Chapter 4 Neyman (1923)

1990 Polish → English

Statistical Science

2023 : 100 J. Causal Inference

CRE n_1 , n_0 固定 随机

$$\vec{Z} = (Z_1 \cdots Z_n) \sim \text{Uniform}$$

$$Pr\left(\vec{Z} = (z_1 \cdots z_n)\right) = \frac{1}{\binom{n}{n_1}}$$

记号 : $\left(Y_i(1), Y_i(0)\right)_{i=1}^{n}$ —— 固定

数据: $\left(Z_i, Y_i = Y_i(Z_i)\right)_{i=1}^{n}$

推断 $$\tau = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i(1) - Y_i(0)\right)$$

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

$$= \bar{Y}(1) - \bar{Y}(0)$$

方差记号: $$S^2(\underset{0}{1}) = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i(\underset{0}{1}) - \bar{Y}(\underset{0}{1})\right)^2$$

$$S^2(\tau) = \frac{1}{n-1}\sum_{i=1}^{n}(\tau_i - \tau)^2$$

τ_i → 个体 $Y_i(1) - Y_i(0)$

τ → 平均

Neyman 定理: CRE

1. $\hat{\tau} = \hat{\bar{Y}}(1) - \hat{\bar{Y}}(0)$ 是 τ 无偏估计

其中 $\hat{\bar{Y}}(1) = \frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i$

$\hat{\bar{Y}}(0) = \frac{1}{n}\sum_{i=1}^{n}(1-Z_i)Y_i$

2. $\mathrm{var}(\hat{\tau}) = \frac{S^2(1)}{n_1} + \frac{S^2(0)}{n_0} - \frac{S^2(\tau)}{n}$

$\overset{\text{等价}}{=} \frac{n_0}{n_1 n}S^2(1) + \frac{n_1}{n_0 n}S^2(0) + \frac{2}{n}S(1,0)$

$+$

Lemma $\quad 2S(1,0) = S^2(1) + S^2(0) - S^2(\tau)$

3. 方差的保守估计: 高估

$$\hat{V} = \frac{\hat{S}^2(1)}{n_1} + \frac{\hat{S}^2(0)}{n_0}$$

其中 $\hat{S}^2(1) = \frac{1}{n_1 - 1} \sum_{i=1}^{n} Z_i \left(Y_i - \hat{\bar{Y}}(1)\right)^2$

$\hat{S}^2(0) = \frac{1}{n_0 - 1} \sum_{i=1}^{n} (1 - Z_i) \left(Y_i - \hat{\bar{Y}}(0)\right)^2$

4. 置信区间

$$\hat{\tau} \pm 1.96 \sqrt{\hat{V}}$$

— $\hat{\tau}$ CLT

— $\hat{V} \Big/ \left(\frac{S^2(1)}{n_1} + \frac{S^2(0)}{n_0}\right) \xrightarrow{P} 1$

— 保守：以大于等于 95% 概率覆盖 τ

Design-based inference
Randomization-based inference
Finite-population inference

} 统计推断基础是随机化实验

证明 1. $\hat{\tau} = \frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n} (1-Z_i) Y_i(0)$

$$E(\hat{\tau}) = \cancel{\frac{1}{n_1}}\sum_{i=1}^{n} \frac{\cancel{n_1}}{n} Y_i(1) - \cancel{\frac{1}{n_0}}\sum_{i=1}^{n} \frac{\cancel{n_0}}{n} Y_i(0)$$

$$= \bar{Y}(1) - \bar{Y}(0) = \tau \quad \text{无偏}$$

2. $$\hat{\tau} = \sum_{i=1}^{n} \underbrace{Z_i}_{\text{随机，相关}} \underbrace{\left(\frac{Y_i(1)}{n_1} + \frac{Y_i(0)}{n_0}\right)}_{\text{固定}} + ⊛$$

