

STAT6061/STAT5008 – Causal Inference

Part 2-2. Completely Randomized Experiment

An-Shun Tai

*¹Department of Statistics
National Cheng Kung University*

*²Institute of Statistics and Data Science
National Tsing Hua University*

Key questions for inference in randomized experiments

- How can statistical methods be applied to assess causal effects in randomized experiments?
- What is the causal hypothesis underlying inference in randomized experiments?
- Should prognostic covariates be adjusted in the analysis of randomized experiments?

Inference for completely randomized experiments

“Experiments should be analyzed as experiments, not as observational studies”

— Freedman (2006, p. 691)

➤ Methods of causal inference in randomized experiments

	Randomization-Based Inference (Design-Based Inference)	Sampling-Based Inference
Features	✓ Assumes potential outcomes are fixed for each subject. ✓ Treats treatment assignment as random. ✓ Used for causal inference based on the experimental design.	✓ Assumes treatment assignments are fixed. ✓ Treats outcomes as random, considering subjects as a random sample from a larger population. ✓ Used for generalizing results beyond the study sample.
Methods	1. Fisher randomization test 2. Neyman repeated sampling inference	3. Regression methods 4. Model-based imputation

Fisher randomization test

➤ Randomized Experiment on Nocturnal Cough (Paul *et al.*, 2007)

Study Overview:

- Investigates effects of buckwheat honey vs. no active treatment
- Focus on nocturnal cough and sleep difficulties in children with upper respiratory tract infections

Dataset Details:

- 72 children participated
- Measured cough frequency and cough severity
- Scale: 0 (not at all) to 6 (extremely severe)

Table. Cough frequency for the first six units from the honey study

Unit	Potential outcomes		Observed variables	
	Potential (A=1)	Control (A=0)	Treatment	Observed outcome Y
1	$Y_1(1) = 3$	$Y_1(0) = ?$	A=1	$Y_1 = 3$
2	$Y_2(1) = 5$	$Y_2(0) = ?$	A=1	$Y_2 = 5$
3	$Y_3(1) = 0$	$Y_3(0) = ?$	A=1	$Y_3 = 0$
4	$Y_4(1) = ?$	$Y_4(0) = 4$	A=0	$Y_4 = 4$
5	$Y_5(1) = ?$	$Y_5(0) = 0$	A=0	$Y_5 = 0$
6	$Y_6(1) = ?$	$Y_6(0) = 1$	A=0	$Y_6 = 1$

Question:

What types of null hypotheses can be considered, and how can corresponding inferences be made?

$$H_0: E(Y(1)) = E(Y(0))$$

or

$$H_0: Y(1) = Y(0)$$

Fisher randomization test (1): Sharp null hypothesis

1. Specify the null hypothesis

Sharp null hypothesis of no treatment/causal effect (Rubin, 1980)

$$H_{0F}: Y_i(0) = Y_i(1) \text{ for all units } i = 1, \dots, N$$

- Under the sharp null hypothesis, it allows each unit to be assigned a hypothetical value (observed outcome) for its unobserved potential outcome.

Table. Cough frequency for the first six units from the honey study

Unit	Potential outcomes		Observed variables		Potential outcomes under the <i>sharp null hypothesis</i>	
	Potential (A=1)	Control (A=0)	Treatment	Observed outcome	Potential (A=1)	Control (A=0)
1	$Y_1(1) = 3$	$Y_1(0) = ?$	A=1	$Y_1 = 3$	$Y_1(1) = 3$	$Y_1(0) = 3$
2	$Y_2(1) = 5$	$Y_2(0) = ?$	A=1	$Y_2 = 5$	$Y_2(1) = 5$	$Y_2(0) = 5$
3	$Y_3(1) = 0$	$Y_3(0) = ?$	A=1	$Y_3 = 0$	$Y_3(1) = 0$	$Y_3(0) = 0$
4	$Y_4(1) = ?$	$Y_4(0) = 4$	A=0	$Y_4 = 4$	$Y_4(1) = 4$	$Y_4(0) = 4$
5	$Y_5(1) = ?$	$Y_5(0) = 0$	A=0	$Y_5 = 0$	$Y_5(1) = 0$	$Y_5(0) = 0$
6	$Y_6(1) = ?$	$Y_6(0) = 1$	A=0	$Y_6 = 1$	$Y_6(1) = 1$	$Y_6(0) = 1$

