Randomized Control Trials and Policy Evaluation

Matteo Bobba
matteo.bobba@tse-fr.eu
Office: T.353

Toulouse School of Economics (TSE)

M2 PPD-EEP-EEE

Spring 2026

Part 2: Econometrics of RCTs

1. The basic framework
   - $\Rightarrow$ Potential outcomes, SUTVA (IR, Ch 1) and covariates
   - $\Rightarrow$ Assignment mechanisms and randomized experiments (IR, Ch 3,4)
2. Statistical analysis of experiments
   - $\Rightarrow$ Completely randomized experiments (IR Ch 5,7)
   - $\Rightarrow$ Stratified randomized experiments (IR Ch 9)
   - $\Rightarrow$ Pairwise randomized experiments (IR Ch 10 & AI Section 6.2)
   - $\Rightarrow$ Clustered randomized experiments (AI Section 8)
   - $\Rightarrow$ Adaptive randomized experiments

Potential outcomes and SUTVA

Causal Inference as a Missing Data Problem

- Population of units, indexed by $i = 1, ..., N$
- Treatment indicator W_i taking values 0 and 1
- For $i \in \{1, ..., N\}$ $\exists$ one realized outcome and one missing potential outcome

$$Y_i^{\text{obs}} = Y_i(W_i) = \begin{cases} Y_i(0) \text{ if } W_i = 0 \\ Y_i(1) \text{ if } W_i = 1 \end{cases}$$

$$Y_i^{\text{miss}} = Y_i(1 - W_i) = \begin{cases} Y_i(1) \text{ if } W_i = 0 \\ Y_i(0) \text{ if } W_i = 1 \end{cases}$$

$\Rightarrow$ Unit-level causal effect $Y_i(1) - Y_i(0)$ is unobserved

Causal Inference as a Missing Data Problem

- Potential outcomes in terms of the observed and missing outcomes

$$Y_i(0) = \begin{cases} Y_i^{\text{miss}} & \text{if } W_i = 1 \\ Y_i^{\text{obs}} & \text{if } W_i = 0 \end{cases}$$

$$Y_i(1) = \begin{cases} Y_i^{\text{miss}} & \text{if } W_i = 0 \\ Y_i^{\text{obs}} & \text{if } W_i = 1 \end{cases}$$

$\Rightarrow$ If we impute the missing outcomes we know all the potential outcomes

$\Rightarrow$ and thus the value of any causal estimand in the population of N units

Potential Vs. Observed Outcomes: An Example

Unit	Potential Outcomes		Causal Effect
	$Y_i(0)$	$Y_i(1)$	$Y_i(1) - Y_i(0)$
Patient #1	1	7	6
Patient #2	6	5	−1
Patient #3	1	5	4
Patient #4	8	7	−1
Average	4	6	2

Unit i	Treatment W_i	Observed Outcome Y_i^{obs}
Patient #1	1	7
Patient #2	0	6
Patient #3	1	5
Patient #4	0	8

- $E(Y_i(1) - Y_i(0)) > 0$, while $E(Y^{\text{obs}} \mid W_i = 1) - E(Y^{\text{obs}} \mid W_i = 0) < 0$

  $\Rightarrow$ Selection bias: $W_i \not\perp (Y_i(1), Y_i(0))$

The Stable Unit Treatment Value Assumption (SUTVA)

- We will generally need to predict the missing potential outcomes
- To do so, we need the following assumption:

  The potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.

SUTVA: No interference

- $W_{-i} = (W_j)_{j \neq i}$: treatment status of all other obs. in the sample except i
- The no interference part of SUTVA requires that

$$W_{-i} \perp (Y_i(1), Y_i(0))$$

$\Rightarrow$ For all y_1, y_0 and w:

$$\Pr(Y_i(1) \leq y_1, Y_i(0) \leq y_0, W_{-i} = w) =$$
$$\Pr(Y_i(1) \leq y_1, Y_i(0) \leq y_0) \Pr(W_{-i} = w)$$

SUTVA: Scale invariance

- Imagine two versions of treatment: $W = \{0, 1, 2\}$
- The scale invariance part of SUTVA requires that

$\Rightarrow Y_i(1) = Y_i(2)$

$$\Rightarrow \begin{cases} \{Y_i(1) & \forall i = 1, ..., M\} \\ \{Y_i(2) & \forall i = M+1, ..., N, M < N\} \end{cases}$$

$\Rightarrow W$ randomized across three values (stochastic treatment)

Example of Possible SUTVA Violations

1. No interference
   - $\Rightarrow$ Fertilizer in one plot may affect yields in contiguous plots
   - $\Rightarrow$ Immunization efficacy may depend on the number of people immunized
   - $\Rightarrow$ Prob(job) after training may be affected by the number of people trained
2. Scale invariance
   - $\Rightarrow$ Endogenous compliance to treatment assignment
   - $\Rightarrow$ Unobserved differences in the method of administering the treatment

The Role of Covariates in RCTs

- Background attributes can help predicting the missing potential outcomes
  - $\Rightarrow$ Test assumptions about the assignment mechanism
  - $\Rightarrow$ Increase estimates' precision by explaining some of the variation in outcomes
  - $\Rightarrow$ Causal effect of the treatment on sub-groups in the population of interest
- Notice that $W_i \perp (Y_i(1), Y_i(0)) \Rightarrow W_i \perp (Y_i(1), Y_i(0) \mid \boldsymbol{X}_i)$
  - $\Rightarrow$ But $W_i \perp (Y_i(1), Y_i(0) \mid \boldsymbol{X}_i) \not\Rightarrow W_i \perp (Y_i(1), Y_i(0))$

Covariance Balance: Sanity Checks for the Randomization

- The distribution of covariates should be the same under treat and control

$$E(X_i \mid W_i = 1) \approx E(X_i \mid W_i = 0),\ \forall X_i \in \boldsymbol{X}_i$$

- A useful implication is that W_i is not predictable by $\boldsymbol{X}_i$

$$E(W_i \mid \boldsymbol{X}_i) = E(W_i)$$

$\Rightarrow$ Both conditions are testable

Improving Precision

- Lets assume that the conditional expectation function (CEF) is linear:

$$E(Y_i \mid W_i, \boldsymbol{X}_i) = \alpha + \beta W_i + \boldsymbol{\gamma}' \boldsymbol{X}_i,\ E(\boldsymbol{X}_i) = 0$$

- OLS estimation recovers ATE even if the regression is incorrectly specified

$$\begin{aligned} \beta &= E(Y_i(1) - Y_i(0)) \\ &= E(Y_i \mid W_i = 1) - E(Y_i \mid W_i = 0) \end{aligned}$$

$\Rightarrow$ $W_i \perp \boldsymbol{X}_i$, even though in finite sample this correlation may differ from zero

Improving Precision

- The Variance of the ATE is

$$V(\beta) = \frac{\sigma^2_{Y|W,X}}{\sum_{i=1}^{N}(W_i - \overline{W})^2}$$

- $\sigma^2_{Y|W,X} < \sigma^2_{Y|W}$: covariates increase precision of the ATE estimator

  $\Rightarrow$ Improvement in precision is not guaranteed in the linear model

- At the cost of loosing (exact) unbiasedness in finite samples

Heterogenous Treatment Effects

- Lets consider the interactive linear regression model

$$E(Y_i \mid W_i, \boldsymbol{X}_i) = \alpha + \beta W_i + \boldsymbol{\delta}' \boldsymbol{X}_i W_i + \boldsymbol{\gamma}' \boldsymbol{X}_i, \; E(\boldsymbol{X}_i) = 0$$

- The Conditional Average Treatment Effect (CATE) is

$$\begin{aligned} \boldsymbol{\delta}(\boldsymbol{X}) &= E(Y_i(1) \mid \boldsymbol{X}_i) - E(Y_i(0) \mid \boldsymbol{X}_i) \\ &= E(Y_i \mid W_i = 1, \boldsymbol{X}_i) - E(Y_i \mid W_i = 0, \boldsymbol{X}_i) \end{aligned}$$

$\Rightarrow$ The vector $\boldsymbol{\delta}(\boldsymbol{X})$ describes the deviation of CATE from the ATE, β

$\Rightarrow$ The interactive approach always delivers improvements in precision

Assignment Mechanisms

Assignment Mechanism

- N units, set $\mathbb{W} = \{0,1\}^N$ of N-vectors with all elements equal to 0 or 1

$\Rightarrow$ The assignment mechanism is a function $P(\mathbf{W}|\mathbf{Y}(0), \mathbf{Y}(1)) \in [0,1]$ such that

$$\sum_{W \in \{0,1\}^N} P(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = 1$$

- $P(\cdot)$ is the probability that a value for the joint assignment will occur
  - $\Rightarrow$ It is not the probability of a particular unit receiving the treatment
  - $\Rightarrow$ Some assignment vectors $\mathbf{W}$ may have zero probability

Assignment Probability and Propensity Score

- The unit-level assignment probability is:

$$p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = \sum_{\mathbf{W}: W_i = 1} P(\mathbf{W} | \mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$$

- The propensity score at x is the average $p_i(\cdot)$ for units with $X_i = x$

$$e(x) = \frac{1}{N(x)} \sum_{i: X_i = x} p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$$

$\Rightarrow$ $N(x) = \sum_{i: X_i = x} \mathbf{1}_{X_i = x}$

$\Rightarrow$ For values x with $N(x) = 0$, the propensity score is defined to be zero

$\Rightarrow$ $p_i(\mathbf{Y}(0), \mathbf{Y}(1)) = e$ in randomized experiments with no covariates

Example of Assignment Mechanism 1

- Two units, and so (2^2) possible values for $\mathbf{W}$

$$\mathbb{W} \in \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}$$

$\Rightarrow$ The assignment mechanism is equal to $P(\mathbf{W}|\mathbf{Y}(0), \mathbf{Y}(1)) = \frac{1}{4}$

