

# Causal Inference Methods and Case Studies

STAT24630

Jingshu Wang

# Lecture 3

Topic: Classical randomized experiments

- Five examples of randomized experiment mechanisms
- Fisher's exact p-value

# Treatment assignment mechanism

- Assignment vector for binary treatment with N units:  $\mathbf{W} = (W_1, \dots, W_N) \in \{0,1\}^N$
- **Unconfoundedness property:**  $P(\mathbf{W}|\mathbf{X}, Y(0), Y(1)) = P(\mathbf{W}|\mathbf{X})$ 
  - Assignment mechanism does not depend unobserved  $\mathbf{U}$  pretreatment confounders
  - $\mathbf{U}$  includes potential outcomes  $(Y_i(0), Y_i(1))$
  - We can alternatively understand it as

$$W_i \perp (Y_i(0), Y_i(1)) \mid \mathbf{X}_i$$

- Make the treatment and control groups “identical”

$$P(Y_i(0), Y_i(1) \mid \mathbf{X}_i, W_i = 0) = P(Y_i(0), Y_i(1) \mid \mathbf{X}_i, W_i = 1)$$

- Identify conditional average treatment effect under unconfoundedness

Causal estimand that  
involve the unobserved  
potential outcomes

$$\tau(\mathbf{x}) = \mathbb{E}(Y_i(1) - Y_i(0) \mid \mathbf{X}_i = \mathbf{x})$$

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

Conditional expectations  
that we can evaluate  
based on observed data

# Common designs of randomized experiments

- Five examples of randomized experiment mechanisms
  - Bernoulli trial
  - Completely randomized experiment
  - Stratified randomized experiment
  - Paired randomized experiment
  - Rerandomization
- Key differences: the set of assignment vectors  $\mathbf{W}$  with positive probability
- The purpose of restricting the assignment mechanism is to eliminate assignment vectors that are less desirable for estimating causal effects
  - Examples: all males get treatment; all females get control

# Bernoulli trial

- Simplest Bernoulli experiment tosses a (fair) coin for each unit
  - If the coin is heads, then unit receive treatment
  - Otherwise, the unit receive control
- For each  $\mathbf{w} \in \{0,1\}^N$ ,  $P(\mathbf{W} = \mathbf{w}|\mathbf{X}) = P(\mathbf{W} = \mathbf{w}) = 0.5^N$
- $W_1, \dots, W_N \sim \text{Bernoulli}(0.5)$  and are independent
- More generally, we can toss a specialized coin for each unit depending on its covariates
  - Define propensity score  $e(\mathbf{X}_i) = P(W_i = 1 | \mathbf{X}_i)$
  - Assignment property:  $P(\mathbf{W} = \mathbf{w}|\mathbf{X}) = \prod_{i=1}^N [e(\mathbf{X}_i)^{W_i} (1 - e(\mathbf{X}_i))^{1-W_i}]$
  - $W_1, \dots, W_N$  are still independent and each  $W_i \sim \text{Bernoulli}(e(\mathbf{X}_i))$
  - Example: when trying to induce people with serious disease to enroll for the trial of a promising drug, we give them a higher probability to receive the treatment
- Drawback of the design: always a positive probability that all units receive the same treatment

# Completely randomized experiment

- A fixed number of subjects  $N_t$  is assigned to receive the active treatment
- Assignment probability

$$P(\mathbf{W} = \mathbf{w}|\mathbf{X}) = \begin{cases} \binom{N}{N_t}^{-1} & \text{if } \sum_{i=1}^N w_i = N_t \\ 0 & \text{otherwise} \end{cases}$$

- Completely randomized experiment guarantees that there are exactly  $N_t$  individuals receiving the treatment and  $N - N_t$  individuals receiving the control
- $W_1, \dots, W_N$  are slightly negatively associated
- There is still positive probability that all females receive the control and all females receive treatment  $\rightarrow$  extreme covariate imbalance after randomization
- In that case, average differences between groups could be due to sex differences
- For this single experiment, we can get a terrible estimate and wrong judgement

