STA 640 — Causal Inference

Chapter 2.1. Randomized Experiments: Fisher's and Neyman's Mode of Inference

Fan Li

Department of Statistical Science
Duke University

Role of Randomization

- In randomized experiments, assignment mechanism is known and controlled by investigators
- Randomization does:
  - balance observed covariates: $Z \perp\!\!\!\perp X$
  - balance unobserved covariates: $Z \perp\!\!\!\perp U$
  - balance potential outcomes, i.e. guarantee ignorability or unconfoundedness (Rubin 1978)

$$Z \perp\!\!\!\perp (Y(1), Y(0))$$

- Note on terminology: the terms *covariates, pretreatment variables, baseline characteristics* are used exchangeably in the literature, referring to X

Role of Randomization

- Under randomization, unconfoundedness holds by design (without conditioning on covariates X.
- Causal effects are (nonparametrically) identified, because we can show

$$\Pr(Y_i(z)) = \Pr(Y_i | Z_i = z), \quad z = 0, 1$$

- Under randomization, *association does imply causation* (of course within the potential outcome framework with assumptions), for example:

$$\tau^{\text{ATE}} \equiv \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y_i | Z_i = 1] - \mathbb{E}[Y_i | Z_i = 0]$$

- Randomization ensures *model-free validity for model-based statistics* (explain later)

Types of Randomized Experiments

1. Bernoulli trials
2. Completely randomized experiment
3. Stratified (or block) randomized experiment
4. Paired randomized experiment
5. Cluster randomized experiment

Bernoulli Trials

Notation: we use bold font to denote the vector of all units of a variables, e.g. $\mathbf{Z} = \{Z_1, ..., Z_N\}, \mathbf{Y}(0) = \{Y_1(0), ..., Y_N(0)\}$

1. *Bernoulli trials*: for each unit we flip a fair coin to determine their assignment
   - The assignment mechanism is

$$p(\mathbf{Z}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = \left(\frac{1}{2}\right)^N$$

   - Here $p(Z_i = 1|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = \frac{1}{2}$
   - The number of possible assignments is 2^N

Completely Randomized Experiment

2. *Completely randomized experiment*: N_1 out of N units are randomly chosen to receive the treatment ($N_0 = N - N_1$) is the no. of control)
   - The assignment mechanism is

$$p(\mathbf{Z} \mid \mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = \begin{cases} \binom{N_1}{N}^{-1} \left(\frac{N_1}{N}\right)^{N_1} \left(\frac{N_0}{N}\right)^{N_0} & \text{if } \sum_{i=1}^{N} Z_i = N_1 \\ 0 & \text{otherwise.} \end{cases}$$

   - Here $p(Z_i = 1|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = \frac{N_1}{N}$
   - Most often $N_1 = N/2$, so that half the units receive the active treatment and half receive the control treatment
   - The number of possible assignments is $\binom{N}{N_1}$

Stratified (block) Randomized Experiment

3. *Stratified or block randomized experiment*: First partition into blocks or strata, similar in some sense (e.g. covariates), and then within each block, perform a completely randomized trial
   - Assignments are independent across blocks
   - Example: block 1 - women; block 2- men
   - *"Block what you can; randomize what you cannot" - George Box*

4. *Paired randomized experiments*: a special case of stratified trials, with 2 units within each block. Total number of possible assignments - $2^{N/2}$.

5. *Clustered randomized experiments*: treatment assigned at cluster level, all units in one unit get the same treatment (requires additional notations, more later)

Fisher's Randomization Inference

- Fundamental idea: inference based *solely on the assignment mechanism*
- $Y(0), Y(1)$ are all fixed, randomness only comes from the assignment of Z
- RA Fisher was the first to grasp the importance of randomization for credibly assessing causal effects (1925, 1935)
- Given data from such a randomized setting, Fisher was intent on testing *sharp null hypotheses*

Fisher's Exact Test of Sharp Null Hypothesis

- Null Hypothesis (H_0): An initial supposition regarding the nature of the treatment effect, usually specified to test its consistency against observed data
- Sharp Null Hypothesis: A null hypothesis articulated with specificity sufficient to allow the researcher to fill in a hypothetical value for each unit's missing potential outcome
- Test Statistic ($T = T(\mathbf{Y}(1), \mathbf{Y}(0), \mathbf{Z})$): A function of the data that the researcher uses to determine the consistency of the null hypothesis with observed data

Fisher's Exact Test: Three Steps

1. Specify the null hypothesis
2. Choose the statistic
3. Measure the extremeness by the p-value (under the randomization distribution)

Fisher's Exact Test of Sharp Null: Example

Imbens and Rubin (2009): A small subset of data from a randomized job training experiment — the Greater Avenues to INdependence (GAIN) programs. Earnings in First Year Post-randomization:

Individual	Potential Outcomes		Actual	Observed Outcome
	$Y_i(0)$	$Y_i(1)$	Treatment	Y_i
1	66	?	0	66
2	0	?	0	0
3	0	?	0	0
4	?	0	1	0
5	?	607	1	607
6	?	436	1	436