$\Rightarrow$ Unit assignment probability $p_i = p = \frac{1}{2}$ for $i = 1, 2$

$\Rightarrow$ No covariates, and so propensity score $e = p = \frac{1}{2}$

Example of Assignment Mechanism 2

- Two units where only assignments with one treated and one control unit

$\Rightarrow$ The assignment mechanism is

$$P(\mathbf{W}|\mathbf{Y}(0),\mathbf{Y}(1)) = \begin{cases} 1/2 & \text{if } \mathbf{W} \in \left\{ \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\} \\ 0 & \text{if } \mathbf{W} \in \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\} \end{cases}$$

$\Rightarrow$ As before, $e = p = \frac{1}{2}$

Restrictions on the Assignment Mechanism

1. Individualistic

$$p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = q(X_i, Y_i(0), Y_i(1)),\ q(\cdot) \in [0, 1]$$

2. Probabilistic

$$0 < p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) < 1$$

3. Unconfounded

$$P(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = P(\mathbf{W}|\mathbf{X}, \mathbf{Y}'(0), \mathbf{Y}'(1)) = P(\mathbf{W}|\mathbf{X})$$

Randomized Experiments

- A regular assignment mechanism satisfies the three restrictions
- Regular assignment mechanisms with known $P(\cdot)$ are randomized experiments
- Individualistic + unconfounded $\rightarrow$ the assignment mechanism is

$$P(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = c \cdot \prod_{i=1}^{N} q(X_i)^{W_i} (1 - q(X_i))^{1-W_i}$$

$\Rightarrow$ The constant c ensures that the probabilities add to unity

$\Rightarrow$ $e(x) = p_i(x) = q(x)$

An Example: Bernoulli Trials (Coin Tossing)

- $p = e = 0.5, \mathbb{W}^+ = \{0,1\}^N$

$$P(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = 0.5^N$$

- More generally, with probability of assignment to treatment $\neq 0.5$

$$P(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = q^{N_t}(1-q)^{N_c}$$

$\Rightarrow$ No way to ensure "enough" treated and control units under each assignment

Taxonomy of Randomized Experiments

Basic Designs

- Let the set of possible assignments $\mathbf{W}$ with positive prob. be denoted by $\mathbb{W}^+$
- Randomization designs can be characterized by restrictions on $\mathbb{W}^+$

  $\Rightarrow$ Completely randomized experiments

  $\Rightarrow$ Stratified randomized experiments

  $\Rightarrow$ Pairwise randomized experiments

Completely Randomized Experiments

- Draw N_t units at random, such that $1 \leq N_t \leq N-1$
- $q = \frac{N_t}{N}$, and the number of possible assignments is $\binom{N}{N_t}$
- The set of possible assignment vectors is

$$\mathbb{W}^+ = \left\{ \mathbf{W} \in \mathbb{W} \mid \sum_{i=1}^{N} W_i = N_t \right\}$$

$\Rightarrow$ Possible issue with covariate unbalancedness after treatment assignment

Stratified Randomized Experiments

- The population of units is first partitioned into blocks or strata $B_i = B(\boldsymbol{X}_i)$
- Within each block, we conduct a completely randomized experiment
- The set of possible assignment vectors is

$$\mathbb{W}^+ = \left\{ \mathbf{W} \in \mathbb{W} \mid \sum_{i:B_i=j}^{N} W_i = N_t(j) \right\}$$

$\Rightarrow$ More precise inference as units of same type are in different treat arms

Paired Randomized Experiments

- As many units as treatments within each block
- $N(j) = 2$ and $N_t(j) = 1$ for $j = 1, ..., N/2$, so that

$$\mathbb{W}^+ = \left\{ \mathbf{W} \in \mathbb{W} \mid \sum_{i:B_i=j}^{N} W_i = 1 \right\}.$$

$\Rightarrow$ Useful design when N is small and/or J is large

Number of Possible Values for the Assignment Vector

Type of Experiment and Design	Number of Possible Assignments Cardinality of $\mathbb{W}^+$	Number of Units (N) in Sample			
		4	8	16	32
Bernoulli trial	2^N	16	256	65,536	4.2×10^9
Completely randomized experiment	$\binom{N}{N/2}$	6	70	12,870	0.6×10^9
Stratified randomized experiment	$\binom{N/2}{N/4}^2$	4	36	4,900	0.2×10^9
Paired randomized experiment	$2^{N/2}$	4	16	256	65,536

More Sophisticated Designs

- Popular experimental designs in the social sciences

  $\Rightarrow$ Clustered randomized experiments

  $\Rightarrow$ Randomized saturation experiments

  $\Rightarrow$ Adaptive randomized experiments

Clustered Randomized Experiments

- Clusters are defined by partitioning the covariate space $G_{ig} = G(\boldsymbol{X}_i)$
- $\overline{W}_g = \sum_{i:G_{ig}=1} \frac{W_i}{N_g} \in \{0, 1\}$ is the average value of W_i for units in cluster g
- The assignment mechanisms concerns groups of units (clusters)

$$\mathbb{W}^+ = \left\{ \mathbf{W} \in \mathbb{W} \mid \sum_{g=1}^{G} \overline{W}_g = G_t \right\}$$

$\Rightarrow$ Relax SUTVA within clusters, but maintain it across clusters

Randomized Saturation Experiments

- A variant of clustered randomization where
  - $\Rightarrow$ Assign each cluster to a treatment saturation, $S_g = \sum_{i \in g} W_{ig} \in [0, 1)$
  - $\Rightarrow$ Assign each individual to a treatment status $W_{ig} = \{0, 1\}$ according to S_g
- This design is aimed at measuring local spillovers/equilibrium effects
- Optimal combination of clustered and stratified designs

Adaptive Randomized Experiments

- Start with an initial treatment assignment on a small wave of data
- Repeated cross-sections $t = 1, ..., T$ of sample sizes N_t
- Treatment assignment in wave t depend on earlier outcomes $(Y^1_{it-1}, ..., Y^k_{it-1})$
  - $\Rightarrow$ Rely on algorithms to shift to the best performing arm(s) at each round
  - $\Rightarrow$ Detect the best-performing treatment more efficiently than static design

Statistical Analysis of Experiments

- For each randomization design, we'll consider two complementary approaches

1. Asymptotic inference
   - $\Rightarrow$ (semi-) Parametric models for the conditional mean of observed outcomes
   - $\Rightarrow$ Observed outcomes vary through random sampling from a population of units
   - $\Rightarrow$ Approximate distribution of the test statistic through large-sample properties

2. Randomization inference (aka Fisher's exact p-values)
   - $\Rightarrow$ Nonparametric (no restrictions on the distribution of the potential outcomes)
   - $\Rightarrow$ Treatment assignments are the sole source of randomness
   - $\Rightarrow$ Assignment mechanism determines exact distribution of the test statistic

Completely Randomized Experiments

An Example

Unit	Potential Outcomes				
	Cough Frequency (cfa)		Observed Variables		
	$Y_i(0)$	$Y_i(1)$	W_i	X_i (cfp)	Y_i^{obs} (cfa)
1	?	3	1	4	3
2	?	5	1	6	5
3	?	0	1	4	0
4	4	?	0	4	4
5	0	?	0	1	0
6	1	?	0	5	1

The Sharp Null

$$H_0 \ : \ Y_i(0) = Y_i(1) \ \forall i = 1, ..., 6.$$

- One test statistics from example above is

$$\begin{aligned} T(\mathbf{W}, \mathbf{Y}^{\text{obs}}) =& \mid \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} \mid \\ =& \mid (Y_1^{\text{obs}} + Y_2^{\text{obs}} + Y_3^{\text{obs}})/3 - (Y_4^{\text{obs}} + Y_5^{\text{obs}} + Y_6^{\text{obs}})/3 \mid \\ =& \mid 8/3 - 5/3 \mid = 1.00 \end{aligned}$$

- Under H_0, all missings in potential outcomes inferred from obs. outcomes

Treatment Assignments Generate Test Distribution

- We can re-do this for $\binom{6}{3} = 20$ permutations of treatment assignments
  - $\Rightarrow$ E.g. instead of $\mathbf{W}^{\text{obs}} = (1, 1, 1, 0, 0, 0)$ take $\tilde{\mathbf{W}} = (0, 1, 1, 0, 0, 1)$
  - $\Rightarrow$ No change in observed outcomes since $Y_i(0) = Y_i(1) = Y_i^{\text{obs}}$ under H_0

- The value of the test statistic for $\tilde{W}$ is

$$\begin{aligned} T(\tilde{\mathbf{W}}, \mathbf{Y}^{\text{obs}}) =& \mid (Y_2^{\text{obs}} + Y_3^{\text{obs}} + Y_6^{\text{obs}})/3 - (Y_1^{\text{obs}} + Y_4^{\text{obs}} + Y_5^{\text{obs}})/3 \mid \\ =& \mid 6/3 - 7/3 \mid = 0.33 \end{aligned}$$

Treatment Assignments Generate Test Distribution

						Statistic: Absolute Value of Difference in Average	
W_1	W_2	W_3	W_4	W_5	W_6	Levels (Y_i)	Ranks (R_i)
0	0	0	1	1	1	−1.00	−0.67
0	0	1	0	1	1	−3.67	−3.00
0	0	1	1	0	1	−1.00	−0.67
0	0	1	1	1	0	−1.67	−1.67
0	1	0	0	1	1	−0.33	0.00
0	1	0	1	0	1	2.33	2.33
0	1	0	1	1	0	1.67	1.33
0	1	1	0	0	1	−0.33	0.00
0	1	1	0	1	0	−1.00	−1.00
0	1	1	1	0	0	1.67	1.33
1	0	0	0	1	1	−1.67	−1.33
1	0	0	1	0	1	1.00	1.00
1	0	0	1	1	0	0.33	0.00
1	0	1	0	0	1	−1.67	−1.33
1	0	1	0	1	0	−2.33	−2.33
1	0	1	1	0	0	0.33	0.00
1	1	0	0	0	1	1.67	1.67
1	1	0	0	1	0	1.00	0.67
1	1	0	1	0	0	3.67	3.00
1	1	1	0	0	0	**1.00**	**0.67**

Note: Observed values in boldface (R_i is rank(Y_i)). Data based on cough frequency for first six units from honey study.