Fisher randomization test (2): test statistic

2. Choose a test statistic

- Test statistic $T = T(\tilde{A}, \tilde{Y})$: a real-valued function of **observed outcomes** and **treatment assignments**
- Any test statistic for quantifying the contrast between the treatment and control groups can be used.

Difference-in-means

$$T_1(\tilde{A}, \tilde{Y}) = \frac{\sum_{i=1}^N A_i Y_i}{N_1} - \frac{\sum_{i=1}^N (1 - A_i) Y_i}{N_0}$$

where $N_1 = \sum_{i=1}^N A_i$ and $N_0 = \sum_{i=1}^N (1 - A_i)$

Wilcoxon rank sum

$$T_2(\tilde{A}, \tilde{Y}) = \frac{\sum_{i=1}^N A_i R_i}{N_1} - \frac{\sum_{i=1}^N (1 - A_i) R_i}{N_0}$$

where $R_i = \sum_{j=1}^N I(Y_j \leq Y_i)$

Table. Cough frequency for the first six units from the honey study

Unit	Potential outcomes		Observed variables		Potential outcomes under the <i>sharp null hypothesis</i>	
	Potential (A=1)	Control (A=0)	Treatment	Observed outcome	Potential (A=1)	Control (A=0)
1	$Y_1(1) = 3$	$Y_1(0) = ?$	A=1	$Y_1 = 3$	$Y_1(1) = 3$	$Y_1(0) = 3$
2	$Y_2(1) = 5$	$Y_2(0) = ?$	A=1	$Y_2 = 5$	$Y_2(1) = 5$	$Y_2(0) = 5$
3	$Y_3(1) = 0$	$Y_3(0) = ?$	A=1	$Y_3 = 0$	$Y_3(1) = 0$	$Y_3(0) = 0$
4	$Y_4(1) = ?$	$Y_4(0) = 4$	A=0	$Y_4 = 4$	$Y_4(1) = 4$	$Y_4(0) = 4$
5	$Y_5(1) = ?$	$Y_5(0) = 0$	A=0	$Y_5 = 0$	$Y_5(1) = 0$	$Y_5(0) = 0$
6	$Y_6(1) = ?$	$Y_6(0) = 1$	A=0	$Y_6 = 1$	$Y_6(1) = 1$	$Y_6(0) = 1$

Fisher randomization test (3): randomization distribution

treatment assignments						Statistics	
A_1	A_2	A_3	A_4	A_5	A_6	Difference-in-means	Wilcoxon rank sum
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

*Observed values are shown in red.

3. Generate the randomization distribution

Definition 2.1 (Randomization Distribution)

The randomization distribution is the distribution of a test statistic (such as the difference in means) obtained by

- (1) *permuting the treatment assignments*
- (2) while *keeping the observed outcomes fixed*.

Fisher randomization test (3): randomization distribution

treatment assignments						Statistics	
A_1	A_2	A_3	A_4	A_5	A_6	Difference-in-means	Wilcoxon rank sum
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

*Observed values are shown in red.

3. Generate the randomization distribution

Procedure:

(i) Conduct all possible treatment assignments:

$$\tilde{A}^{(1)}, \tilde{A}^{(2)}, \dots, \tilde{A}^{(M)}$$

where $M = C_{N_1}^N$

(For example, in the randomized experiment on nocturnal cough, there are $C_3^6 = 20$ possible treatments assignments)

(ii) Calculate the values of the test statistic for each possible treatment assignment:

$$\mathbb{F} = \{T(\tilde{A}^{(1)}, \tilde{Y}), T(\tilde{A}^{(2)}, \tilde{Y}), \dots, T(\tilde{A}^{(M)}, \tilde{Y})\}$$

(Note: In a CRE, each possible treatment assignment to units has an *equal probability*.)