Fisher's Exact Test: Example

- We first test the sharp null hypothesis that the program had absolutely no effect on reading scores:

$$H_0 : Y_i(0) = Y_i(1) \text{ for all } i = 1, \dots, 6$$

- Under H_0, the unobserved potential outcomes are equal to the observed outcomes for each unit

Individual	Potential Outcomes $Y_i(0)$	$Y_i(1)$	Actual Treatment	Observed Outcome Y_i
1	66	(66)	0	66
2	0	(0)	0	0
3	0	(0)	0	0
4	(0)	0	1	0
5	(607)	607	1	607
6	(436)	436	1	436

Fisher's Exact Test: Example

- Consider testing the null hypothesis against the alternative hypothesis that $Y_i(0) \neq Y_i(1)$ for some units
- We use the difference in average outcomes by treatment status as test statistic:

$$T = T(\mathbf{Z}, Y^{obs}) = \frac{1}{3}\sum_{i=1}^{6} Z_i \cdot Y_i^{obs} - \frac{1}{3}\sum_{i=1}^{6}(1 - Z_i) \cdot Y_i^{obs}$$

- For the observed data the value of the test statistic is 325.6
- Under the null hypothesis, we can calculate the value of this statistic under each vector of treatment assignments $\mathbf{Z}$ (20 different assignment vectors)

Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	levels
0	0	0	1	1	1	325.6
0	0	1	0	1	1	325.6
0	0	1	1	0	1	-79.0
0	0	1	1	1	0	35.0
0	1	0	0	1	1	325.6
0	1	0	1	0	1	-79.0
0	1	0	1	1	0	35.0
0	1	1	0	0	1	-79.0
0	1	1	0	1	0	35.0
0	1	1	1	0	0	369.6
1	0	0	0	1	1	369.6
1	0	0	1	0	1	-35.0
1	0	0	1	1	0	79.0
1	0	1	0	0	1	-35.0
1	0	1	0	1	0	79.0
1	0	1	1	0	0	325.6
1	1	0	0	0	1	-35.0
1	1	0	0	1	0	79.0
1	1	0	1	0	0	325.6
1	1	1	0	0	0	325.6

Fisher's Exact Test: Example

- How likely is (under the randomization distribution) to observe a value of the test statistic that is as large in absolute value as the one actually observed?
- Exact p-value: $\Pr(T \geq T^{obs}) = 8/20 = 0.40$ for our sharp null hypothesis, our test statistic, and our definition of unusual
- In this example, no difference between one-way and two-way tests. But usually matters
- Under the null hypothesis of no effect of the program the observed difference could therefore well be due to chance

Fisher's Exact Test: Definition

- *Randomization Distribution*: All possible values of the test statistic and the probabilities associated with each such value under the (randomized) assignment mechanism, assuming the null hypothesis to be true
- *p-value*: Under the assumption that the null hypothesis is correct, the probability of observing a value of the test statistic as extreme or more extreme than the one actually observed
- *Fisher Interval*: The set of possible values of the causal quantity of interest corresponding to test statistics with p-values that fall within some range set by the researcher; usually interpreted to be a set of plausible values of the average causal effect

Fisher's Exact Test: Choice of Statistic

- t statistic

$$T = \frac{1}{n_1} \sum_{i=1}^{n} Z_i Y_i - \frac{1}{n_0} \sum_{i=1}^{n} (1 - Z_i) Y_i$$

- Wilcoxon statistic based on ranks R_i's of Y_i's

$$W = \frac{1}{n_1} \sum_{i=1}^{n} Z_i R_i - \frac{1}{n_0} \sum_{i=1}^{n} (1 - Z_i) R_i$$

- Special case with binary outcome: equivalent to Fisher's exact test

Covariate adjustment in randomization test

- What if there is chance imbalance in baseline (pretreatment) covariates?
- Regression adjustment:
  - OLS (or other model) fitting of the Y_i's on pretreatment covariates
  - obtain residuals ϵ_i's
- Key facts under the sharp null $Y_i(1) = Y_i(0)$ for all i
  - values of Y_i's and X_i's are fixed
  - values of ϵ_i's are fixed
- Replace Y_i by ϵ_i in the test statistic
- Or replace R_i by the rank of ϵ_i
- Validity of randomization test does not rely on the linear model assumption (Rosenbaum 2002 Stat Sci)

Pros and cons of randomization test

- Pros
  - exact, no need for asymptotics
  - works for any test statistic
  - test statistics = any functions of data (fitted parameters of any models)
  - can deal with heavy-tailed data
- Cons
  - exact inference is attractive, but sharp null seems dull (and restrictive). Extended to less restrictive null, i.e. $H_0 : Y(1) = Y(0) + \tau$ by Ding et al. (2016), but still dull and restrictive
  - computationally intensive, especially in large experiments
  - often less powerful than methods mentioned later
  - no guidance for choice of test statistic (power?)
  - confidence interval is hard unless constant causal effects

Neyman's Repeated Sampling Approach

- Focus on methods for estimating *average treatment effects*
- Neyman's two questions:
  - What would the average outcome be if all units were exposed to the treatment?
  - How did that compare to the average outcome if all units were exposed to the control?
- Neyman's procedure (classic frequentist's view):
  1. Find an *unbiased estimator* of the ATE
  2. Obtain variance and interval estimates of the unbiased estimator (derive the estimator's distribution under *repeated sampling* based on the randomization distribution of the assignment vector Z)