Computation of Exact p-values

- Calculate the value of the statistic for each assignment vector
- In previous example, each assignment vector has prior probability=1/20
- How unusual is $T(\mathbf{W}, \mathbf{Y}^{\text{obs}}) = 1.00$ under the sharp null hypothesis?
  - $\Rightarrow$ There are 16/20 assignments with $T(\tilde{\mathbf{W}}, \mathbf{Y}^{\text{obs}}) > T(\mathbf{W}, \mathbf{Y}^{\text{obs}})$: p-value $= 0.80$
  - $\Rightarrow$ In this example, the observed difference could well be due to chance

Computation of Approximate p-values

- Recall that the number of distinct values of the treatment vector is $\binom{N_c+N_t}{N_t}$
  - $\Rightarrow$ For instance, if $N = 100$ and $q = 0.5$ then $dim(\mathbb{W}^+) = e^{29}$
- We thus need to rely on numerical approximations to calculate the p-value
  - $\Rightarrow$ Draw vector $\mathbf{W_k}$ with $N - N_t$ zeros and N_t ones from $\mathbb{W}^+$
  - $\Rightarrow$ Compute $T^{\text{dif},k} = T(\mathbf{W_k}, \mathbf{Y}^{\text{obs}}) = \mid \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} \mid$
  - $\Rightarrow$ Repeat this process $K - 1$ times and approximate the p-value by:

$$\widehat{p} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}_{T^{\text{dif},k} \geq T^{\text{dif,obs}}}$$

What is a Reasonable Choice of K?

Number of Simulations	P-Value	$\widehat{(\text{s. e.})}$
100	0.010	(0.010)
1,000	0.044	(0.006)
10,000	0.044	(0.002)
100,000	0.042	(0.001)
1,000,000	0.043	(0.000)

Note: Statistic is absolute value of difference in average ranks of treated and control cough frequencies. P-value is proportion of draws at least as large as observed statistic.

The Choice of the Null Hypothesis

- Fisher's sharp null hypothesis $\neq$ average null hypothesis
  - $\Rightarrow$ Treat effect is positive for some units and negative for others
  - $\Rightarrow$ This does not imply that the average null hypothesis is less relevant
- Fisher's approach can accommodate other sharp null hypotheses. E.g.:

$$H_0 \ : \ Y_i(1) = Y_i(0) + C_i \ \forall i = 1, ..., N$$

  - $\Rightarrow$ We will focus on the sharp null hypothesis of no effect whatsoever

Test Statistic

- Test statistic is any scalar function $T(\mathbf{W}, \mathbf{Y}^{\text{obs}}, \mathbf{X})$ used to find a p-value
- Not all statistics have same ability to distinguish null and alternative hp
  - $\Rightarrow$ A statistic has power if it takes values that are large when the null is false
  - $\Rightarrow$ Use only one statistic (specified before seeing the data) and its p-value

The Choice of Statistic

1. Absolute values of the difference in average outcomes

$$T^{\text{dif}} = \mid \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} \mid = \left| \frac{\sum_{i:W_i=1} Y_i^{\text{obs}}}{N_t} - \frac{\sum_{i:W_i=0} Y_i^{\text{obs}}}{N_c} \right|$$

$\Rightarrow$ Additive treatment effect and distributions of $Y_i(0)$ and $Y_i(1)$ have few outliers

2. Log transform of T^{dif}

$$T^{\text{log}} = \left| \frac{\sum_{i:W_i=1} ln(Y_i^{\text{obs}})}{N_t} - \frac{\sum_{i:W_i=0} ln(Y_i^{\text{obs}})}{N_c} \right|$$

$\Rightarrow$ Multiplicative treatment effect and distributions of $Y_i(0)$ and $Y_i(1)$ are skewed

The Choice of Statistic

3. Quantiles

$$T^{\text{median}} = \mid \text{med}_t(Y_i^{\text{obs}}) - \text{med}_c(Y_i^{\text{obs}}) \mid$$

$\Rightarrow$ More robust to outliers

4. T-Statistics

$$T^{\text{t-stat}} = \left| \frac{\overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}}{\sqrt{\sigma_t^2/N_c + \sigma_c^2/N_t}} \right|$$

$\Rightarrow$ Allows direct comparison with asymptotic student-t (Randomization-t)

The Choice of Statistic

5. Rank Statistic

$$T^{\text{rank}} = \mid \overline{R_t} - \overline{R_c} \mid = \left| \frac{\sum_{i:W_i=1} R_i}{N_t} - \frac{\sum_{i:W_i=0} R_i}{N_c} \right|$$

$\Rightarrow$ $R_i = \sum_{j=1}^{N} \mathbf{1}_{Y_j^{\text{obs}} \leq Y_i^{\text{obs}}} - \frac{N+1}{2}$ (without ties in outcomes)

$\Rightarrow$ $R_i = \sum_{j=1}^{N} \mathbf{1}_{Y_j^{\text{obs}} \leq Y_i^{\text{obs}}} + \frac{1}{2}\left(1 + \sum_{j=1}^{N} \mathbf{1}_{Y_j^{\text{obs}} = Y_i^{\text{obs}}}\right) - \frac{N+1}{2}$ (with ties)

6. The Kolmogorov-Smirnov Statistic

$$T^{\text{ks}} = \sup_y \left| \hat{F}_t(y) - \hat{F}_c(y) \right|$$

$\Rightarrow$ $\hat{F}_t(y) = 1/N_t \sum_{i:W_i=1} \mathbf{1}_{Y_i^{\text{obs}} \leq y^\star}$ and $\hat{F}_c(y) = 1/N_c \sum_{i:W_i=0} \mathbf{1}_{Y_i^{\text{obs}} \leq y^\star}$

Linear Regression with No Covariates

$$Y_i^{\text{obs}} = \alpha + \tau W_i + \epsilon_i$$

- OLS solves

$$(\widehat{\tau}^{\text{ols}}, \widehat{\alpha}^{\text{ols}}) = \text{argmin} \sum_{i=1}^{N} (Y_i^{\text{obs}} - \alpha - \tau W_i)^2$$

- Which gives

$$\widehat{\tau}^{\text{ols}} = \frac{\sum_{i=1}^{N} (W_i - \overline{W})(Y_i^{\text{obs}} - \overline{Y}^{\text{obs}})}{\sum_{i=1}^{N} (W_i - \overline{W})^2}$$
$$\widehat{\alpha}^{\text{ols}} = \overline{Y}^{\text{obs}} - \widehat{\tau}^{\text{ols}} \overline{W}$$

- Hence

$$\widehat{\tau}^{\text{ols}} = \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$$

Asymptotic Inference

- Assuming $\tau = Y_i(1) - Y_i(0)\ \forall i$, the estimated variance of the OLS residuals is

$$\widehat{\sigma}^2_{Y|W} = \frac{1}{N-2}\sum_{i=1}^{N}\widehat{\epsilon}_i^2 = \frac{1}{N-2}\sum_{i=1}^{N}(Y_i^{\text{obs}} - \widehat{Y}_i^{\text{obs}})^2$$

- The estimator for the variance of τ^{ols} is

$$\widehat{V}_{\text{homosk}} = \frac{\widehat{\sigma}^2_{Y|W}}{\sum_{i=1}^{N}(W_i - \overline{W})^2} = \widehat{\sigma}^2_{Y|W}\left\{\frac{1}{N_t} + \frac{1}{N_c}\right\}$$

- Where we have used the fact that $\sigma^2_{Y|W} = \sigma^2_t = \sigma^2_c$ (homoskedasticity)

Asymptotic Inference

- Estimator for the sampling variance of $\widehat{\tau}^{\text{OLS}}$ that allows for heteroskedasticity

$$\widehat{V}_{\text{robust}} = \frac{\sum_{i=1}^{N} \widehat{\epsilon}_i^2 \cdot \left(W_i - \overline{W}\right)^2}{\left(\sum_{i=1}^{N} (W_i - \overline{W})^2\right)^2} = \frac{\widehat{\sigma}_t^2}{N_t} + \frac{\widehat{\sigma}_c^2}{N_c}$$

- where

$$\widehat{\sigma}_t^2 = \frac{1}{N_t - 1} \sum_{i:W_i=1} (Y_i^{\text{obs}} - \widehat{Y}_t^{\text{obs}})^2$$

$$\widehat{\sigma}_c^2 = \frac{1}{N_c - 1} \sum_{i:W_i=0} (Y_i^{\text{obs}} - \widehat{Y}_c^{\text{obs}})^2$$

Linear Regression with Covariates

$$Y_i^{\text{obs}} = \alpha + \tau W_i + \mathbf{X_i}\beta + \epsilon_i$$

- Same as above after conditioning out $\mathbf{X_i}$

$$\widehat{V}_{\mathbf{X}}^{\text{homosk}} = \frac{\hat{\sigma}^2_{Y|W,\mathbf{X}}}{\sum_{i=1}^{N}(W_i - \bar{W})^2} = \hat{\sigma}^2_{Y|W,\mathbf{X}} \left\{ \frac{1}{N_t} + \frac{1}{N_c} \right\}$$

$$\widehat{V}_{\mathbf{X}}^{\text{robust}} = \frac{\sum_{i=1}^{N} \widehat{\epsilon}^2_{\mathbf{X},i} \cdot \left(W_i - \overline{W}\right)^2}{\left(\sum_{i=1}^{N} (W_i - \overline{W})^2 \right)^2}$$

Testing for Treatment Effects

$$Y_i^{\text{obs}} = \alpha + \tau W_i + \mathbf{X_i}\beta + W_i(\mathbf{X_i} - \overline{\mathbf{X}})\gamma + \epsilon_i$$

1. Zero average treatment effect (the "average null")

$$H_0 \ : \ \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x] = 0, \forall x$$

   - $Q_{\text{zero}} = \begin{pmatrix} \hat{\tau}^{\text{ols}} \\ \hat{\gamma}^{\text{ols}} \end{pmatrix}' \widehat{V}_{\tau,\gamma}^{-1} \begin{pmatrix} \hat{\tau}^{\text{ols}} \\ \hat{\gamma}^{\text{ols}} \end{pmatrix}$

2. Constant average treatment effect

$$H_0 \ : \ \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x] = \tau, \forall x$$

