SISCER Module 13
Lecture 1: Randomization inference

Ting Ye & Qingyuan Zhao

University of Washington & University of Cambridge

July 23, 2025

Plan

- Correlation does not imply causation.
- General theory for randomization inference for randomized experiments.
- Example 1: Fisher's exact test for 2×2 contigency tables.
- Example 2: Stepped-wedge cluster randomized trials.
- Example 3: Matched observational studies.

Recommended references for this lecture

Paul R. Rosenbaum (2002). *Observational Studies*. Springer Series in Statistics. New York: Springer. DOI: 10.1007/978-1-4757-3692-2, Chap. 2;

Yao Zhang and Qingyuan Zhao (Apr. 2023). "What Is a Randomization Test?" In: *Journal of the American Statistical Association* 118.544, pp. 2928–2942. DOI: 10.1080/01621459.2023.2199814.

Outline

Causality and association

- Causality is central to human knowledge.
- The major part of classic statistics is about association (e.g., Pearson correlation, regression coefficient) rather than causation.
  - Association/correlation describes the **statistical relationship** in the data, indicating **difference in one variable is associated with difference in another**.
  - Association / correlation does not imply causation.
  - May be good for prediction but not enough for causation.
- Causation requires **mechanistic understanding**, indicating whether **intervention in one variable leads to change in another**.

Yule-Simpson paradox

	Success	Failure
Open surgical procedure	273	77
Small puncture procedure	289	61

- From Clive R Charig et al. (1986). "Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy.". In: *Br Med J (Clin Res Ed)* 292.6524, pp. 879–882. DOI: 10.1136/bmj.292.6524.879.
- Estimated risk difference: $\widehat{\mathrm{RD}} = \underbrace{\frac{273}{273+77}}_{\text{open}} - \underbrace{\frac{289}{289+61}}_{\text{small}} = 78\% - 83\% = -5\% < 0.$
- Success rate is higher among the small puncture group **(association)**
- But is small puncture procedure better? **(causation)**

Yule-Simpson paradox

- Patients were **not randomized** into the two procedures
- Patients receiving **open surgical** tend to have **large** stones, whereas patients receiving **small puncture** tend to have **small** stones.

With small stones	Success	Failure
Open surgical	81	6
Small puncture	234	36

With large stones	Success	Failure
Open surgical	192	71
Small puncture	55	25

- **Yule-Simpson's paradox**:

$$\widehat{\mathrm{RD}}_S = \frac{81}{8+6} - \frac{234}{234+46} = 6\% > 0,\ \widehat{\mathrm{RD}}_L = \frac{192}{192+71} - \frac{55}{55+25} = 4\% > 0,\ \text{but } \widehat{\mathrm{RD}} = -5\% < 0.$$

- **Confounding**: stone size affects both treatment assignment and success rate.

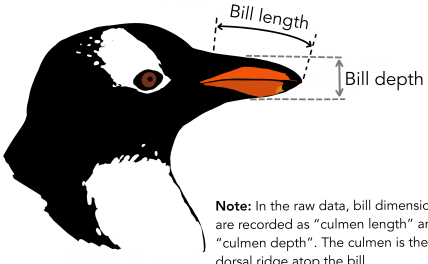
Another example with linear regression

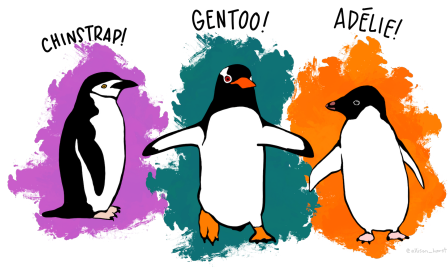
- The kidney stone example shows that marginal association and conditional association may have **different signs**.
- Here is another example (non-causal but very memorable).[1]

[1] Dataset and R code for the figures below can be found at `https://allisonhorst.github.io/palmerpenguins/articles/examples.html`.