$$\mathrm{var}(\hat{\tau}) = \ ?$$

抽样调查： 简单随机抽样 A3

SRS

$\{ C_1 \ \cdots \ C_n \}$ — 固定

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i$$

$\{ Z_1 \ \cdots \ Z_n \}$ 抽样指示函数

n_1 样本

$$\hat{C} = \frac{1}{n_1}\sum_{i=1}^{n} Z_i C_i$$

1. $E(\hat{c}) = \bar{c}$

fpc 有限样本修正

2. $\text{var}(\hat{c}) = \left(1 - \frac{n_1}{n}\right) \frac{S_c^2}{n_1} = \frac{n_0}{n n_1} S_c^2$

其中 $S_c^2 = \frac{1}{n-1} \sum_{i=1}^{n} (c_i - \bar{c})^2$

3. $\hat{S}_c^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n} Z_i (c_i - \hat{c})^2$

$E(\hat{S}_c^2) = S_c^2$ 无偏

$$\hat{\tau} = n_1 \frac{1}{n_1} \sum_{i=1}^{n} Z_i \underbrace{\left(\frac{Y_i(1)}{n_1} + \frac{Y_i(0)}{n_0}\right)}_{= c_i} + \circledast$$

$$\text{var}(\hat{\tau}) = n_1^2 \left(1 - \frac{n_1}{n}\right) \frac{S_c^2}{n_1}$$

$$= \frac{n_1 n_0}{n(n-1)} \sum_{i=1}^{n} \left(\frac{Y_i(1)}{n_1} + \frac{Y_i(0)}{n_0} - \frac{\bar{Y}(1)}{n_1} - \frac{\bar{Y}(0)}{n_0}\right)^2$$

$a_i + b_i$

$$= \frac{n_1 n_0}{n(n-1)}$$

$$\times \left[\frac{1}{n_1^2} \sum_{i=1}^{n} \left(Y_i(1) - \bar{Y}(1) \right)^2 + \frac{1}{n_0^2} \sum_{i=1}^{n} \left(Y_i(0) - \bar{Y}(0) \right)^2 + \frac{2}{n_1 n_0} \sum_{i=1}^{n} \left(Y_i(1) - \bar{Y}(1) \right) \left(Y_i(0) - \bar{Y}(0) \right) \right]$$

$$= \frac{\cancel{n_1} n_0}{n\, n_1^{\cancel{2}}} S_{(1)}^2 + \frac{n_1 \cancel{n_0}}{n\, n_0^{\cancel{2}}} S_{(0)}^2 + \frac{2\cancel{n_1 n_0}}{n\, \cancel{n_1 n_0}} S(1,0)$$

等价形式 2

3. 处理组 $\xleftarrow{SRS}$ $\{1 \cdots n\}$

$$E\left(\hat{S}_{(1)}^2\right) = S_{(1)}^2$$

对照组 $\xleftarrow{SRS}$ $\{1 \cdots n\}$

$$E\left(\hat{S}^2(0)\right) = S^2(0)$$

4. CLT 证明?

$$\underbrace{\hat{\tau} = \sum_{i=1}^{n} Z_i C_i}_{\text{随机}} + ⊛$$

Hoeffding combinatorial CLT

$A = \boxed{(a_{ij})}$

$\pi: \{1 \dots n\} \longrightarrow \{1 \dots n\}$

1-1 映射

$i \longmapsto \pi(i)$

$$\sum_{i=1}^{n} a_{i\pi(i)} \sim \text{CLT}$$

Charles Stein : Stein's method

被 Hoeffding CLT

启发

1972 Berkeley Symposium

实际中如何分析 RCT

$$\mathrm{lm}(Y_i \sim Z_i) \Rightarrow \mathrm{coef}(Z_i) = \hat{c}$$

$$\Downarrow$$

$$\hat{V}_{OLS} = \frac{\cancel{n}(\cancel{n_1-1})}{\cancel{(n-2)}\cancel{n_1}n_0}\hat{S}^2(1) + \frac{\cancel{n}(\cancel{n_0-1})}{\cancel{(n-2)}n_1\cancel{n_0}}\hat{S}^2(0)$$

$$\approx \frac{\hat{S}^2(1)}{n_0} + \frac{\hat{S}^2(0)}{n_1}$$

不对：异方差

异方差稳健方差

Eicker-Huber-White (EHW)