# Stratified randomized experiment

- Basic procedure:
  1. Blocking (Stratification): create groups of similar units based on pre-treatment covariates, let  $B_i \in \{1, \dots, J\}$  be the block indicator
  2. Block (Stratified) randomization: completely randomize treatment assignment within each group
- Blocking can improve the efficiency by minimizing the variance of the potential outcomes within each strata

*“Block what you can and randomize what you cannot”*

Box, et al. (2005). Statistics for Experimenters. 2nd eds. Wiley

- Assignment probability

$$P(\mathbf{W} = \mathbf{w}|\mathbf{X}) = \begin{cases} \prod_{j=1}^J \binom{N(j)}{N_t(j)}^{-1} & \text{if } \sum_{i:B_i=j} w_i = N_t(j) \text{ for } j = 1, \dots, J \\ 0 & \text{otherwise} \end{cases}$$

# Examples

- Randomized trial for the Moderna vaccine

[Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *New England journal of medicine*, 2020.]

- Participants were randomly assigned in a 1:1 ratio, through the use of a centralized interactive response technology system, to receive vaccine or placebo.
- Assignment was stratified into the following risk groups: persons 65 years of age or older, persons younger than 65 years of age who were at heightened risk (at risk) for severe Covid-19, and persons younger than 65 years of age without heightened risk (not at risk).

- Experiment of women policy makers in India

[Women as policy makers: Evidence from a randomized policy experiment in India. *Econometrica*, 2004]

- Each Gram Panchayat (GP) encompasses 10,000 people in several villages (between 5 and 15)
- Starting 1993, in a third of the villages in each GP, only women could be candidates for the position of councilor for the area.
- Random selection: villages are ranked in consecutive order according to an administrative number, every third village is reserved for a woman



# Paired randomized experiment

- Can we keep blocking until we cannot block any further?
- Procedure:
  1. Create  $J = N/2$  pairs of similar units
  2. Randomize treatment assignment within each pair
- Example: evaluation of health insurance policy

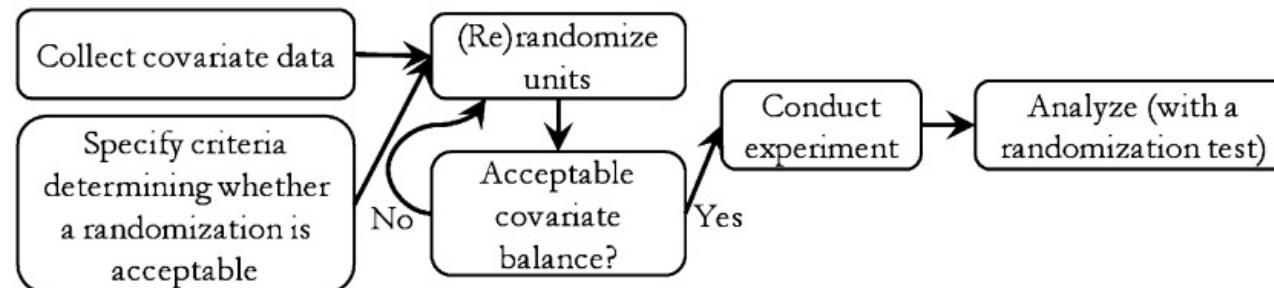
[Public policy for the poor? A randomised assessment of the Mexican universal health insurance programme. *The lancet*, 2009.]

  - Seguro Popular, a programme aimed to deliver health insurance, regular and preventive medical care, medicines, and health facilities to 50 million uninsured Mexicans
  - Units: health clusters = predefined health facility catchment areas
  - Randomization within 74 matched pairs of “similar” health clusters
  - Outcome: proportion of households within each health cluster who experienced catastrophic medical expenditure

# Rerandomization

[Morgan and Rubin. 2012. Ann. Stat., Li et al. 2018. PNAS]