(iii) $\mathbb{F}$ forms the randomization distribution

Fisher randomization test (4): Fisher's exact p-value

treatment assignments						Statistics	
A_1	A_2	A_3	A_4	A_5	A_6	Difference-in-means	Wilcoxon rank sum
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

*Observed values are shown in red.

4. Compute Fisher's exact p-value

One-sided

$$P_F = \frac{1}{M} \sum_{i=1}^M I\left(T(\tilde{A}^{obs}, \tilde{Y}) \geq T(\tilde{A}^{(i)}, \tilde{Y})\right)$$

or

$$P_F = \frac{1}{M} \sum_{i=1}^M I\left(T(\tilde{A}^{obs}, \tilde{Y}) \leq T(\tilde{A}^{(i)}, \tilde{Y})\right)$$

Two-sided

$$P_F = \frac{1}{M} \sum_{i=1}^M I(|T(\tilde{A}^{obs}, \tilde{Y})| \leq |T(\tilde{A}^{(i)}, \tilde{Y})|)$$

Experiment on nocturnal cough

Difference-in-means: $P_F = 16/20 = 0.8$

Wilcoxon rank sum: $P_F = 16/20 = 0.8$

Neyman vs. Fisher

- The limitations of Fisher randomization test
 1. Do not account for treatment effect heterogeneity
 2. Do not support inference at the population level

- During the development of statistical inference methods, Fisher and Neyman introduced distinct frameworks that have significantly influenced causal inference.

- Neyman's two questions:
 1. What would the average outcome be if all units were exposed to the treatment?
 2. How did that compare to the average outcome if all units were exposed to the control?

Neyman repeated sampling inference

- Fisher emphasizes hypothesis testing, while Neyman focuses on parameter estimation.

Fisher randomization test

- Sharp null hypothesis (**No individual treatment effect** within the analytic sample)
- Considers only randomness from treatment assignment
- Derives exact **p-values**

Neyman repeated sampling inference

- Focuses on **population/sample** average treatment effect
- Accounts for two sources of randomness:
 1. Random sampling from a (super-) population
 2. Random treatment assignment
- Derives unbiased estimators and confidence intervals (considers variability across both **repeated treatment assignments** and **repeated sampling**)

- The causal estimands of interest in the Neyman inference:
 - Sample average treatment effect (SATE)
 - Population average treatment effect (PATE)

SATE and PATE

➤ Sample average treatment effect (SATE)

$$\tau_s \equiv \frac{1}{N} \sum_{i=1}^N (Y_i(1) - Y_i(0)) = \bar{Y}(1) - \bar{Y}(0)$$

where N represents the total number of units in the experiment and $\bar{Y}(a) = N^{-1} \sum_{i=1}^N Y_i(a)$.

➤ Population average treatment effect (PATE)

$$\tau_p \equiv \mathbb{E}(Y_i(1) - Y_i(0))$$

- Note 1: ITE is $\tau_i \equiv Y_i(1) - Y_i(0)$
- Note 2: SATE is precisely the ATE as previously defined.

➤ Questions for SATE

1. Is SATE a **parameter** or an **estimator**?
2. In SATE, are $\{Y_i(1), Y_i(0)\}_{i=1}^N$ **random** or **fixed**? How does this compare to PATE
3. Who cares about SATE, and why it is important?

