Selective Randomization Inference for Adaptive Studies

Tobias Freidling

Statistical Laboratory, DPMMS
University of Cambridge

American Causal Inference Conference 17/05/2024

Collaborators

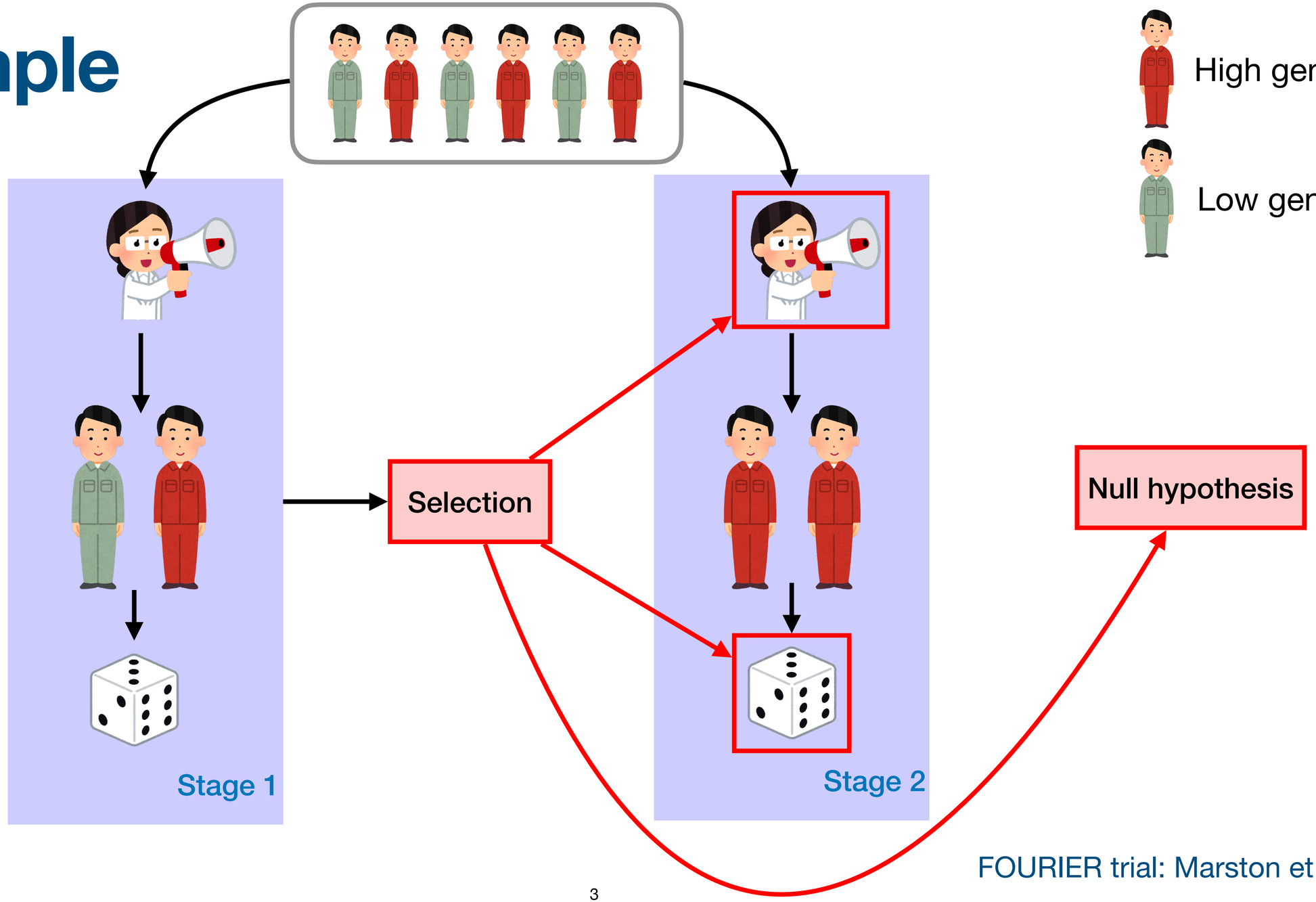
Qingyuan Zhao

Statistical Laboratory, DPMMS
University of Cambridge

Zijun Gao

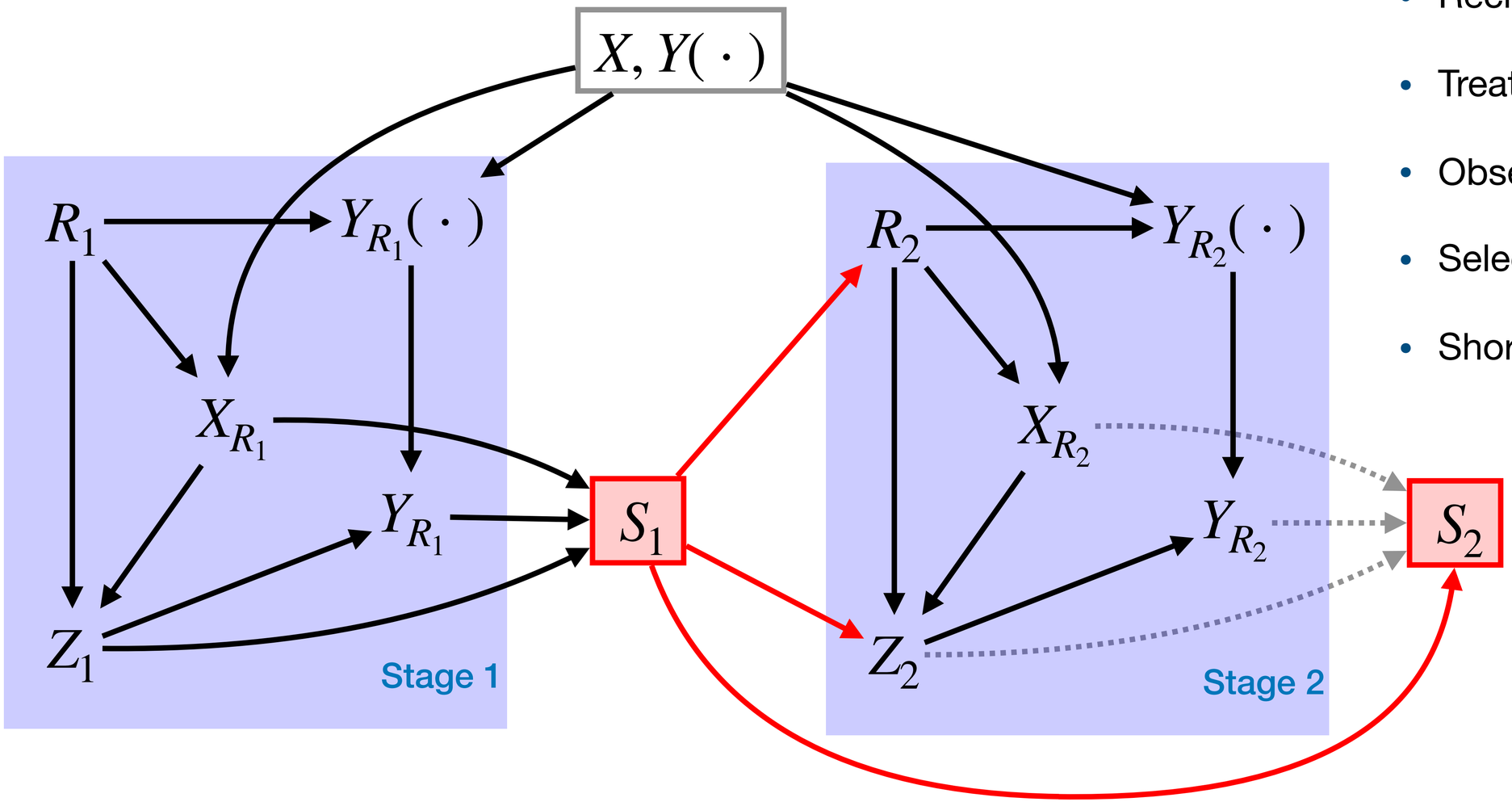
Marshall School of Business
University of Southern California