$Y_i \sim X_i$ 线性回归

$$\hat{\beta} = (X'X)^{-1}X'Y$$

$$\hat{\mathrm{cov}}(\beta) = \hat{\sigma}^2 (X'X)^{-1}, \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{n-p}$$

$$\hat{\mathrm{cov}}_{EHW}(\beta) = (X'X)^{-1}X'\hat{\Omega}X(X'X)^{-1}$$

其中 $\hat{\Omega} = \mathrm{Diag}(\hat{\varepsilon}_1^2 \cdots \hat{\varepsilon}_n^2)$

回归 CRE: $Y_i \sim 1 + Z_i$

$$\hat{V}_{EHW} = \frac{\hat{S}^2(1)}{n_1}\cancel{\frac{n_1-1}{n_1}} + \frac{\hat{S}^2(0)}{n_0}\cancel{\frac{n_0-1}{n_0}}$$

$$\approx \frac{\hat{S}^2(1)}{n_1} + \frac{\hat{S}^2(0)}{n_0}$$

Fisherian Principles for Experimental Design

— Randomization 随机化 : FRT, Neyman inference

— Replication 重复

Neyman : CLT $n_1 \to \infty$
$n_0 \to \infty$

FRT: P_{FRT} 可能不够小

例子: 人参 1 treatment
1 control

$$P_{FRT} = \frac{1}{2}$$

例子: Lady Tasting Tea

4+4 cups $P_{FRT} = \frac{1}{\binom{8}{4}} = 0.014$

3+3 cups $P_{FRT} \geq \frac{1}{\binom{6}{3}} = 0.05$

— Blocking 分区组 (农业实验术语)

实验 / 试验

Chapter 5 SRE

Stratified Randomized Experiment

strata

RBD

Randomized Block Design

blocks

discrete X 离散协变量

Box: Block what you can

and randomize what you can not.

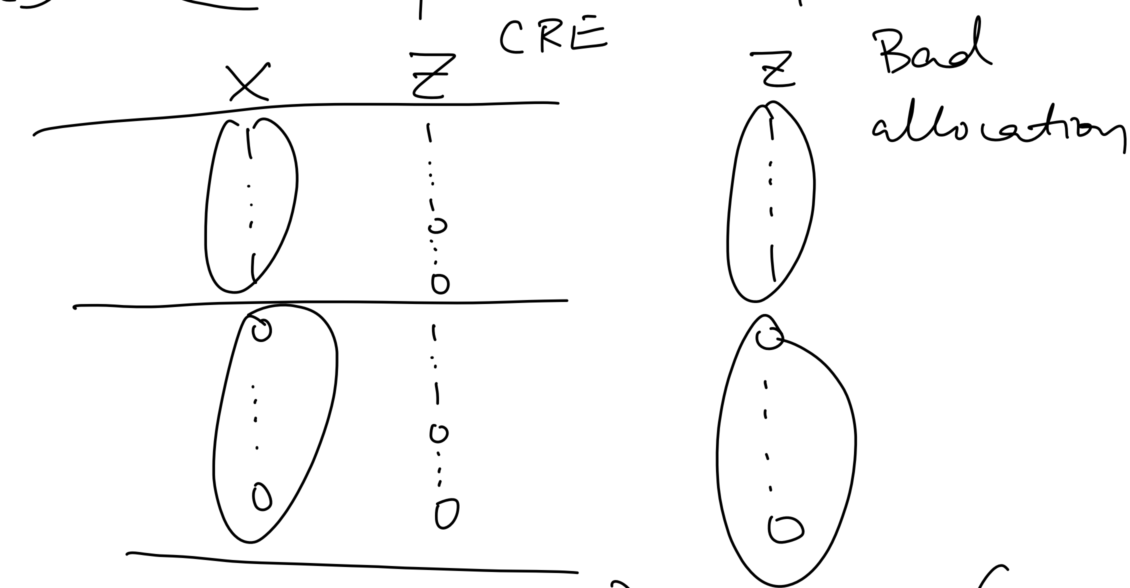
X
- 1 CRE
- 2 CRE
- ⋮
- K CRE

independent across strata

SRE

why SRE?

