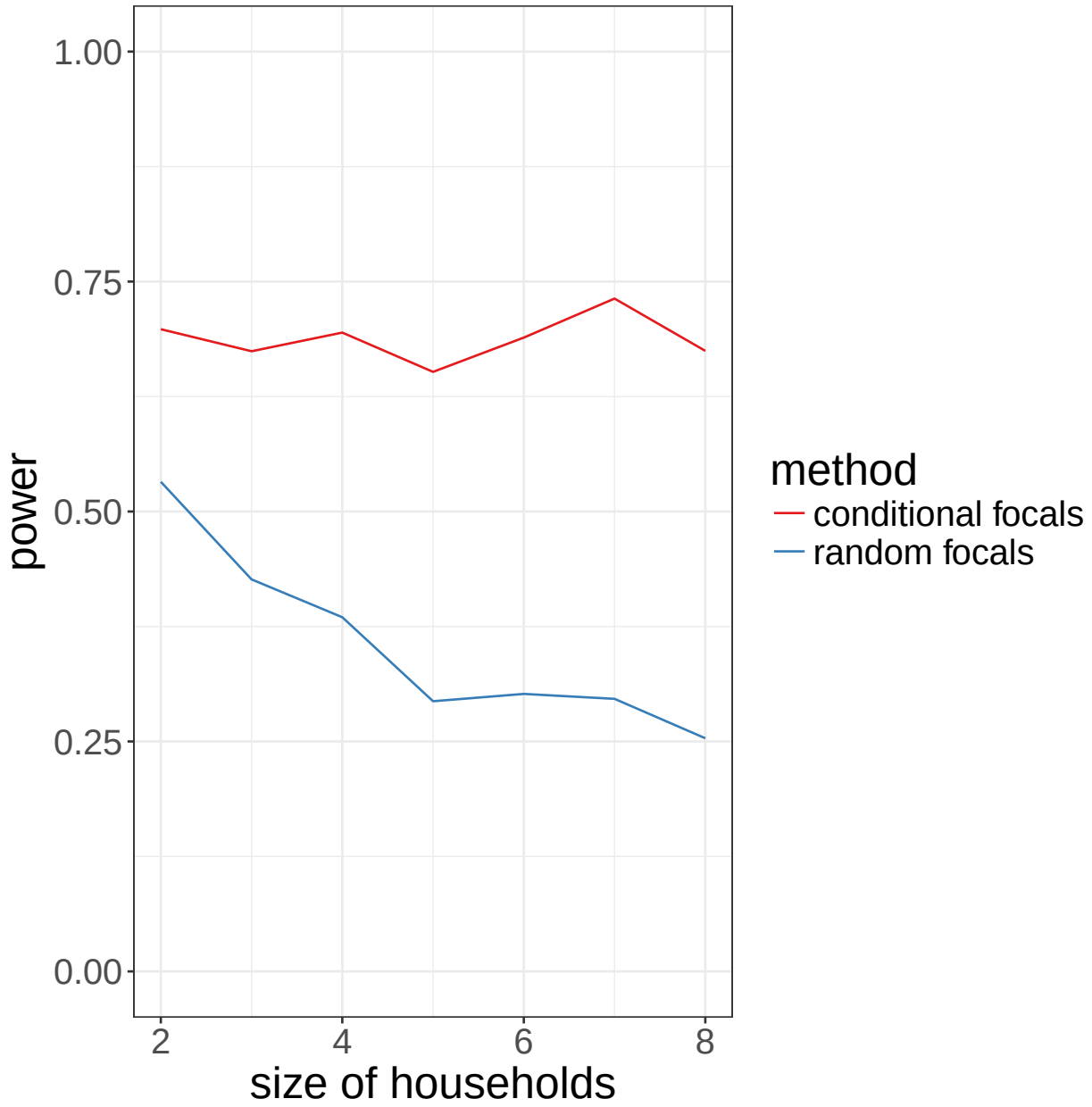
- In our framework we can choose the focal units **conditional on** Z^{obs}. This way we can maximize the number of focal units and assignments considered in the test, and thus improve power.