Another example with linear regression

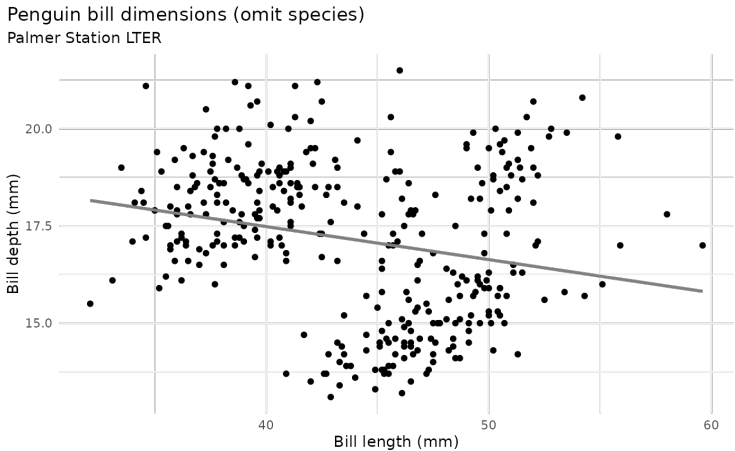
- The kidney stone example shows that marginal association and conditional association may have **different signs**.
- Here is another example (non-causal but very memorable).[1]



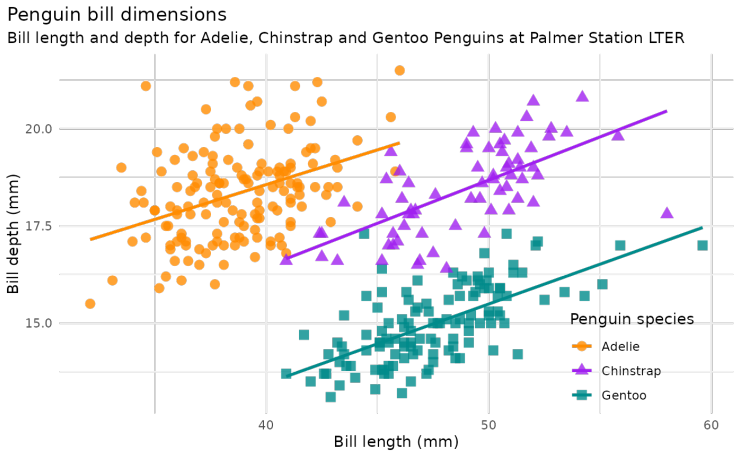


[1] Dataset and R code for the figures below can be found at `https://allisonhorst.github.io/palmerpenguins/articles/examples.html`.

Marginal association



Conditional association



Outline

Setup

Suppose there are n units (e.g. clinics or patients) in an experiment.

- **Covariates** $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots \boldsymbol{X}_n)$.
- **Treatment** $\boldsymbol{Z} \in \mathcal{Z}$ is randomly determined (e.g. by tossing coins or using the RNG in R).
- **Exposure** $\boldsymbol{A} = (A_1, \ldots, A_n)$ is determined by $\boldsymbol{Z}$.
  - Semantically, "treatment" speaks from the investigator's perspective and "exposure" from the experimental unit's perspective.
  - Often (but not always), $\boldsymbol{A} = \boldsymbol{Z}$ and these terms are used interchangeably.
- **Outcome** $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$.

Conceptulizing causality

Every possible treatment assignment $\boldsymbol{z}$ corresponds to a vector of **potential outcomes**

$$\boldsymbol{Y}(\boldsymbol{z}) = (Y_1(\boldsymbol{z}), \ldots, Y_n(\boldsymbol{z})).$$

Assumptions

- The observed outcomes are **consistent** with the potential outcomes: $\boldsymbol{Y} = \boldsymbol{Y}(\boldsymbol{Z})$.
- The **exposure map is valid**: if $A_i(\boldsymbol{z}) = A_i(\tilde{\boldsymbol{z}})$, then $Y_i(\boldsymbol{z}) = Y_i(\tilde{\boldsymbol{z}})$.
  - Under this assumption, the potential outcome is also denoted as $Y_i(a_i)$.

The Neyman-Rubin causal model

- Further assumes $A_i(\boldsymbol{z}) = z_i$ (**no interference**).
- Often $z_i \in \{0\,(\text{control}), 1\,(\text{treatment})\}$ is binary, so the **individual treatment effect** of unit i is $Y_i(1) - Y_i(0)$.

Fundamental problem of causal inference

- Only one potential outcome can ever be observed.
- But we would like to infer the full *potential outcomes schedule* $(\boldsymbol{Y}(\boldsymbol{z}))_{\boldsymbol{z}\in\mathcal{Z}}$.

i	$Y_i(0)$	$Y_i(1)$	A_i	Y_i
1	?	**1**	1	1
2	**0**	?	0	0
3	?	**0**	1	0
⋮	⋮	⋮	⋮	⋮

- $A_i = 0$: open surgical procedure; $A_i = 1$: small puncture procedure.
- $Y_i = 0$: failure; $Y_i = 1$: success.