Neyman's Repeated Sampling: Basic concepts

- Estimand: average treatment effects (ATE).
- Note: in randomized experiments, ATE is the same as ATT or ATC (average treatment effect for the control). Why?
- Unbiased estimator of the estimand: A statistic whose average value over all possible randomizations equals the true treatment effect
- Confidence (Neyman) Interval: An interval which, over all possible randomizations, includes the true value of the quantity of interest at least as often as advertised. For example, a 95% confidence interval includes the true value 95% of the time or more
- Variance: find the *sampling* variance of the unbiased estimator of the effect and an unbiased estimator of this variance

Estimand: Population Average Treatment Effect (PATE)

- Population Average Treatment Effect (PATE):

$$\tau^{\text{PATE}} \equiv \mathbb{E}[Y_i(1) - Y_i(0)]$$

- The sample under study is a simple random sample from a super population with sample size $\tilde{N} \gg N_0 + N_1$

Estimand: Sample Average Treatment Effect (SATE)

- Another estimand is Sample Average Treatment Effect (SATE)

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

- SATE focuses on the sample (size $N = N_0 + N_1$) under study
- SATE is often of interest in randomized trials, but rarely in observational studies

- Subtle theoretical difference between PATE and SATE, see Imbens and Rubin (2015, Chapter 6), but little difference in practice

- We will mostly focus on PATE (referred to simply as ATE)

PATE: Unbiased Estimator

- An unbiased estimator of PATE is the difference-in-means between groups, also known as the unadjusted estimator:

$$\hat{\tau}^{\text{unadj}} = \sum_{i=1}^{N} \left(\frac{Z_i \cdot Y_i}{N_1} \right) - \sum_{i=1}^{N} \left(\frac{(1 - Z_i) \cdot Y_i}{N_0} \right) = \bar{Y}_1 - \bar{Y}_0$$

- Unbiasedness is with respect to *all possible randomization assignments* $\mathbf{Z}$ *and sampling*

$$\mathbb{E}[\hat{\tau}^{\text{unadj}} | Y(0), Y(1)] = \tau^{\text{PATE}}$$

PATE: Variance of Unadjusted Estimator

- Variance of the unadjusted estimator $\hat{\tau}^{\text{unadj}}$

$$\mathbb{V}[\hat{\tau}^{\text{unadj}}] = \frac{S_0^2}{N_0} + \frac{S_1^2}{N_1} - \frac{S_{01}^2}{\tilde{N}}$$

where $\tilde{N}$ is the size of the super population, and

$$S_z^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i(z) - \bar{Y}(z))^2 \quad z = 0, 1$$

$$S_{01}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i(1) - Y_i(0) - (\bar{Y}(1) - \bar{Y}(0)))^2$$

- This estimator of the variance is unbiased if the super population size $\tilde{N}$ is infinitely large

(Sample) Estimator of Variance

- Because we never observe both $Y_i(1)$ and $Y_i(0)$, S_{01} is not estimable from data
- So, as Neyman suggested, usually one just ignores S_{01} in estimating variance of ATE:

$$\mathbb{V}^{\text{Neyman}} = \frac{S_0^2}{N_0} + \frac{S_1^2}{N_1}$$

- A unbiased estimator of the Neyman variance $\mathbb{V}^{\text{Neyman}}$ is to replace S_z by its sample version

$$\hat{\mathbb{V}}^{\text{Neyman}} = \frac{s_0^2}{N_0} + \frac{s_1^2}{N_1}$$

where

$$s_z^2 = \frac{1}{N_z - 1} \sum_{i:Z_i=1} \left(Y_i - \bar{Y}_z\right)^2, \quad z = 0, 1$$

Confidence Interval

- Usually based on a normal approximation to the randomization (and sampling) distribution of the estimate $\hat{\tau}^{\text{unadj}}/\mathbb{V}_{\text{Neyman}}$
- A standard confidence interval of level $(1-\alpha)\times 100\%$ is

$$CI^{1-\alpha}(\hat{\tau}) = \left(\hat{\tau}^{\text{unadj}} - z_{1-\alpha/2}\cdot\sqrt{\hat{\mathbb{V}}}, \hat{\tau}^{\text{unadj}} + z_{1-\alpha/2}\cdot\sqrt{\hat{\mathbb{V}}}\right)$$

- Can be improved by a t-distribution

References

Fisher, R. A. (1925). Statistical Methods for Research Workers. Edinburgh, UK: Oliver and Boyd.

Fisher, R. A. (1935). The Design of Experiments (1st ed.), Edinburgh, UK: Oliver and Boyd.

Imbens, G. W. and Rubin, D. B. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge Univ. Press, New York.

Neyman J. (1990). On the application of probability theory to agricultural experiments: Essay on Principles, Section 9. Masters Thesis. Portions translated into English by D. Dabrowska and T. Speed (1990). Statistical Science. 1923(5): 465-472