Misunderstandings between experimentalists and observationalists

(Imai, King, and Stuart, 2008)

	SATE (τ_s)	PATE (τ_p)
Scope	Applies to study sample	Applies to entire population
Estimability	Directly computed from study data	Requires extrapolation beyond sample
Generalizability	Limited to sample	Requires assumptions for external validity
Bias risks	Susceptible to selection bias if sample is not random	More representative but harder to estimate

- **SATE is useful for internal validity** (causal inference within a study).
- **PATE is necessary for external validity** (generalizing findings to a broader population).
- **Bridging the gap** between SATE and PATE requires either:
 1. Random sampling,
 2. Statistical adjustments (e.g., weighting, regression models),
 3. Replicating studies in diverse settings.

Estimator of SATE

- Reminder: In SATE inference under CRE, $\{\tilde{Y}(1), \tilde{Y}(0)\} = \{Y_i(1), Y_i(0)\}_{i=1}^N$ are fixed, and the only source of randomness comes from the random treatment assignments.

- **The difference-in-means estimator of SATE**

$$\hat{\tau}_s = \frac{1}{N_1} \sum_{i=1}^N A_i Y_i - \frac{1}{N_0} \sum_{i=1}^N (1 - A_i) Y_i$$

where N_1 and N_0 represent the number of units in the treatment and control groups, respectively.

- Is $\hat{\tau}_s$ an unbiased estimator for SATE?
- How can you conduct inference for SATE by using $\hat{\tau}_s$, such as deriving a confidence interval?

Estimator of SATE

Theorem 2.1 (Unbiasedness and sampling variance in estimating SATE)

Under a CRE, the difference-in-means estimator $\hat{\tau}_s$

(1) is unbiased for τ_s (SATE) and

(2) has sampling variance

$$\text{Var}\left(\hat{\tau}_s | \tilde{Y}(1), \tilde{Y}(0)\right) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\tau^2}{N}$$

where

$$S_a^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i(a) - \bar{Y}(a))^2 \text{ and } S_\tau^2 = \frac{1}{N-1} \sum_{i=1}^N (\tau_i - \tau_s)^2.$$

- It demonstrates that $\hat{\tau}_s$ is a *reasonable estimator* for SATE under a CRE.
- S_τ^2 includes both potential outcomes $Y_i(1)$ and $Y_i(0)$, making it non-estimable from observed data.
- Neyman suggested that, in practice, S_τ^2 is often ignored when estimating the sampling variance of $\hat{\tau}_s$.

Estimator of the sampling variance

Theorem 2.2

Under *the constant additive treatment effect assumption* (i.e., the individual treatment effect $\tau_i (\equiv Y_i(1) - Y_i(0))$ is constant), we have

- (1) $S_\tau^2 = 0$ and
- (2) an unbiased estimator for the sampling variance

$$\widehat{V} = \frac{\hat{S}_1^2}{N_1} + \frac{\hat{S}_0^2}{N_0}$$

where

$$\hat{S}_1^2 = \frac{1}{N_1 - 1} \sum_{i=1}^N A_i (Y_i - \bar{Y}_{A=1})^2,$$

$$\hat{S}_0^2 = \frac{1}{N_0 - 1} \sum_{i=1}^N (1 - A_i) (Y_i - \bar{Y}_{A=0})^2,$$

$$\bar{Y}_{A=1} = \frac{1}{N_1} \sum_{i=1}^N A_i Y_i, \text{ and } \bar{Y}_{A=0} = \frac{1}{N_0} \sum_{i=1}^N (1 - A_i) Y_i$$

Further insights on $\widehat{V}$

- It is important to note that if there is heterogeneity in treatment effects—meaning the constant additive treatment effect assumption does not hold—then $\widehat{V}$ is a *conservative estimator* of the sampling variance:

$$\mathbb{E} \left(\widehat{V} | \tilde{Y}(1), \tilde{Y}(0) \right) - \text{Var} \left(\hat{\tau}_s | \tilde{Y}(1), \tilde{Y}(0) \right) = \frac{S_{\tau}^2}{N} \geq 0.$$

- Even when the assumption of an additive treatment effect may be known to be inaccurate, $\widehat{V}$ remains widely used (in two-sample testing).
 1. The confidence interval for SATE constructed using $\widehat{V}$ is conservative.
 2. For PATE, in a (super-)population, $\widehat{V}$ is an unbiased estimator for the sampling variance of $\hat{\tau}_s$.