   - $Q_{\text{const}} = \left(\hat{\gamma}^{\text{ols}}\right)' \widehat{V}_{\gamma}^{-1} \hat{\gamma}^{\text{ols}}$

Testing for Treatment Effects: An Example

Table 7.3. ***Regression Estimates for Average Treatment Effects on Post-Cholesterol Levels for the PRC-CPPT Cholesterol Data from Table 7.1***

Covariates	Model for Levels		Model for Logs	
	Est	$\widehat{(\text{s. e.})}$	Est	$\widehat{(\text{s. e.})}$
Assignment	−25.04	(2.56)	−0.098	(0.010)
Intercept	−3.28	(12.05)	−0.133	(0.233)
`chol1`	0.98	(0.04)	−0.133	(0.233)
`chol2-chol1`	0.61	(0.08)	0.602	(0.073)
`chol1` × Assignment	−0.22	(0.09)	−0.154	(0.107)
(`chol2-chol1`) × Assignment	0.07	(0.14)	0.184	(0.159)
R-squared	0.63		0.57	

Table 7.4. ***P-Values for Tests for Constant and Zero Treatment Effects, Using*** `chol1` ***and*** `chol2-chol1` ***as Covariates for the PRC-CPPT Cholesterol Data from Table 7.1***

		Post-Cholesterol Level	Compliance
Zero treatment effect	$\mathcal{X}^2(3)$ approximation	<0.001	<0.001
	Fisher exact p-value	<0.001	0.001
Constant treatment effect	$\mathcal{X}^2(2)$ approximation	0.029	0.270

Stratified Randomized Experiments

What's the Point of Stratification?

- Units grouped according to some pre-treatment characteristics into strata
- Rules out substantial imbalances in covariates that could arise by chance
- Within each stratum, a completely randomized experiment is conducted

The Benefits of Stratification

- Consider a DGP with $i = 1, ..., N$ and one covariate $G_i \in \{f, m\}$

  $\Rightarrow p(G_i = f) = p = \frac{N(f)}{N}$

  $\Rightarrow p(G_i = m) = 1 - p = \frac{N(m)}{N}$

- Completely randomized design

  $\Rightarrow N_t = qN$ and $N_c = (1 - q)N$

- ATE and its sampling variance

$$\hat{\tau}^{\text{dif}} = \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}}$$

$$\mathbb{V}(\hat{\tau}^{\text{dif}}) = \frac{\sigma_t^2}{N_t} + \frac{\sigma_c^2}{N_c}$$

The Benefits of Stratification

- Stratified design

  $\Rightarrow$ $N_t(f) = pqN$ and $N_c(f) = p(1-q)N$

  $\Rightarrow$ $N_t(m) = (1-p)qN$ and $N_c(m) = (1-p)(1-q)N$

- CATE (G_i)

$$\hat{\tau}^{\text{dif}}(f) = \overline{Y}_t^{\text{obs}}(f) - \overline{Y}_c^{\text{obs}}(f)$$
$$\hat{\tau}^{\text{dif}}(m) = \overline{Y}_t^{\text{obs}}(m) - \overline{Y}_c^{\text{obs}}(m)$$

- ATE and its sampling variance

$$\hat{\tau}^{\text{strat}} = p\hat{\tau}^{\text{dif}}(f) + (1-p)\hat{\tau}^{\text{dif}}(m)$$
$$\mathbb{V}(\hat{\tau}^{\text{strat}}) = \frac{p}{N}\left(\frac{\sigma_t^2(f)}{p} + \frac{\sigma_c^2(f)}{1-p}\right) + \frac{1-p}{N}\left(\frac{\sigma_t^2(m)}{p} + \frac{\sigma_c^2(m)}{1-p}\right)$$

The Benefits of Stratification

- The difference in the two variances is

$$\mathbb{V}(\hat{\tau}^{\text{dif}}) - \mathbb{V}(\hat{\tau}^{\text{strat}}) = \frac{p(1-p)}{N}((\mu_c(f) - \mu_c(m))^2 + (\mu_t(f) - \mu_t(m))^2) \geq 0$$

$\Rightarrow$ The extra-variance in dif vs. strat comes from the unbalancedness of G_i

$\Rightarrow$ In principle, $\hat{V}^{\text{dif}} < \hat{V}^{\text{strat}}$ (within-stratum potential outcome variances)

An Alternative to Stratification: Re-randomization

- What if after the random draw some (important) covariates are unbalanced?
- Randomize many times and select the draw that achieves better balance
  - $\Rightarrow$ E.g. pick the draw with the minimum maximum t-stat across covariates
- Preferred over stratification if need balance among several variables
- Inference is tricky: not every $\mathbf{W}$ is ex-post equally probable !
  - $\Rightarrow$ p-values need to be adjusted for the re-randomization
  - $\Rightarrow$ Ignoring the adjustment leads to conservative p-values

Re-randomization: Example

- $N = 100$ individuals, with 50 women and 50 men
- Completely randomize 60 individuals to treatment

  $\Rightarrow$ Reject and re-randomize many times until we get 30 men and 30 women
- This is a stratified experiment

  $\Rightarrow$ Need to track the entire sequence of assignment vectors that led to $\mathbf{W}$

The Structure of Stratified Randomized Experiments

- Let J be the number of strata/blocks, and $N(j), N_c(j), N_t(j)$
- Let $S_i \in \{1, ..., J\}$ be the stratum for unit i
- Let $B_i(j) = \mathbf{1}_{S_i=j}$ be the stratum indicator for unit i
- The assignment mechanism is

$$P(\mathbf{W}|\mathbf{B}, \mathbf{Y}(0), \mathbf{Y}(1)) = \prod_{j=1}^{J} \binom{N(j)}{N_t(j)}^{-1} \text{ for } \mathbf{W} \in \mathbb{W}^+$$

$\Rightarrow$ $\mathbb{W}^+ = \{\mathbf{W} \in \mathbb{W} \mid \sum_{i=1}^{N} B_i(j) \cdot W_i = N_t(j) \text{ for } j = 1, ..., J\}$

Example: Tennessee Project Star

School/ Stratum	No. of Classes	Regular Classes ($W_i = 0$)	Small Classes ($W_i = 1$)
1	4	−0.197, 0.236	0.165, 0.321
2	4	0.117, 1.190	0.918, −0.202
3	5	−0.496, 0.225	0.341, 0.561, −0.059
4	4	−1.104, −0.956	−0.024, −0.450
5	4	−0.126, 0.106	−0.258, −0.083
6	4	−0.597, −0.495	1.151, 0.707
7	4	0.685, 0.270	0.077, 0.371
8	6	−0.934, −0.633	−0.870, −0.496, −0.444, 0.392
9	4	−0.891, −0.856	−0.568, −1.189
10	4	−0.473, −0.807	−0.727, −0.580
11	4	−0.383, 0.313	−0.533, 0.458
12	5	0.474, 0.140	1.001, 0.102, 0.484
13	4	0.205, 0.296	0.855, 0.509
14	4	0.742, 0.175	0.618, 0.978
15	4	−0.434, −0.293	−0.545, 0.234
16	4	0.355, −0.130	−0.240, −0.150
Average		−0.13	0.09
(S.D.)		(0.56)	(0.61)

Randomization Inference for Stratified Experiments

- Sharp null hypothesis

$$H_0 \ : \ Y_i(1) = Y_i(0) \ \forall i = 1, 2, ..., N.$$

- Define average observed outcomes in stratum j as

$$\overline{Y}_t^{\text{obs}}(j) = \frac{1}{N_t(j)} \sum_{i:S_i=j} W_i Y_i^{\text{obs}}$$

$$\overline{Y}_c^{\text{obs}}(j) = \frac{1}{N_c(j)} \sum_{i:S_i=j} (1 - W_i) Y_i^{\text{obs}}$$

$\Rightarrow$ Strata-level propensity score is

$$e(j) = \frac{N_t(j)}{N(j)}$$

Test Statistics

- Within-stratum test statistic

$$T^{\text{dif}}(j) = \left|\overline{Y}_t^{\text{obs}}(j) - \overline{Y}_c^{\text{obs}}(j)\right|$$

$\Rightarrow$ Not very informative as we are interested in treatment effects across all strata

- Linear combination of the within-stratum statistics

$$T^{\text{dif},\lambda_{RSS}} = \left|\sum_{j=1}^{J} \frac{N_j}{N}\left(\overline{Y}_t^{\text{obs}}(j) - \overline{Y}_c^{\text{obs}}(j)\right)\right|$$

$\Rightarrow$ Need $e(j)$ to vary across j for the test to have power over T^{dif}

Randomization Inference of the Tennessee Project Star

- $B_i(j), i = 1, ..., 68$ (class-level data)

$$H_0 \ : \ Y_i(1) = Y_i(0) \ \forall i = 1, 2, ..., 68.$$

- Total number of possible assignments is $(6^{13}) \times 10^2 \times 15$
  - $\Rightarrow$ 13 Schools with two classes in each group: $\binom{4}{2} = 6$
  - $\Rightarrow$ 2 Schools with three small classes and two regular classes: $\binom{5}{2} = 10$
  - $\Rightarrow$ 1 School with four small classes and two regular classes: $\binom{6}{2} = 15$

$\Rightarrow T^{\text{dif}} = \mid \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} \mid = 0.224$, with $p = 0.034$

$\Rightarrow T^{\text{dif},\lambda_{RSS}} = \mid \sum_{j=1}^{J} \frac{N_j}{N} (\overline{Y}_t^{\text{obs}}(j) - \overline{Y}_c^{\text{obs}}(j)) \mid = 0.241$, with $p = 0.023$

Regression Analysis

$$Y_i^{\text{obs}} = \tau W_i + \sum_{j=1}^{J} \beta(j) B_i(j) + \epsilon_i$$

- Recall that $B_i(j) = \mathbf{1}_{S_i=j}$ is the stratum indicator for unit i

$\Rightarrow$ $\widehat{\tau}^{\text{ols}}$ is not a consistent estimator of ATE if $\tau(j) \neq \tau(j') \; \forall j \neq j'$

$$WATE = \frac{\sum_{j=1}^{J} \omega(j)\tau(j)}{\sum_{j=1}^{J} \omega(j)}$$

- $\omega(j) = \frac{N_j}{N} \frac{N_t(j)}{N(j)} \frac{N(j)-N_t(j)}{N(j)} = q(j)e(j)(1-e(j))$