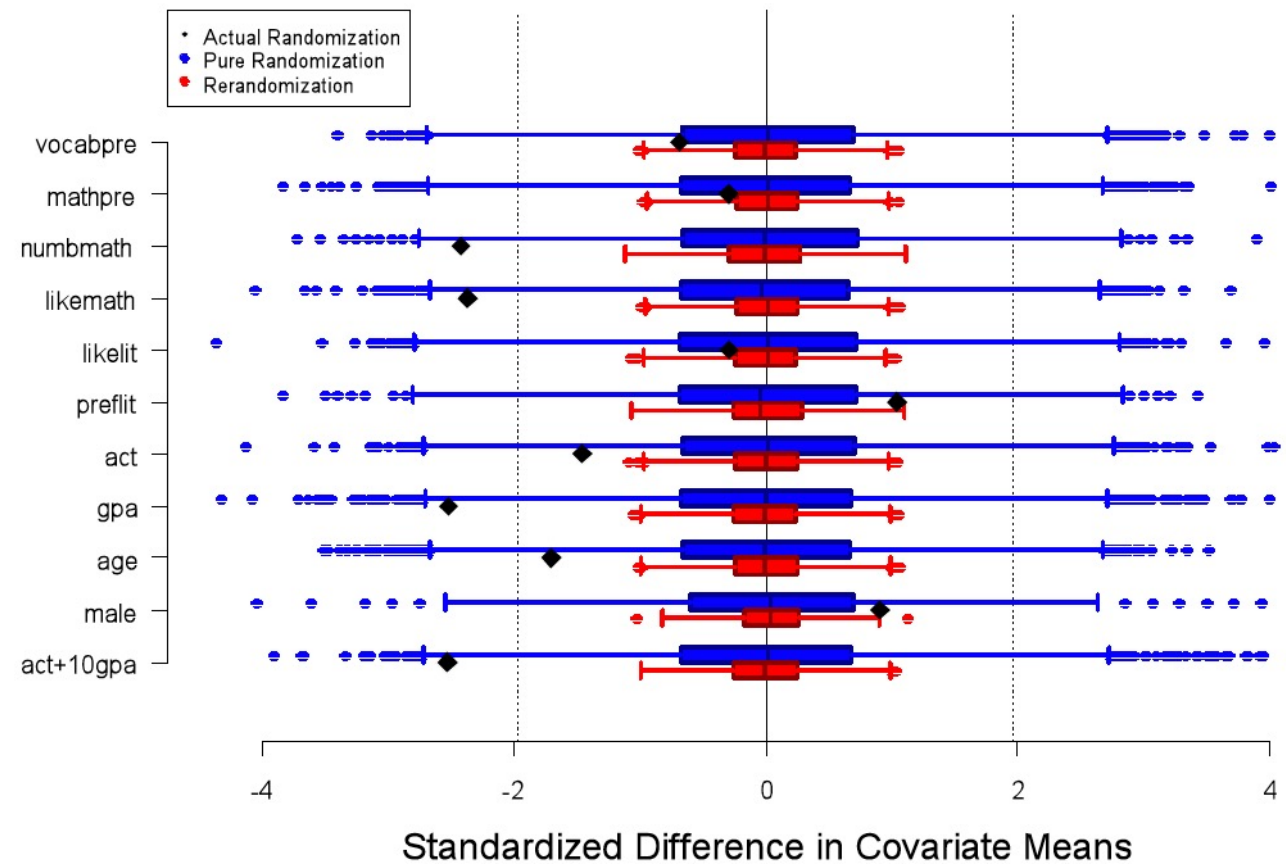
- The more covariates, the more likely at least one covariate will be imbalanced across treatment groups
- Randomization only eliminate confounding factors and yield unbiased on average (over repeated run of experiments)
- For any particular experiment, covariate imbalance is possible
- Procedure:
  1. Specify the acceptable level of covariate balance
  2. Randomize the treatment and check covariate balance
  3. Repeat until the covariate balance criterion is met



# Rerandomization: an example

[Rerandomization to balance tiers of covariates. *Journal of the American Statistical Association*, 2015.]