The role of randomization

Assumption: Exogeneity of randomization

The treatment is independent of the potential outcomes schedule given the covariates:

$$\boldsymbol{Z} \perp\!\!\!\perp (\boldsymbol{Y}(\boldsymbol{z}))_{\boldsymbol{z}\in\mathcal{Z}} \mid \boldsymbol{X}. \tag{1}$$

Furthermore, the conditional distribution of $\boldsymbol{Z}$ given $\boldsymbol{X}$ is **known** (often called the *randomization scheme* or *treatment assignment mechanism*).

Remarks

- If the randomization scheme does not use $\boldsymbol{X}$, we can drop $\boldsymbol{X}$ in (1).

Imputation of potential outcomes

Next we will explore, in the Neyman-Rubin causal model, how to use randomization to test the **sharp null hypothesis** $H_0 : Y_i(0) = Y_i(1)$ for all i.

Key insight

Under H_0, we may impute all the potential outcomes by $Y_i(0) = Y_i(1) = Y_i$.

Example

i	$Y_i(0)$	$Y_i(1)$	A_i	Y_i
1	**1**	1	1	1
2	0	**0**	0	0
3	**0**	0	1	0
⋮	⋮	⋮	⋮	⋮

Randomization distribution

- Consider any test statistic $T = T(\boldsymbol{A}, \boldsymbol{Y})$.
- Under a potential treatment assignment $\boldsymbol{a} = (a_1, \ldots, a_n)$, the corresponding statistic is $T(\boldsymbol{a}) = T(\boldsymbol{a}, \boldsymbol{Y}(\boldsymbol{a}))$.
- The last insight suggests that under H_0, we know the value of $T(\boldsymbol{a})$ for every $\boldsymbol{a}$.
- The *randomization distribution* is that of $T(\boldsymbol{A})$ under the randomization scheme.

Example: An simple estimator of the average treatment effect

$$T = \frac{\sum_{i=1}^n A_i Y_i}{\sum_{i=1}^n A_i} - \frac{\sum_{i=1}^n (1 - A_i) Y_i}{\sum_{i=1}^n (1 - A_i)}.$$

i	$Y_i(0)$	$Y_i(1)$	A_i	Y_i
1	**1**	1	1	1
2	0	**0**	0	0
3	**0**	0	1	0

Equal probability $(1/3)$ on $T(1,0,1) = 1/2$; $T(1,1,0) = 1/2$; $T(0,1,1) = -1$.

Randomization tests

We may reject the null hypothesis H_0 if the observed $T(\boldsymbol{A})$ is too extreme when compared to its potential values.

- The p-value is the probability that T exceeds the observed value:

$$P = \Pr(T(\boldsymbol{A}^*, \boldsymbol{Y}(\boldsymbol{A}^*)) > T(\boldsymbol{A}, \boldsymbol{Y}(\boldsymbol{A})) \mid \boldsymbol{Z}, \boldsymbol{Y}(\cdot)),$$

where $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{Z})$, $\boldsymbol{A}^* = \boldsymbol{A}(\boldsymbol{Z}^*)$, and $\boldsymbol{Z}^*$ is an independent and identically distributed copy of $\boldsymbol{Z}$.

Remarks

- Compared the usual formulation of hypothesis testing (e.g. t-test and F-test), randomization test uses a **reference distribution that is entirely based on randomization generated by the experiment**.
- Randomization tests are particularly attractive for small sample sizes and complex design (e.g. repeated measurements, individuals from the same household).

Outline

Example: Fisher's exact test and lady tasting tea

- A basic yet important statistical problem is hypothesis testing in 2×2 contingency tables.
- This is illustrated by a famous example in Fisher's 1935 book *The Design of Experiments*.

  A lady declares that by tasing a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup... Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment in a random order. The subject has been told in advance of what the test will consist... Her task is to divide the 8 cups into two sets of 4.

- Exercise: What are the units, treatment, exposure, and outcome in this experiment?