Simulation – Power for test of no primary effect

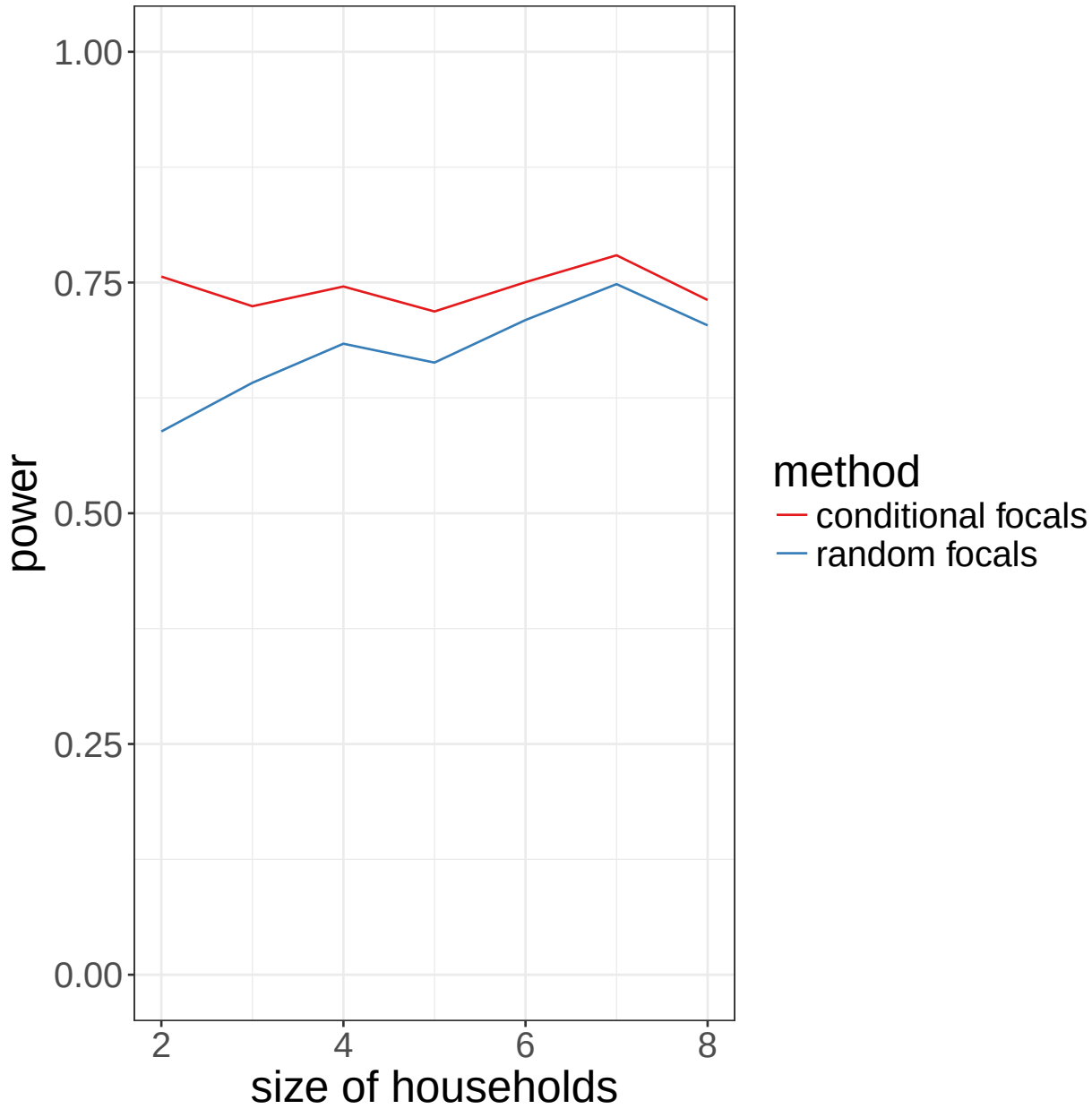
We set 500 households with 10 units each.
Outcome model: $Y_i(1,1) = Y_i(0,0) + \tau$, and $Y_i(0,0) \sim \mathcal{N}(0, \sigma^2)$.