- The study aim to examine whether observational studies can be analyzed to yield valid estimates of causal effects.
- Undergraduate psychology students at a particular college were randomized to be in one of two arms: a randomized experiment ( $n_r = 235$ ) or an observational study ( $n_o = 210$ ).
- In the randomized experiment were randomized to take either a vocabulary or mathematics course



# Randomization Inference vs. Model-based Inference

- Randomization as the “**reason basis for inference**” (Fisher)
- Randomness comes from the physical act of randomization, which then can be used to make statistical inference
- Also called **design-based inference**
- Advantage: design justifies analysis
- model-based inference: assume a distribution for potential outcomes (at least the i.i.d. assumptions)
- Advantage of model-based inference: flexibility
- Two types of classical randomization inference
  - Fisher’s exact p-values
  - Neyman’s repeated sampling approach

# Fisher's original experiment: Lady tasting tea

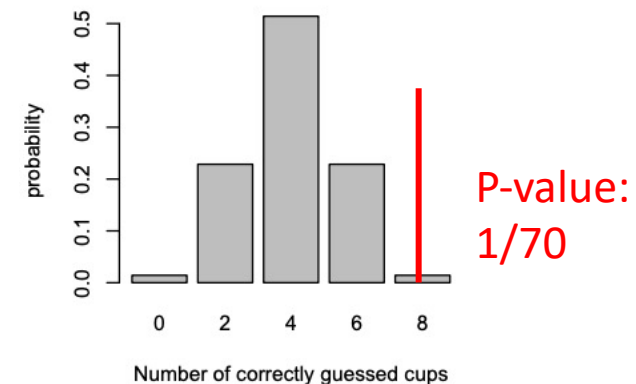
[Fisher, 1935]

- The lady in question (Muriel Bristol) claimed to be able to tell whether the tea or the milk was added first to a cup.
- Fisher proposed to give her **eight cups, four of each variety**, in random order.
- **Null hypothesis:** the lady cannot tell the difference
- **Sharp null:**  $H_0: Y_i(0) \equiv Y_i(1)$  for all  $i = 1, \dots, 8$
- **Test statistics:** the number of correctly classified cups
- The lady classified all 8 cups correctly! Did this happen by chance?

cups	guess	actual	scenarios		...
1	M	M	T	T	
2	T	T	T	T	
3	T	T	T	T	
4	M	M	T	M	
5	M	M	M	M	
6	T	T	M	M	
7	T	T	M	T	
8	M	M	M	M	
correctly guessed		8	4	6	

- Completely randomized experiment:  $\binom{8}{4} = 70$  possible scenarios with equal probability
- Under the sharp null, the lady will always have the same guesses under all scenarios

Null distribution of the test statistics



# Cough frequency example with 6 units

**Table 5.3.** *Cough Frequency for the First Six Units from the Honey Study*

Unit	Potential Outcomes				
	Cough Frequency (cfa)		Observed Variables		
	$Y_i(0)$	$Y_i(1)$	$W_i$	$X_i$ (cfp)	$Y_i^{\text{obs}}$ (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

Imputation under the sharp null

	Cough Frequency (cfa)		$Y_i^{\text{obs}}$	$\text{rank}(Y_i^{\text{obs}})$
	$Y_i(0)$	$Y_i(1)$		
1	(3)	3	3	4
2	(5)	5	5	6
3	(0)	0	0	1.5
4	4	(4)	4	5
5	0	(0)	0	1.5
6	1	(1)	1	3

- **Sharp null:**  $H_0: Y_i(0) \equiv Y_i(1)$  for all  $i = 1, \dots, 6$   
absolutely no causal effect of the treatment
- Test statistics:  $|\bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}}|$  or  $|\overline{\text{rank}}(Y_t^{\text{obs}}) - \overline{\text{rank}}(Y_c^{\text{obs}})|$

# Cough frequency example with 6 units

- If following completely randomized experiment:  $\binom{6}{3} = 20$  assignments with equal probability

Table 5.5. Randomization Distribution for Two Statistics for the Honey Data from Table 5.3

$W_1$	$W_2$	$W_3$	$W_4$	$W_5$	$W_6$	Statistic: Absolute Value of Difference in Average	
						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	<b>1.00</b>	<b>0.67</b>

- P-value based on test statistics  $|\bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}}|: \frac{16}{20} = 0.8$
- P-value based on test statistics  $|\overline{\text{rank}}(Y_t^{\text{obs}}) - \overline{\text{rank}}(Y_c^{\text{obs}})|: \frac{16}{20} = 0.8$
- The most extreme p-value we can get:  $2/20 = 0.1$
- $N = 6$  is too small to obtain statistically significant rejections