Example

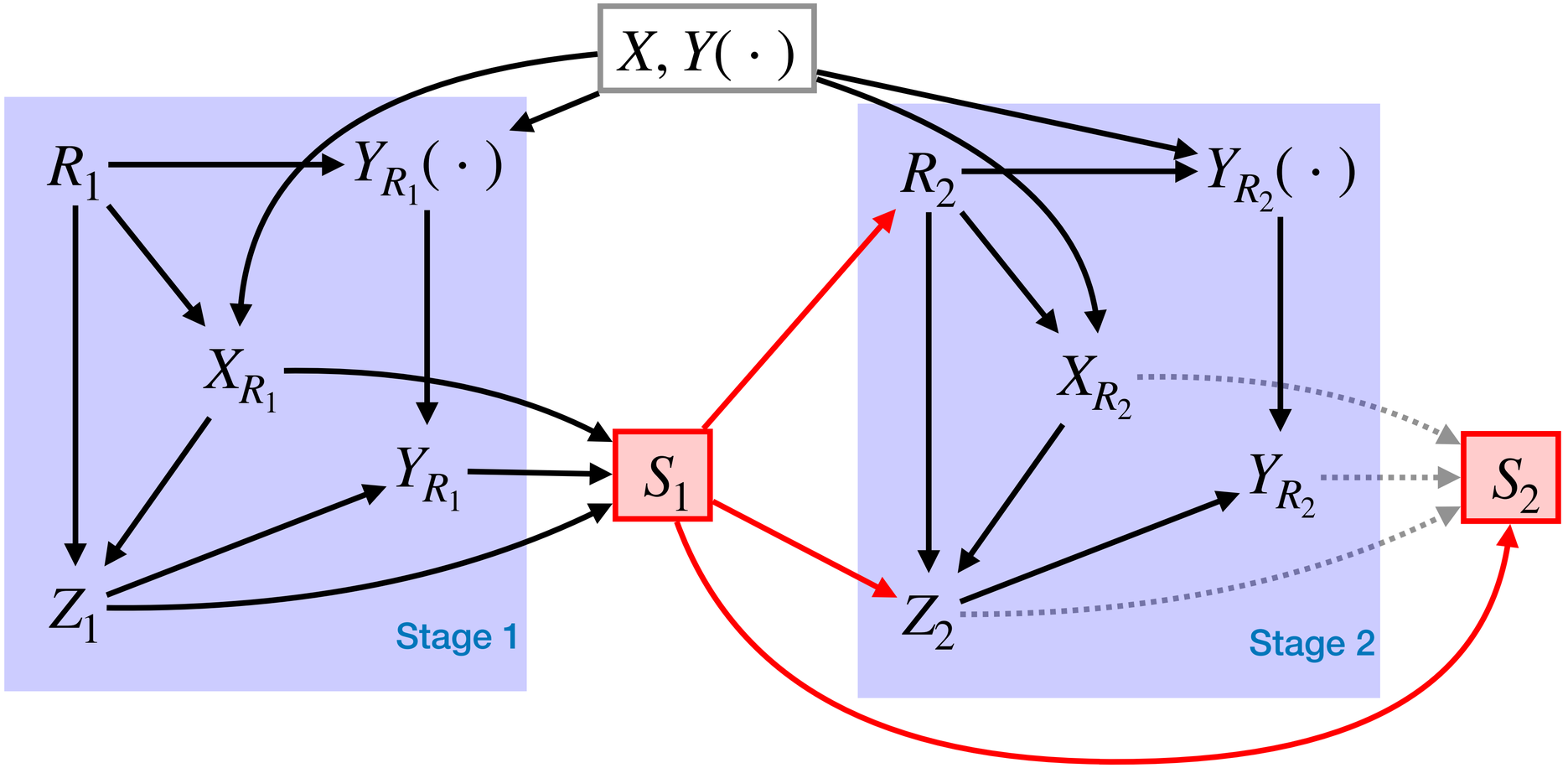


FOURIER trial: Marston et al. (2020)