Asymptotic Inference

- The asymptotic variance of the WATE is

$$\mathbb{V}^{\text{strat}} = \frac{\sum_{i=1}^{N} \epsilon_i^2 \cdot \left(W_i - \sum_{j=1}^{J} q(j) B_i(j)\right)^2}{\left(\sum_{j=1}^{J} \omega(j)\right)^2}$$

$\Rightarrow$ Variance weights more those $\widehat{\tau}(j)$ that are more precisely estimated

$$\mathbb{V}^{j} = \frac{\sigma^2/N}{q(j)e(j)(1-e(j))}$$

Fully-interacted Model

$$Y_i^{\text{obs}} = \tau W_i \frac{B_i(j)}{N(j)/N} + \sum_{j=1}^{J} \beta(j) B_i(j) + \sum_{j=1}^{J-1} \gamma(j) W_i \left(B_i(j) - B_i(J) \frac{N(j)}{N(J)} \right) + \epsilon_i$$

$\Rightarrow$ In this case, $\widehat{\tau}^{\text{ols}}$ is a consistent estimator of ATE

- With asymptotic variance

$$\mathbb{V}^{\text{strat,inter}} = \sum_{i=1}^{N} q(j)^2 \cdot \left(\frac{\sigma_c^2(j)}{q(j)(1-e(j))} + \frac{\sigma_t^2(j)}{q(j)e(j)} \right)$$

$\Rightarrow$ In general, $\mathbb{V}^{\text{strat,inter}} > \mathbb{V}^{\text{strat}}$

Regression Analysis of the Tennessee Project Star

- The point estimate of τ in the standard model is
  - $\widehat{\tau}^{\text{ols}} = 0.238$ $(\widehat{s.e.} = 0.103)$
- The point estimate of τ in the fully-interacted model is
  - $\widehat{\tau}^{\text{ols,inter}} = 0.241$ $(\widehat{s.e.} = 0.095)$

$\Rightarrow$ Limited heterogeneity in the treatment effects across strata

Pairwise Randomized Experiments

What is a Pairwise Experiment?

- Stratified experiments with exactly two units in each stratum
- Units are matched to other units based on their similarity in covariates, with the expectation that this similarity corresponds to similarity in the potential outcomes under each treatment

$\Rightarrow$ Each stratum has the same proportion of treated units, and so the natural estimator for the average treatment effect weights each stratum equally

Example: Children's Television Workshop Experiment

Pair	Unit A					Unit B				
	$Y_{i,A}(0)$	$Y_{i,A}(1)$	$W_{i,A}$	$Y_{i,A}^{\text{obs}}$	$X_{i,A}$	$Y_{i,B}(0)$	$Y_{i,B}(1)$	$W_{i,B}$	$Y_{i,B}^{\text{obs}}$	$X_{i,B}$
1	54.6	?	0	54.6	12.9	?	60.6	1	60.6	12.0
2	56.5	?	0	56.5	15.1	?	55.5	1	55.5	13.9
3	75.2	?	0	75.2	16.8	?	84.8	1	84.8	17.2
4	76.6	?	0	75.6	15.8	?	101.9	1	101.9	18.9
5	55.3	?	0	55.3	13.9	?	70.6	1	70.6	15.3
6	59.3	?	0	59.3	14.5	?	78.4	1	78.4	16.6
7	87.0	?	0	87.0	17.0	?	84.2	1	84.2	16.0
8	73.7	?	0	73.7	15.8	?	108.6	1	108.6	20.1

The Structure of Pairwise Randomized Experiments

- The number of units, N, is even. The number of strata is $J = N/2$
- There is one treated unit and one control unit in each stratum, $N_t(j) = N_c(j) = 1$, and $N(j) = 2$ for all $j = 1, ..., J$
- Let G_i be the variable indicating the pair, with $G_i \in \{1, ..., N/2\}$, which is a function of covariates $\mathbf{X}_i$
- Within each pair there are $\binom{N(j)}{N_t(j)} = \binom{2}{1} = 2$ possible assignments, so the assignment mechanism is

$$P(\mathbf{W}|\mathbf{G}, \mathbf{Y}(0), \mathbf{Y}(1)) = \prod_{j=1}^{N/2} \binom{N(j)}{N_t(j)}^{-1} = \prod_{j=1}^{N/2} \frac{1}{2} = 2^{-N/2}, \text{ for } \mathbf{W} \in \mathbb{W}^+$$

$\Rightarrow$ $\mathbb{W}^+ = \{\mathbf{W} \in \mathbb{W} \mid \sum_{i:G_i=j} W_i = 1, \text{ for } j = 1, ..., N/2\}$

Potential Outcomes

- For all pairs j, $W_{j,A} = 1 - W_{j,B}$ and $P(W_{j,A} \mid \mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}) = 1/2$

$\Rightarrow$ Potential outcomes are

$$Y_{j,A}^{\text{obs}} = \begin{cases} Y_{j,A}(0) \text{ if } W_{j,A} = 0 \\ Y_{j,A}(1) \text{ if } W_{j,A} = 1 \end{cases}$$

$$Y_{j,B}^{\text{obs}} = \begin{cases} Y_{j,B}(0) \text{ if } W_{j,A} = 1 \\ Y_{j,B}(1) \text{ if } W_{j,A} = 0 \end{cases}$$

Estimand

- The average treatment effect within pair j is

$$\begin{aligned}\tau^{\text{pair}}(j) &= \frac{1}{2}\sum_{i:G_i=j}(Y_i(1)-Y_i(0)) \\ &= \frac{1}{2}\left\{(Y_{j,A}(1)-Y_{j,A}(0))+(Y_{j,B}(1)-Y_{j,B}(0))\right\}\end{aligned}$$

$\Rightarrow$ The average treatment effect is

$$\begin{aligned}\tau &= \frac{1}{N}\sum_{i=1}^{N}(Y_i(1)-Y_i(0)) \\ &= \frac{2}{N}\sum_{j=1}^{N/2}\tau^{\text{pair}}(j)\end{aligned}$$

Randomization Inference

$$H_0 \ : \ Y_i(1) = Y_i(0), \ \ \forall i = 1, 2, ..., N.$$

- Usual "absolute difference" statistic across pairs is

$$\begin{aligned} T^{\text{dif}} &= \left| \frac{1}{J} \sum_{j=1}^{J} (Y_{j,t}^{\text{obs}} - Y_{j,c}^{\text{obs}}) \right| \\ &= \left| \frac{2}{N} \sum_{j=1}^{N/2} (W_{i,A}(Y_{j,A}^{\text{obs}} - Y_{j,B}^{\text{obs}})(1 - W_{i,A})(Y_{j,B}^{\text{obs}} - Y_{j,A}^{\text{obs}})) \right| \\ &= \left| \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} \right| \end{aligned}$$

$\Rightarrow$ The associated p-value is different than that calculated under a completely randomized design due to fewer elements in $\mathbb{W}^+$

Randomization Inference (cont'd)

- Alternative statistics include

$$
\begin{aligned}
T^{\text{rank}} &= \left|\overline{R}_t - \overline{R}_c\right| \\
&= \left|\frac{2}{N}\sum_{j=1}^{N/2}\left(W_{i,A}\left(R_{j,A} - R_{j,B}\right) + \left(1 - W_{i,A}\right)\left(R_{j,B} - R_{j,A}\right)\right)\right| \\
T^{\text{rank,pair}} &= \left|\frac{2}{N}\sum_{j=1}^{N/2}\left(\mathbf{1}_{Y_{j,1}^{\text{obs}} > Y_{j,0}^{\text{obs}}} - \mathbf{1}_{Y_{j,1}^{\text{obs}} < Y_{j,0}^{\text{obs}}}\right)\right|
\end{aligned}
$$

- Both statistics are robust to the presence of outliers in observed outcomes

$\Rightarrow$ When there is substantial variation in outcomes across pairs $T^{\text{rank,pair}}$ has more power than T^{rank}

Normalized Rank: Children's Television Example

Pair G_i	Treatment W_i	Pre-Test Score X_i	Post-Test Score Y_i^{obs}	Normalized Rank Post-Test Score R_i
1	0	12.9	54.6	−7.5
1	1	12.0	60.6	2.5
2	0	15.1	56.5	−4.5
2	1	12.3	55.5	5.5
3	0	16.8	75.2	0.5
3	1	17.2	84.8	4.5
4	0	15.8	75.6	1.5
4	1	18.9	101.9	7.5
5	0	13.9	55.3	−6.5
5	1	15.3	70.6	−1.5
6	0	14.5	59.3	−3.5
6	1	16.6	78.4	2.5
7	0	17.0	87.0	5.5
7	1	16.0	84.2	3.5
8	0	15.8	73.7	−0.5
8	1	20.1	108.6	7.5

Randomization Inference: Children's Television Example

$$T^{\text{dif}} = 13.4, \quad p\text{-value} = 0.031$$

$$T^{\text{rank}} = 3.8, \quad p\text{-value} = 0.031$$

$$T^{\text{rank,pair}} = 0.5, \quad p\text{-value} = 0.145$$

$\Rightarrow$ $T^{\text{rank,pair}}$ is less significant than the other statistics because for the two pairs where $Y_{j,1}^{\text{obs}} < Y_{j,0}^{\text{obs}}$ the difference in outcomes is small

Regression Methods

- Primary outcome of interest is the within-pair difference in observed outcomes of the treated an the control unit in the pair,

$$\widehat{\tau}^{\text{pair}}(j) = Y_{j,1}^{\text{obs}} - Y_{j,0}^{\text{obs}}$$

- Then consider the following (trivial) regression

$$\widehat{\tau}^{\text{pair}}(j) = \tau + \epsilon_j$$

- The standard estimator for the average treatment effect is the simple average of the within-pair differences:

$$\widehat{\tau}^{\text{ols}} = \frac{2}{N} \sum_{j=1}^{N/2} \widehat{\tau}^{\text{pair}}(j)$$