2×2 contigency tables

- Let A_i be the exposure of the i-th cup (0/1 if milk/tea was added first).
- Let Y_i be the outcome of the i-th cup (0/1 if the lady guesses milk/tea was added first).
- Let N_{ay} be the number of cups with $A_i = a$ and $Y_i = y$, $a, y = 0, 1$.
- The outcome of this experiment can be summarized by the following 2×2 table.

		Outcome Y		
		0	1	Total
Treatment A	0	N_{00}	N_{01}	$N_{0\cdot}$
	1	N_{10}	N_{11}	$N_{1\cdot}$
	Total	$N_{\cdot 0}$	$N_{\cdot 1}$	N

Fisher's exact test (abstract)

		Outcome Y 0	1	Total
Treatment A	0	N_{00}	N_{01}	$N_{0\cdot}$
	1	N_{10}	N_{11}	$N_{1\cdot}$
	Total	$N_{\cdot 0}$	$N_{\cdot 1}$	N

- Null hypothesis $H_0 : Y_i(0) = Y_i(1)$ for all i, meaning the lady's guess is random.
- $N_{0\cdot} = N_{1\cdot} = 4$ by design and $N_{\cdot 0} = N_{\cdot 1}$ by H_0.
- So there is only one degree of freedom: Given N_{00}, the entire table is known.
- Fisher showed that the probability of observing $(N_{00}, N_{01}, N_{10}, N_{11})$ is given by

$$\frac{\binom{N_{0\cdot}}{N_{01}}\binom{N_{1\cdot}}{N_{11}}}{\binom{N}{N_{0\cdot}}} = \frac{N_{0\cdot}!N_{1\cdot}!N_{\cdot 0}!N_{\cdot 1}!}{N_{00}!N_{01}!N_{10}!N_{11}!N!}$$

- So we may reject H_0 if N_{00} is large (compared to this hypergeometric distribution).

Fisher's exact test (example)

- Suppose the lady gave random guesses and got 3 milk-first cups correct.

		Outcome Y		
		0	1	Total
Treatment A	0	3	1	4
	1	1	3	4
	Total	4	4	8

$$
\begin{aligned}
P &= \Pr(\text{guessed} \geq 3 \text{ correctly} \mid \text{random guesses}) \\
&= \Pr(\text{guessed 4 correctly} \mid \text{random guesses}) + \Pr(\text{guessed 3 correctly} \mid \text{random guesses}) \\
&= \frac{\binom{4}{4}\binom{4}{4}}{\binom{8}{4}} + \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = \frac{1}{70} + \frac{16}{70} = \frac{17}{70} = 0.243.
\end{aligned}
$$

Example: Stepped-wedge design and a real clinical trial

Haines et al., *PLOS Medicine*, 2017, DOI:10.1371/journal.pmed.1002412.

- The goal was to investigate the impact of disinvestment from weekend allied health services across acute medical and surgical wards.
- 12 wards in 2 hospitals were randomized to switch from an old model of weekend allied health services to no services, before adopting a new model of services. (You can visualize the design in Figure 1 of the article.)
- During this trial, a number of patient characteristics were collected. Of interest is the average length of stay in these wards.
- Exercise: What are the units, treatment, exposure, and outcome in this experiment?
- This example will be further explored in the `R` practical.

Example: Matched observational studies

- By matching units in an observational studies with very similar covariates, the hope is that we reconstruct a block randomized experiment.
- Consider the Neyman-Rubin causal model. Suppose treated observation $i = 1, \ldots, n$ is matched to control observation $i + n$. Define

$$\mathcal{M} = \{\boldsymbol{a}_{[2n_1]} \in \{0,1\}^{2n_1} \mid a_i + a_{i+n_1} = 1, \forall i \in [n_1]\}$$

- Randomization analysis of matched observational studies assumes

$$\mathbb{P}\Big(\boldsymbol{A} = \boldsymbol{a} \,\Big|\, \boldsymbol{X}, \boldsymbol{A} \in \mathcal{M}\Big) = \begin{cases} 2^{-n_1}, & \text{if } \boldsymbol{a} \in M, \\ 0, & \text{otherwise.} \end{cases}$$

- This would be satisfied if $(\boldsymbol{X}_i, A_i)$ are drawn i.i.d. (independent and identically distributed) from a population and the matching is exact.
- By further assuming no unmeasured confounders $A_i \perp\!\!\!\perp Y_i(a) \mid \boldsymbol{X}_i$ for all a, randomization tests can be constructed in the same way as before.