Graphical Model



- Covariates: X
- Potential outcomes: $Y(\cdot)$
- Recruitment: $R_k \subseteq [n]$
- Treatments: Z_k
- Observed outcomes: $Y = Y(Z)$
- Selective choice: S_k
- Short-hand: $W = (R, X_R, Y_R(\cdot))$



- Assumption (A1): $Z_k \perp\!\!\!\perp Y_{R_{[k]}}(\cdot) \mid R_{[k]}, X_{R_{[k]}}, Y_{R_{[k-1]}}, Z_{[k-1]} \quad \forall\, k \in [K]$
- Assumption (A2): $R_k, X_{R_k}, Y_{R_k}(\cdot) \perp\!\!\!\perp Z_{[k-1]} \mid W_{[k-1]}, S_{k-1} \quad \forall\, k \in [K]$
- Assumption (A1*): $Z_k \perp\!\!\!\perp Z_{[k-1]}, W_{[k-1]}, Y_{R_k}(\cdot) \mid R_k, X_{R_k}, S_{k-1} \quad \forall\, k \in [K]$

Randomization Inference

- Strength: no modelling assumptions, no i.i.d. data
- Distribution of $Z = (Z_1, \ldots, Z_K)$ is known
- Null hypothesis: $Y_i(1) - Y_i(0) = 0$ for all/subset of units
- Condition on W and compare observed value of statistic $T(Z, W)$ against values $T(Z^*, W)$ under alternative treatment assignments Z^*.
- $\mathbb{P}(\, T(Z^*, W) \leq T(Z, W) \mid W, Z), \quad$ where $\quad Z^* \overset{D}{=} Z$ and $Z^* \perp\!\!\!\perp Z \mid W$
- Is there a problem when the experiment is adaptive?

Fisher (1935), Pitman (1937), Zhang & Zhao (2023)

Selective Randomization Inference

- Using data twice (double dipping)
- Comparing to Z^* that choose different stage-II design or null hypothesis than Z
- Result: Type-I error inflation
- Solutions:
  - Data splitting (Cox, 1975): $\mathbb{P}(\, T(Z^*, W) \leq T(Z, W) \mid W, Z, Z_1^* = Z_1)$, where $K = 2$
  - Selective inference (Lee et al., 2016; Fithian et al., 2017): regression models etc.
  - Selective randomization inference:

$$P_{sel} = \mathbb{P}(\, T(Z^*, W) \leq T(Z, W) \mid W, Z,\ S(Z^*) = S(Z)\,)$$

Computability

- Under Assumptions (A1) and (A2), the selective randomization p-value can be computed.
- Formula for the selective randomization distribution under (A1*):
  - $\mathbb{P}(Z = z \mid W = w, S(Z) = s) = \dfrac{q(z \mid w, s)}{\sum_{z'} q(z' \mid w, s)}$, where
  - $q(z \mid w, s) = \mathbf{1}\{S(z) = s\} \cdot \prod_{k=1}^{K} \mathbb{P}(Z_k = z_k \mid R_k = r_k, X_{R_k} = x_{R_k}, S_{k-1} = s_{k-1})$
- Formula for p-value: $P_{sel} = \dfrac{\sum_{z^*} \mathbf{1}\{T(z^*, W) \leq T(Z, W)\}\ q\big(z^* \mid W, S(Z)\big)}{\sum_{z^*} q\big(z^* \mid W, S(Z)\big)}$

Computation

$$P_{sel} = \mathbb{P}(\, T(Z^*, W) \leq T(Z, W) \mid W, Z,\ S(Z^*) = S(Z)\,)$$

- Monte Carlo approximation: Generate M feasible samples $(z_j^*)_{j=1}^M$, i.e. $S(z_j^*) = S(Z)$, and compute

$$\hat{P}_M := \frac{1 + \sum_{j=1}^M \mathbf{1}\{T(z_j^*, W) \leq T(Z, W)\}}{1 + M}\,.$$

- Rejection sampling, Markov Chain Monte Carlo (MCMC)