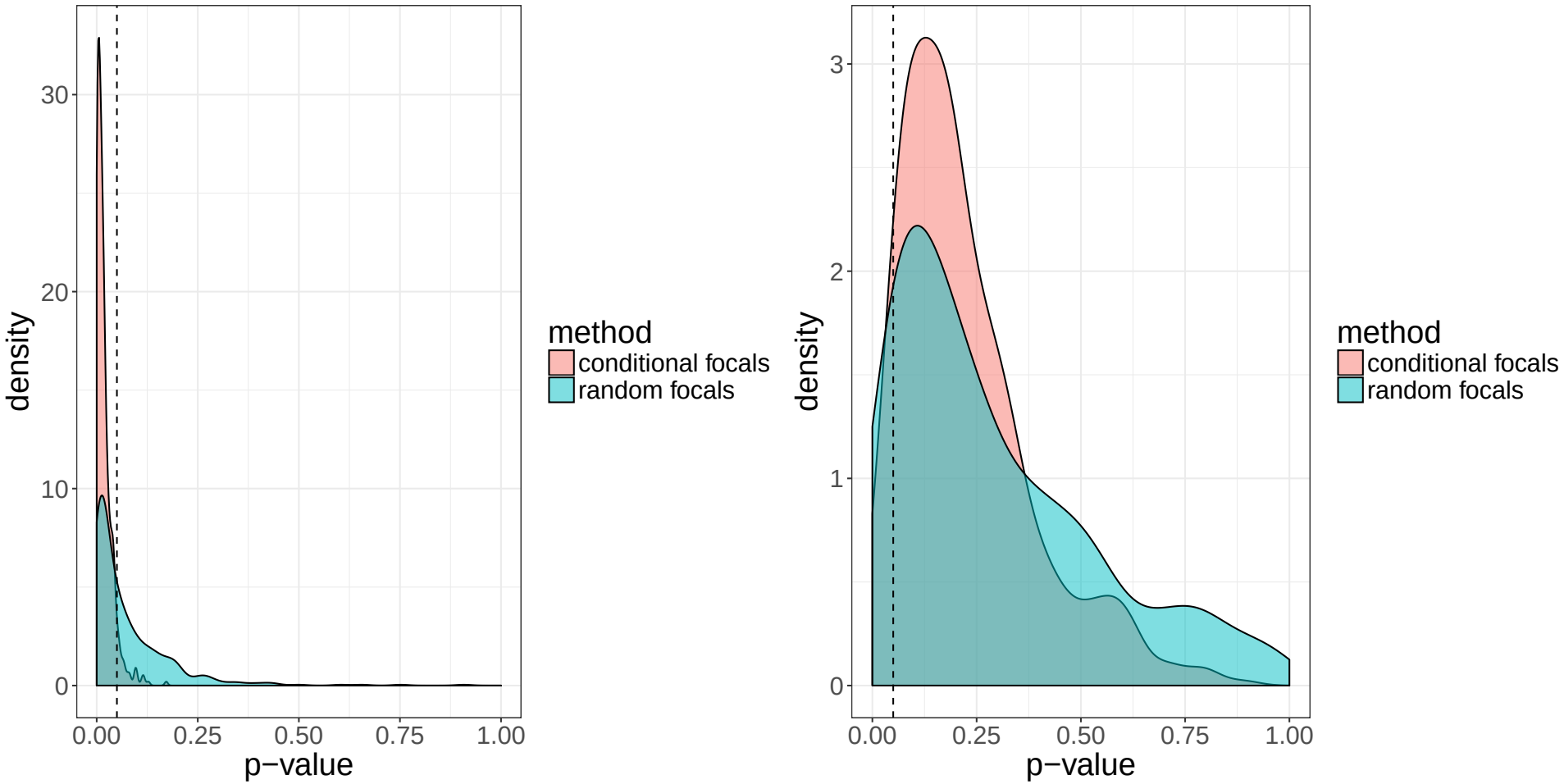
Simulation – Power for test of no primary effect



Simulation – Power for test of no spillover effect



Absenteeism data – distribution of 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: interference presents unique challenges.
- We build a framework that allows flexible conditioning mechanism, which can offer significant increase in testing power.
- Interference is a great application area for such conditional testing mechanisms. In two-stage randomization our conditional testing is simplified to classical permutation tests with restrictions.

Conclusion

- Randomization inference is appealing – makes minimal assumptions.
- But hard: interference presents unique challenges.
- We build a framework that allows flexible conditioning mechanism, which can offer significant increase in testing power.
- Interference is a great application area for such conditional testing mechanisms. In two-stage randomization our conditional testing is simplified to classical permutation tests with restrictions.

Current and future work:

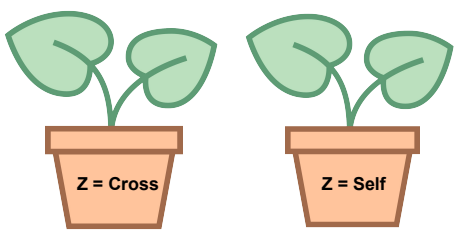
- Aggregate p-values over different selections of focals.
- Extend to more complicated interference.

Conclusion

- Randomization inference is appealing – makes minimal assumptions.
- But hard: interference presents unique challenges.
- We build a framework that allows flexible conditioning mechanism, which can offer significant increase in testing power.
- Interference is a great application area for such conditional testing mechanisms. In two-stage randomization our conditional testing is simplified to classical permutation tests with restrictions.

Current and future work:

- Aggregate p-values over different selections of focals.
- Extend to more complicated interference.



Thanks for your attention!