# Fisher's exact p-value

- **Features**
  - **Justified by randomization alone**: No assumptions about models or asymptotic normality
  - The sharp null may be of little interest
  - P-value is exact for small  $N$
  - Same idea as a permutation test
- **Computation of p-value**
  - Exact computation is difficult when  $n$  is large
  - Monte Carlo approximation
    1. Fill in missing potential outcomes under the sharp null
    2. Sample  $W_i$  according to complete randomization
    3. Compute the test statistic to form a reference distribution
  - Approximation can be arbitrarily accurate by increasing number of draws
- Analytical approximations when  $n$  is large (omitted)



# Fisher's exact p-value and CI

- Choice of the null hypothesis
  - Sharp null of no treatment effect:  $H_0: Y_i(0) \equiv Y_i(1)$  for all  $i = 1, \dots, N$
  - Fisher's approach cannot accommodate a null hypothesis of **zero average effect**
  - Allow more general null hypothesis  $H_0: Y_i(0) = Y_i(1) + C_i$  for all  $i = 1, \dots, N$  with pre-defined  $(C_1, \dots, C_N)$
- Invert Fisher's exact p-values for confidence intervals of  $\tau_0$ 
  - Assume the constant additive effect model  $Y_i(0) - Y_i(1) \equiv \tau_0$
  - Collect all null values  $\tau_0$  that cannot be rejected by  $\alpha$ -level Fisher's exact test
  - Idea: if we cannot reject a null hypothesis with a particular effect size, then the confidence interval should include it

# Cough frequency example with $N = 72$

P-value computation with Monte Carlo approximation

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.

- 95% CI based on statistics  $|\bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}}|$ :  $[-1.44, 0.06]$
- 95% CI based on statistics  $|\overline{\text{rank}}(Y_t^{\text{obs}}) - \overline{\text{rank}}(Y_c^{\text{obs}})|$ :  $[-2, 0]$

Hypothesized Treatment Effect	P-Value (level)	P-Value (rank)
−3.00	0.000	0.000
−2.75	0.000	0.000
−2.50	0.000	0.000
−2.25	0.000	0.000
−2.00	0.001	0.000
−1.75	0.006	0.078
−1.50	0.037	0.078
−1.44	0.050	0.078
−1.25	0.146	0.078
−1.00	0.459	0.628
−0.75	0.897	0.428
−0.50	0.604	0.428
−0.25	0.237	0.429
0.00	0.067	0.043
0.06	0.050	0.043
0.25	0.014	0.001
0.50	0.003	0.000
0.75	0.000	0.001
1.00	0.000	0.000

# Fisher's exact test for binary outcome

	Treated ( $W_i = 1$ )	Control ( $W_i = 0$ )	Total
Success ( $Y_i = 1$ )	$\sum_{i=1}^n W_i Y_i(1)$	$\sum_{i=1}^n (1 - W_i) Y_i(0)$	$m$
Failure ( $Y_i = 0$ )	$\sum_{i=1}^n W_i (1 - Y_i(1))$	$\sum_{i=1}^n (1 - W_i) (1 - Y_i(0))$	$N - m$
Total	$N_1$	$N_1$	$N$

- In the tea tasting example, the lady knows that there are 4 cups for each variety, so  $m$  is also fixed
- Test statistics:  $S = \sum_{i=1}^n W_i Y_i(1)$
- Then under complete randomization and the sharp null,  $S$  follows a hyper-geometric distribution

$$P(S = s) = \frac{\binom{m}{s} \binom{N - m}{N_1 - s}}{\binom{N}{N_1}}$$

- If  $m$  is not naturally fixed, we treat our inference as conditional on  $m$

# The project STAR example

(Mosteller. 1997. Bull. Am. Acad. Arts Sci.)

- The student-Teacher Achievement Ratio Project (1985-1989)
  - More than 10,000 students involved with the cost of \$12 million
  - Effects of class size in early grade levels
  - 3 arms: Small class, Regular-sized class, Regular class with aid
- Long-term impact of class size

	Small class	Regular-sized class
Graduate	754	892
Not graduate	148	189
Total	902	1081

- Exact p-value: 0.28 (one-sided), 0.55 (two-sided)
- Asymptotic p-value: 0.26 (one-sided), 0.53 (two-sided) [using fisher.test function in R]