Inference

$$P_{sel} = \mathbb{P}(\, T(Z^*, W) \leq T(Z, W) \mid W, Z,\ S(Z^*) = S(Z)\,)$$

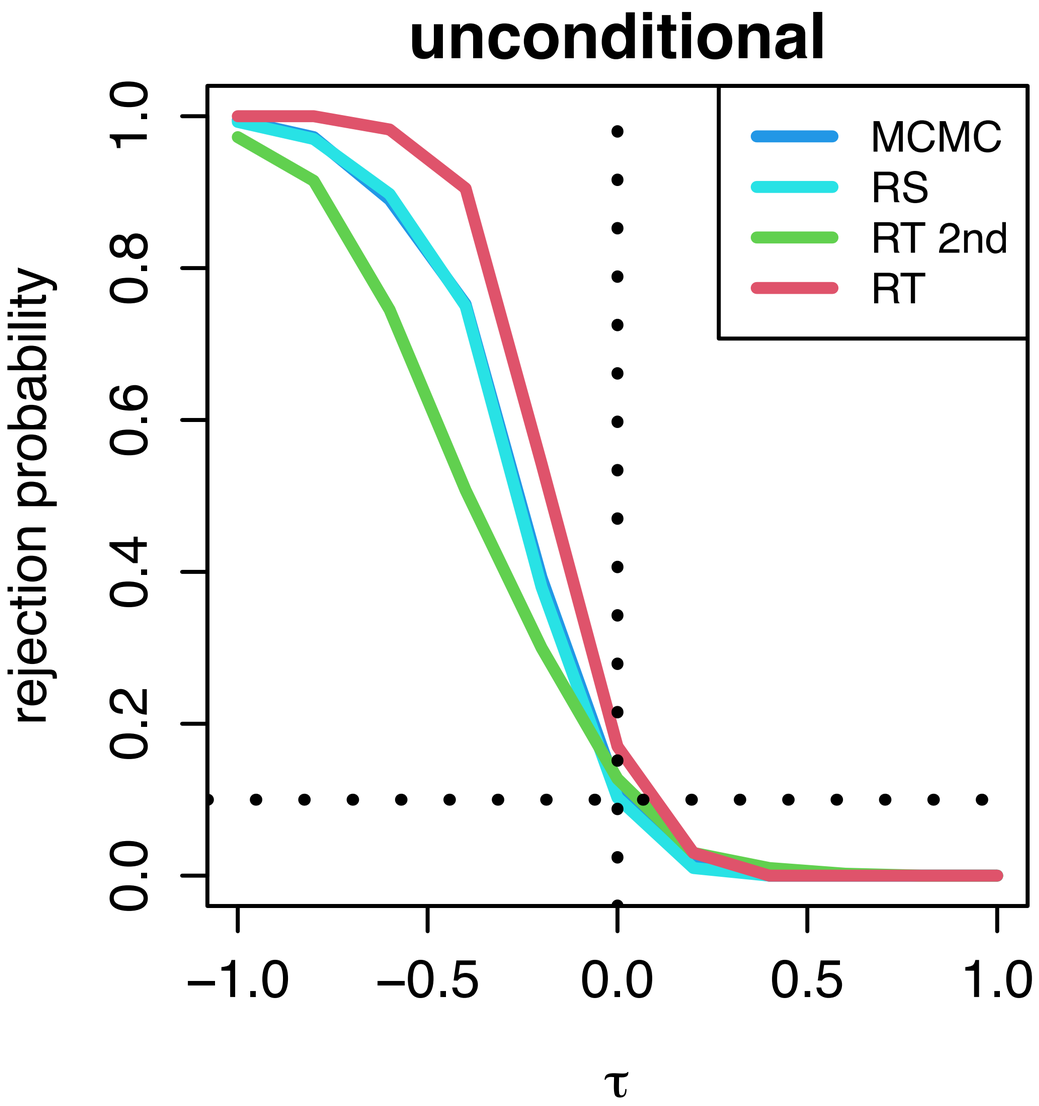
- Confidence intervals:
  - test $Y_i(1) - Y_i(0) = \tau$ for different τ
  - $(1-\alpha)$ confidence interval: $C_{1-\alpha} = \{\tau\colon\ P_{sel}(\tau) \geq \alpha\}$
- Estimation: τ such that $P_{sel}(\tau) = 0.5$
- Data carving: non-adaptive hold-out units