Regression Methods: Inference

- So far, pairwise design is just a special case of stratified designs
- Complications arise when estimating the variance of $\widehat{\tau}^{\text{pair}}(j)$
- Cannot estimate the within-stratum variance, which requires at least two treated and at least two control units in each stratum
- Instead, consider the variance of $\widehat{\tau}^{\text{pair}}(j)$ over the pairs:

$$\widehat{V}^{\text{pair}} = \frac{1}{N/2(N/2-1)} \sum_{j=1}^{N/2} \left(\widehat{\tau}^{\text{pair}}(j) - \widehat{\tau}\right)^2$$

$\Rightarrow$ Typically, $\widehat{V}^{\text{pair}} < \widehat{V}^{\text{strata}} < \widehat{V}^{\text{ols}}$

Regression Methods: Adding Covariates

1. Adding covariates as within-pair difference

$$\widehat{\tau}^{\text{pair}}(j) = \tau + \beta \Delta_{X,j} + \epsilon_j$$

   - Where $\Delta_{X,j} = (W_{j,A} \cdot (X_{j,A} - X_{j,B}) + (1 - W_{j,A}) \cdot (X_{j,B} - X_{j,A}))$

2. Adding covariates as within-pair average

$$\widehat{\tau}^{\text{pair}}(j) = \tau + \gamma \overline{X_j} + \epsilon_j$$

   - Where $\overline{X_j} = (X_{j,A} - X_{j,B})/2$

3. General case

$$\widehat{\tau}^{\text{pair}}(j) = \tau + \beta \Delta_{X,j} + \gamma(\overline{X}_j - \overline{X}) + \epsilon_j$$

Regression Methods: Children's Television Example

- For the regression model with only a constant

$$\widehat{\tau}^{\text{ols}} = 13.4 \; (\widehat{s.e.} = 4.3)$$

- For the regression function that includes the within-pair difference

$$\widehat{\tau}^{\text{ols}} = 9.0 \; (\widehat{s.e.} = 1.5) \text{ and } \widehat{\beta}^{\text{ols}} = 5.4 \; (\widehat{s.e.} = 0.6)$$

- For the regression function that includes the within-pair average

$$\widehat{\tau}^{\text{ols}} = 13.4 \; (\widehat{s.e.} = 3.5) \text{ and } \widehat{\gamma}^{\text{ols}} = 3.9 \; (\widehat{s.e.} = 1.7)$$

- Including both terms

$$\widehat{\tau}^{\text{ols}} = 8.5 \; (\widehat{s.e.} = 1.5), \widehat{\beta}^{\text{ols}} = 5.9 \; (\widehat{s.e.} = 0.8), \text{ and } \widehat{\gamma}^{\text{ols}} = -1.0 \; (\widehat{s.e.} = 0.7)$$

Clustered Randomized Experiments

What's the Point of Clustering?

- Instead of assigning treatments at the unit level, in this setting the population is first partitioned into a number of clusters
- Then all units in a cluster are assigned to the same treatment level
- Given a fixed sample size, this design is in general not as efficient as a completely randomized design or a stratified randomized design

$\Rightarrow$ There may be interference between units at the unit-level violating SUTVA

$\Rightarrow$ In many cases it is easier to sample units at the cluster level

The Structure of Clustered Experiments

- Let G_{ig} be a binary indicator that unit i belongs to cluster $g = 1, ..., G$
- $N_g = \sum_{i=1}^{N} G_{ig}$, so that N_g/N is the share of cluster g in the sample
- $\overline{W}_g \in \{0, 1\}$ is the treatment assignment for all units in cluster g
- The assignment mechanism is

$$P(\mathbf{W}, \mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}) = \binom{G}{G_t}^{-1}$$

$\Rightarrow \mathbb{W}^+ = \{\mathbf{W} \in \mathbb{W} \mid \sum_{g=1}^{G} \overline{W}_g = G_t\}$

Example: The *Progresa* Program

- Clustered RCT during the roll-out of the program in rural areas
  - 506 villages among those eligible to receive the program
  - 320 early treatment and 186 late treatment (control)
- Individual/HH level data for both eligible and non-eligible HHs in each village
  - Approx. 30,000 program eligible children
  - About 50-100 HHs per village

Estimands

- The choice of estimand depends on the choice of the unit of analysis

$\Rightarrow$ Unit-level: a natural estimand is the usual ATE

$$\tau = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0))$$

$\Rightarrow$ Cluster-level: weighted average of the unit-level treatment effect

$$\tau^{\text{clust}} = \frac{1}{G}\sum_{g=1}^{G}\tau_g, \quad \text{where } \tau_g = \frac{1}{N_g}\sum_{i:G_{ig}=1}^{N}(Y_i(1) - Y_i(0))$$

Unit-level Vs. Cluster-level

- Cluster-level analysis is more directly linked to the randomization framework
  - $\Rightarrow$ Differently-sized clusters, such as states or towns
  - $\Rightarrow$ Many units will be in the same treatment group so unit-level inference is tricky
- Unit-level incorporates individual-level covariates, which improve efficiency
  - $\Rightarrow$ More homogenous clusters, such as schools or classrooms

Randomization Inference

- The usual statistic for unit-level analysis

$$T^{\text{dif}} = \mid \overline{Y}_t^{\text{obs}} - \overline{Y}_c^{\text{obs}} \mid$$

- The equivalent statistic for cluster-level analysis

$$T^{\text{clust}} = \left| \frac{1}{G_t} \sum_{g:\overline{W}_g=1} \overline{Y}_g^{\text{obs}} - \frac{1}{G_c} \sum_{g:\overline{W}_g=0} \overline{Y}_g^{\text{obs}} \right|$$

Randomization Inference of *Progresa*

- Children-level analysis on school enrollment (pre-program year 1997)

$$T^{\text{dif}} = 0.0075, \quad p\text{-value} = 0.400$$

- Children-level analysis on school enrollment (program year 1998)

$$T^{\text{dif}} = 0.0388, \quad p\text{-value} < 0.001$$

- Village-level analysis on school enrollment (program year 1998)

$$T^{\text{clust}} = 0.0234, \quad p\text{-value} = 0.0120$$

Regression Methods

- In unit-level analysis, we estimate the following regression

$$Y_i^{\text{obs}} = \alpha + \tau W_i + \bar{X}'\gamma + \epsilon_i$$

$\Rightarrow$ $\widehat{\tau}^{\text{ols}}$ consistently estimates ATE (with centered covariates)

- Use Liang-Zeger clustered Var-Cov:

$$\mathbb{V}_{\text{clust}} = \frac{\sum_{g=1}^{G} \left(\sum_{i:G_{ig}=1} \epsilon_i^2 \cdot \left(W_i - \overline{W}\right)^2 \right)}{\left(\sum_{i=1}^{N} (W_i - \overline{W})^2 \right)^2}$$

Regression Analysis: Cluster-Level

- In cluster-level analysis, consider the following regression

$$\overline{Y}_g^{\text{obs}} = \alpha + \tau \overline{W}_g + \eta_g$$

$\Rightarrow$ $\widehat{\tau}^{\text{ols}}$ consistently estimates ATE

- The sampling variance of τ^{ols} is the usual one

$$\mathbb{V} = \frac{\sum_{g=1}^{G} \eta_g^2}{\sum_{g=1}^{G} \left(\overline{W}_g - \overline{W}\right)^2} = \sigma^2 \left\{ \frac{1}{G_t} + \frac{1}{G_c} \right\}$$

Regression Analysis of *Progresa*

- Children-level analysis on school enrollment (pre-program year 1997)

$$\widehat{\tau}^{\text{ols}} = 0.0075 \ (\widehat{s.e.} = 0.0091)$$

- Children-level analysis on school enrollment (program year 1998)

$$\widehat{\tau}^{\text{ols}} = 0.0388 \ (\widehat{s.e.} = 0.0104)$$

- Village-level analysis on school enrollment (program year 1998)

$$\widehat{\tau}^{\text{ols}} = 0.0234 \ (\widehat{s.e.} = 0.0092)$$

Adaptive Randomized Experiments

What is an Adaptive Randomized Experiment?

- A standard RCT applies the same procedures for allocating treatments throughout the trial

$\Rightarrow$ An adaptive design may, based on interim analysis of the trial's result, change the allocation of subjects to treatment arms

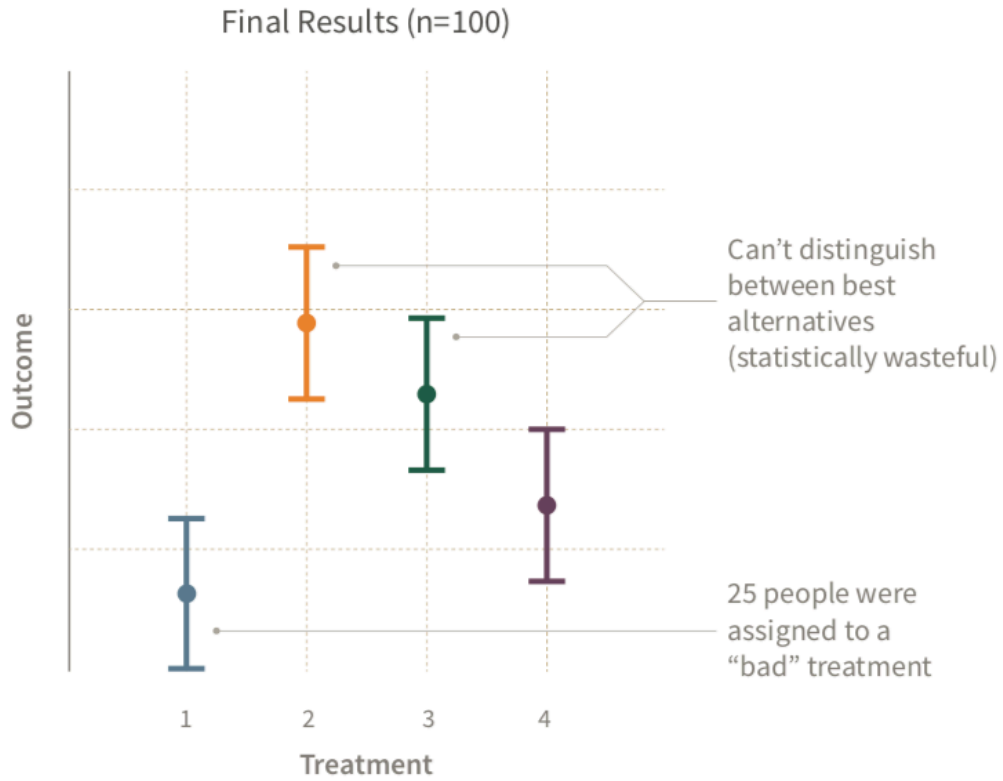
- Adaptive designs require multiple periods of treatment and outcome assessment
  - $\Rightarrow$ Well suited to survey, on-line, and lab experiments, where participants are treated and outcomes measured in batches over time
  - $\Rightarrow$ Some field experiments are conducted in stages (e.g. treatment is to be deployed over time in a series of different regions)