Proof of Theorem 2.1

Unbiasedness

First, the difference-in-means estimator $\hat{\tau}_s$ can be expressed as

$$\hat{\tau}_s = \frac{1}{N_1} \sum_{i=1}^N A_i Y_i(1) - \frac{1}{N_0} \sum_{i=1}^N (1 - A_i) Y_i(0)$$

Then we have

$$\begin{aligned} \mathbb{E}(\hat{\tau}_s | \tilde{\mathbf{Y}}(1), \tilde{\mathbf{Y}}(0)) &= \frac{1}{N_1} \sum_{i=1}^N \mathbb{E}(A_i | \tilde{\mathbf{Y}}(1), \tilde{\mathbf{Y}}(0)) Y_i(1) - \frac{1}{N_0} \sum_{i=1}^N \{1 - \mathbb{E}(A_i | \tilde{\mathbf{Y}}(1), \tilde{\mathbf{Y}}(0))\} Y_i(0) \\ &= \frac{1}{N_1} \sum_{i=1}^N \mathbb{E}(A_i) Y_i(1) - \frac{1}{N_0} \sum_{i=1}^N \{1 - \mathbb{E}(A_i)\} Y_i(0) \\ &= \frac{1}{N_1} \sum_{i=1}^N \frac{N_1}{N} Y_i(1) - \frac{1}{N_0} \sum_{i=1}^N \frac{N_0}{N} Y_i(0) \\ &= \tau_s \end{aligned}$$

Estimator of PATE

Theorem 2.3 (Unbiasedness and sampling variance in estimating PATE)

Under a CRE, the difference-in-means estimator $\hat{\tau}_s$

- (1) is unbiased for τ_p (PATE) and*
- (2) has sampling variance*

$$Var(\hat{\tau}_s) = \frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}$$

where σ_a^2 is the population variance of $Y(a)$ for $a = 1, 2$

- Unbiasedness is straightforward to demonstrate, as the units/samples are randomly drawn from the population.
- **The derivation of sampling variance (over repeated sampling and repeated treatment assignments):**

By the the law of total variance,

$$\begin{aligned} Var(\hat{\tau}_s) &= Var\left(\mathbb{E}(\hat{\tau}_s | \tilde{Y}(1), \tilde{Y}(0))\right) + \mathbb{E}\left(Var(\hat{\tau}_s | \tilde{Y}(1), \tilde{Y}(0))\right) \\ &= \frac{1}{N} Var(Y_i(1) - Y_i(1)) + \frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0} - \frac{1}{N} Var(Y_i(1) - Y_i(1)) = \frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0} \end{aligned}$$

Confidence interval

- Li and Ding (2017) demonstrated the asymptotic Normality of $\hat{\tau}_s$ based on the finite population central limit theorem.

- Wald-type large-sample confidence interval is given by

$$\hat{\tau}_s \pm z_{1-\alpha/2} \sqrt{\hat{V}}$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ upper quantile of the standard Normal distribution.

References

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 171(2), 481-502.

Li, X., & Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520), 1759-1769.

Paul, I. M., Beiler, J., McMonagle, A., Shaffer, M. L., Duda, L., & Berlin, C. M. (2007). Effect of honey, dextromethorphan, and no treatment on nocturnal cough and sleep quality for coughing children and their parents. *Archives of pediatrics & adolescent medicine*, 161(12), 1140-1146.

Rubin, D. B. (1980). Comment on “Randomization analysis of experimental data: The Fisher randomization test”. *Journal of the American statistical association*, 75(371), 591-593.