Simulation Study

- 2 stages, 2 treatments $Z_i \in \{0,1\}$, 2 groups $X_i \in \{\text{low}, \text{high}\}$
- Potential outcomes: $Y_i(0) = Y_i(1) \sim N(0,1)$ i.i.d.
- First stage: 100 patients, Second stage: 40 patients
- $\Delta =$ standardized difference in SATEs between groups
- Selection variable:

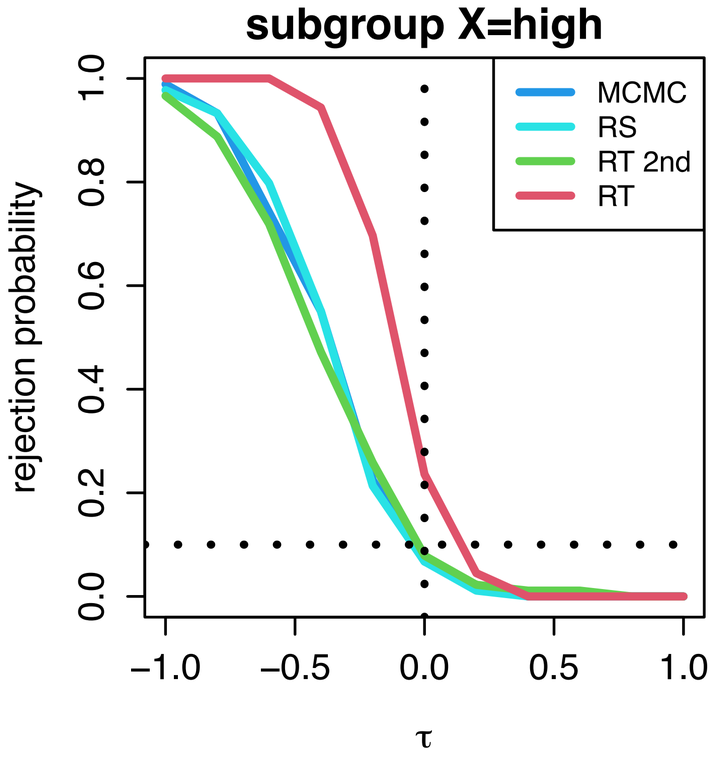
$$S = \begin{cases} \text{only low}, & \Delta < \Phi^{-1}(0.2), & \text{recruit 40 from group } X_i = \text{low} \\ \text{only high}, & \Delta > \Phi^{-1}(0.8), & \text{recruit 40 from group } X_i = \text{high} \\ \text{both}, & \text{otherwise}, & \text{recruit 20 from each group} \end{cases}$$