Illustrative Example

- The goal is to select an optimal website design
  - The treatments are different color schemes
  - The outcome is some measure of visitors' engagement with the website
- In a non-adaptive experiment, we assign each visitor to a particular color scheme and then measure how much she engages with the website
- In an adaptive experiment, we begin with an initial treatment assignment on a small wave of data
  - $\Rightarrow$ Intermediate results give us some idea of the performance of each arm
  - $\Rightarrow$ This informs how the next wave of data should be allocated across arms

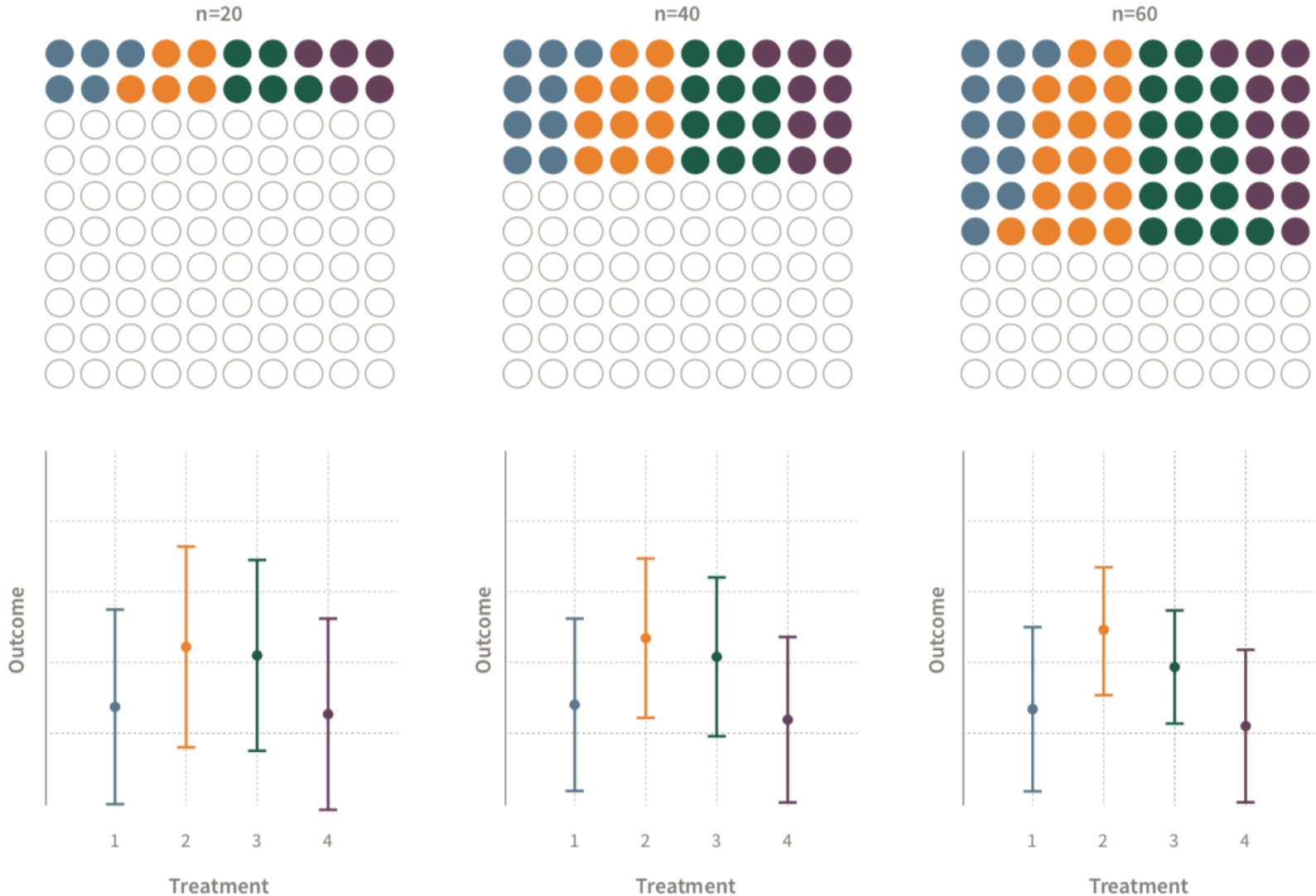
Illustrative Example (cont'd)

- The fraction of observations (number of users) assigned to each treatment is set before the experiment starts



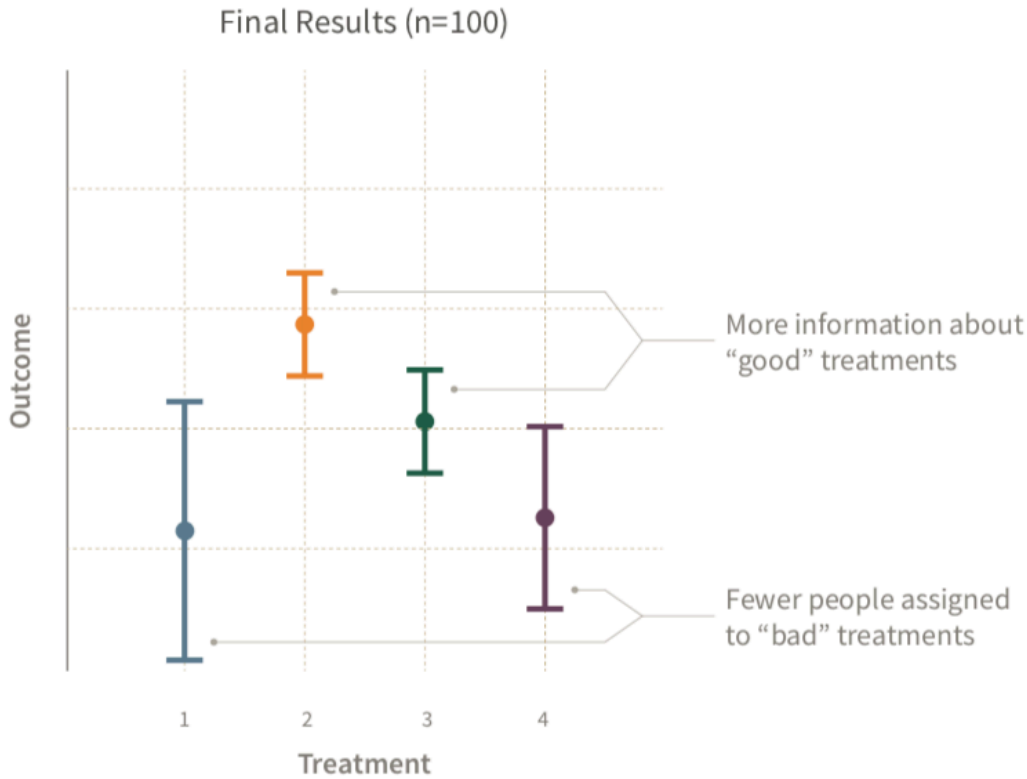
Illustrative Example (cont'd)

⇒ Each wave re-assigns treatment shares based on intermediate results



Illustrative Example (cont'd)

⇒ We assign the arms that seem more promising more often, according to the objective we set out



Setup

- Waves $t = 1, ..., T$, sample sizes N_t
- Treatment $D \in \{1, ..., k\}$, outcomes $Y \in [0, 1]$, covariates X
- Potential outcomes Y^d, and $\theta^{dx} = E[Y_{it}^d | X_{it} = x]$
- Repeated cross-sections: $(Y_{it}^1, ..., Y_{it}^k; X_{it})$ are i.i.d. across both i and t
- Given all information available at time t form posterior beliefs P_t over θ

$\Rightarrow$ Based on beliefs and the objective, decide what share p_t^{dx} of stratum x will be assigned to treatment d in time t

Thompson Sampling

- In each period subjects are assigned to treatment arms in proportion to the posterior probability that a given arm is best

$$p_t^{dx} = P_t\left(d = \underset{d'}{\text{argmax}}\ \theta^{d'x}\right)$$

- Suppose you care about both participant welfare, and precise point estimates/high power for all treatments

$$\tilde{p}_t^{dx} = (1-\gamma)p_t^{dx} + \gamma/k$$

$\Rightarrow$ The designer max participant welfare while learning something about the effectiveness of suboptimal treatments

Exploration Sampling

- Kasy and Sautmann (ECMA, 2021) propose a modification of Thompson sampling probabilities to make them less aggressive

$\Rightarrow$ Increase the expected value of the arm selected at the end of the experiment