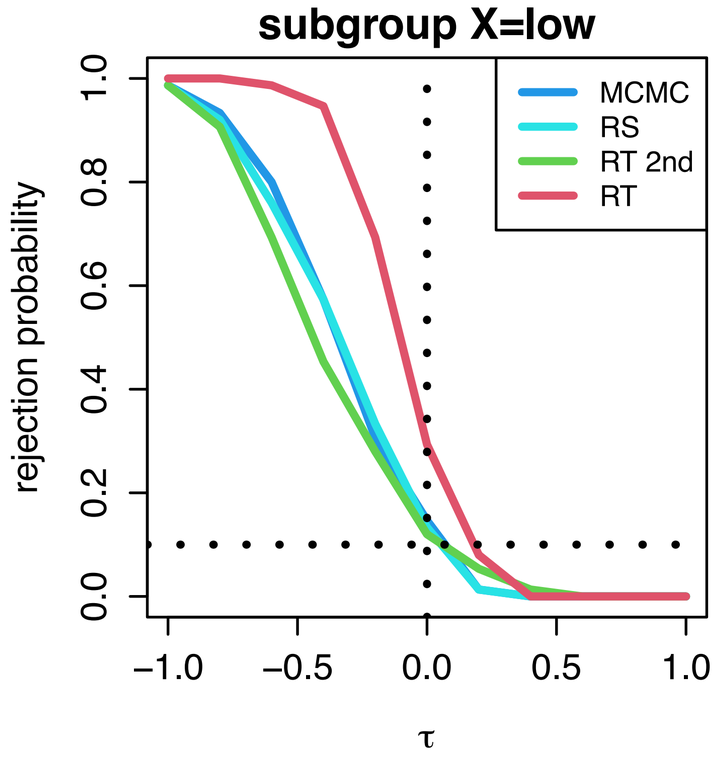
Power Analysis

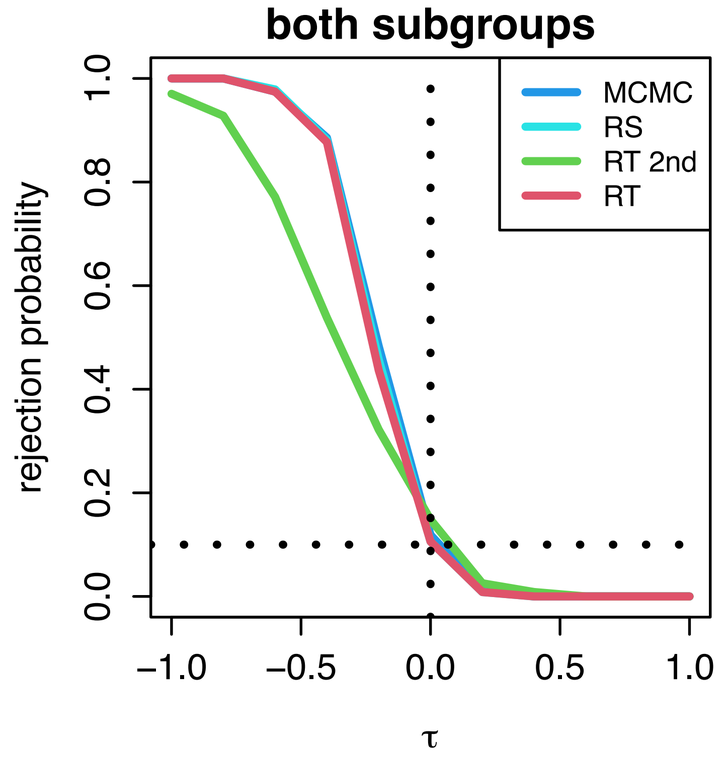


- RT: **no type-I error control**
- RT 2nd: valid but has **low power**
- Selective RT: **valid and more powerful.**
- Rejection sampling and MCMC lead to very similar approximations.

Power Analysis







- Type-I error control in every subgroup
- Gain in power when there is a lot of “randomness left”

Conclusion

- Experiments with adaptive treatments, recruitment and null hypothesis
- Visualization via DAGs
- **Key idea: Conditioning randomization p-value on the selection information**
- Computability under general assumptions
- Approximation via rejection sampling or MCMC

Thanks for your attention!

taf40@cam.ac.uk

References

Cox, D.R. (1975) ‘A note on data-splitting for the evaluation of significance levels’, Biometrika, 62(2), pp. 441–444.

Fisher, R. A. (1935). ‘The design of experiments’, Edinburgh: Oliver & Boyd.

Fithian, W., Sun, D. and Taylor, J. (2017) ‘Optimal Inference After Model Selection’, arXiv:1410.2597

Lee, J.D., Sun, D.L., Sun, Y., Taylor J. (2016) ‘Exact post-selection inference, with application to the lasso’, The Annals of Statistics, 44(3).

Marston, N.A. et al. (2020) ‘Predicting Benefit From Evolocumab Therapy in Patients With Atherosclerotic Disease Using a Genetic Risk Score’, Circulation, 141(8), pp. 616–623.

Pitman, E.J.G. (1937) ‘Significance Tests Which May be Applied to Samples From any Populations’, Supplement to the Journal of the Royal Statistical Society, 4(1), pp. 119–130.

Zhang, Y. and Zhao, Q. (2023) ‘What is a Randomization Test?’, Journal of the American Statistical Association, 0(0), pp. 1–15.

Hold-out Units