- Assigns shares q_t^d of each wave to treatment d, where

$$q_t^d = S_t \cdot p_t^d \cdot (1 - p_t^d)$$

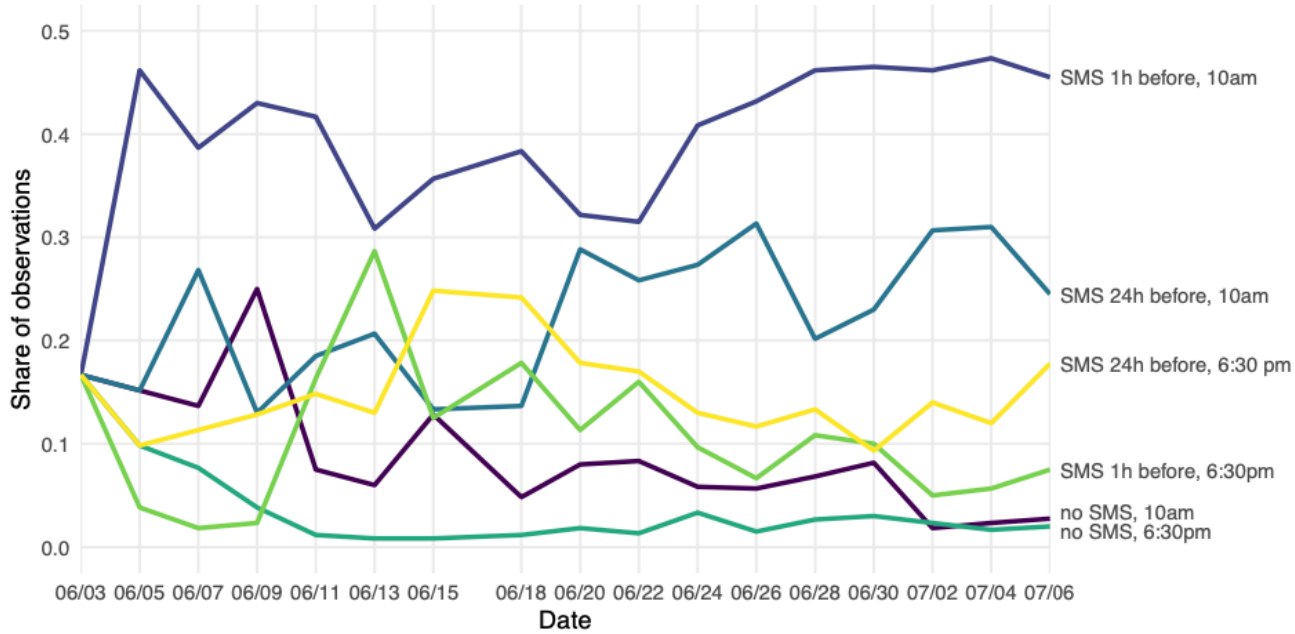
$$p_t^d = P_t\left(d = \underset{d'}{\text{argmax}}\ \theta^{d'}\right)$$

$$S_t = \frac{1}{\sum_d p_t^d (1 - p_t^d)}$$

$\Rightarrow$ Shifts weight away from best performing treatment to its close competitors

Exploration Sampling in Practice

- Kasy and Sautmann (2021) design an experiment using exploration sampling on agricultural extension services for farmers in India
- Six treatments to incentivize phone-call completion
- Outcome is call completion (1=answer five questions asked during the call)
- Daily waves of 600 phone calls randomly selected out of 10,000 valid numbers



Inference

- Inference has to take into account adaptivity. Example:
- Flip a fair coin
- If head, flip again, else stop
- Probability distribution: 50% tail-stop, 25% head-tail, 25% head-head
- Expected share of heads?

$$0.5 \times 0 + 0.25 \times 0.5 + 0.25 \times 1 = 0.375 \neq 0.5$$

$\Rightarrow$ Sample averages by treatment arms are, in general, biased

Randomization inference for Adaptive Designs

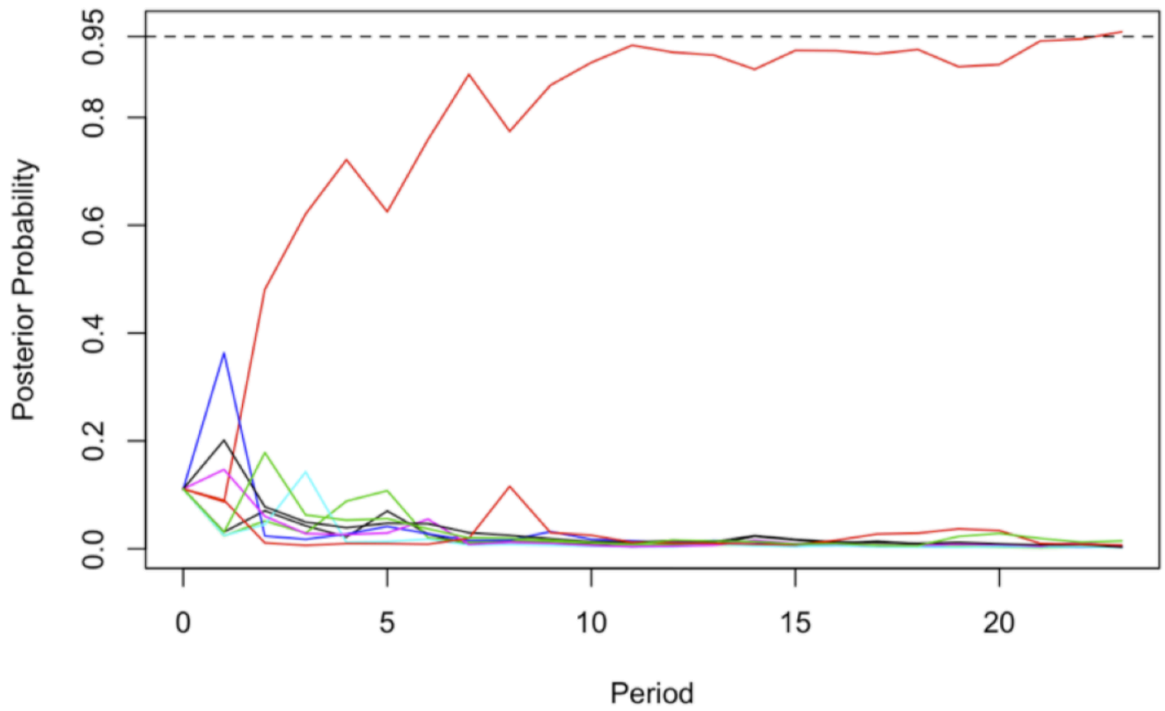
- Sharp null hypothesis: $Y_i^1 = ... = Y_i^k$
- Under this null, it is easy to re-simulate the treatment assignment: just let your assignment algorithm run with the data, switching out the treatments
- Do this many times, re-calculate the test statistic each time
- Take the $1 - \alpha$ quantile across simulations as critical value

$\Rightarrow$ This delivers finite-sample exact inference for any adaptive assignment scheme

A Simple Illustration

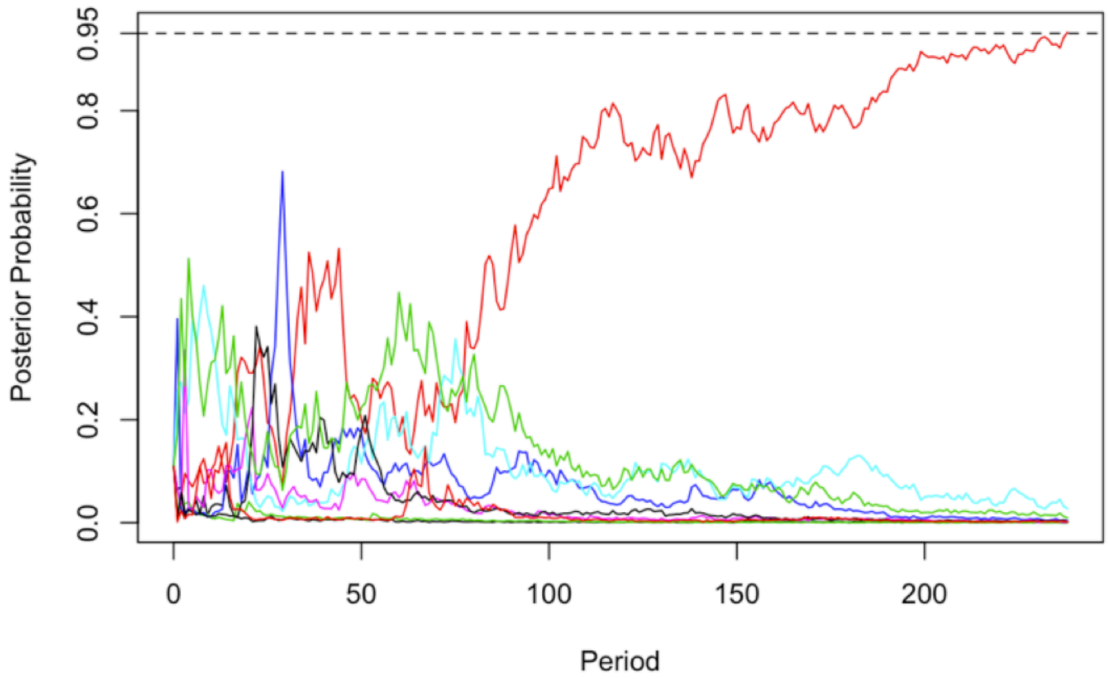
- We simulate an RCT involving a control group and eight treatment arms
- We administer treatments and outcomes for 100 subjects during each period
- The simulation assumes that each subject's outcome is binary (e.g., good versus bad)
- We allocate next period's subjects according to posterior probabilities that a given treatment arm is best
- The stopping rule is that the RCT is halted when one arm is found to have a 95% posterior probability of being best

A Simple Illustration (cont'd)



- The simulation assumes that the probability of success is 0.10 for all arms except one, which is 0.20
- The best arm (the red line) is correctly identified, and the trial is halted after 23 periods (total N=2300)

A Simple Illustration (cont'd)



- All but one of the arms have a 0.10 probability of success, and the superior arm has a 0.12 probability of success
- The design eventually settles on the truly superior arm but only after more than 200 periods (N=23,810)

Wrapping Up on Adaptive Design

- Funders and implementation partners may welcome the idea of an experimental design that responds to on-the-ground conditions such that problematic arms are scaled back

$\Rightarrow$ Even when one arm is clearly superior (inferior), the lead-time necessary to staff or outfit this arm may make it difficult to scale it up (down)

- Adaptive designs add to the complexity of the research design, the implementation and field work, and the ex-post analysis