① 避免那些糟糕的随机化



$(n_{[k]1}, n_{[k]0})$: k层(处理, 对照)样本

$n_{[k]}$: k层一共样本

$\frac{n_{[k]1}}{n_{[k]}}$: k层处理频率

② 一些情况下，在SRE下估计了比在CRE下估计了更精确.

SRE：如何分析？

FRT：　$H_{0F}: Y_i(1) = Y_i(0),\ i = 1 \dots n$

$$T = T(\underbrace{\vec{z}}_{\text{随机}}, \underbrace{\vec{Y}, \vec{X}}_{\text{固定}})$$

$\Downarrow$

群论语言：　在 X 的 K 个层中，移定

子群　也随机置换

$\vec{X}$ 稳定子

也被称为 conditional randomization test

conditional permutation test

n 个经典选择：

eg1.　$$\hat{\tau}_S = \sum_{k=1}^{K} \underbrace{\pi_{[k]}}_{\text{加权平均}} \underbrace{\hat{\tau}_{[k]}}_{\text{分层无偏估计}}$$

e.g. 2

$$t_S = \frac{\hat{\tau}_S}{\sqrt{\hat{V}_S}}$$

其中

$$\hat{V}_S = \sum_{k=1}^{K} \pi_{[k]}^2 \left(\frac{\hat{S}^2_{[k]}(1)}{n_{[k]1}} + \frac{\hat{S}^2_{[k]}(0)}{n_{[k]0}} \right)$$

e.g. 3

$$W_S = \sum_{k=1}^{K} c_{[k]} W_{[k]}$$

其中

$$c_{[k]} = \frac{1}{n_{[k]1}\, n_{[k]0}}$$

或

$$c_{[k]} = \frac{1}{n_{[k]} + 1}$$

e.g. 4 Hodges – Lehmann (1962) aligned rank statistic

K 大时候：Wilcoxon $W_{[k]}$ 进行的比较不够

step1 $\tilde{Y}_i = Y_i - \bar{Y}_{[k]}$ if $X_i = k$

其中 $\bar{Y}_{[k]}$ = k 层 Y_i 均值

step2 $\mathrm{rank}(\tilde{Y}_1 \cdots \tilde{Y}_n)$

$\Rightarrow \tilde{R}_1 \cdots \tilde{R}_n$

step3 $\tilde{W} = \sum_{i=1}^{n} Z_i \tilde{R}_i$

KS k 层

eg5

$n_{[k]}$ 大

$$\begin{cases} D_S = \sum_{k=1}^{K} C_{[k]} D_{[k]} \\ \quad ? \\ D_{max} = \max_{1 \le k \le K} C_{[k]} D_{[k]} \\ \quad ? \end{cases}$$

$n_{[k]}$ 小, K 大

$$D = \max_y \left| \sum_{k=1}^{K} \pi_{[k]} \left(\hat{F}_{[k]1}(y) - \hat{F}_{[k]0}(y) \right) \right|$$

Neyman 估计

k层: CRE

$$\tau_{[k]} = \frac{1}{n_{[k]}} \sum_{i:X_i=k} \left(Y_i(1) - Y_i(0) \right)$$

$\hat{\tau}_{[k]}$ = 无偏估计

$\mathrm{var}\left(\hat{\tau}_{[k]}\right)$ = Neyman 公式

$$\hat{V}_{[k]} = \frac{\hat{S}^2_{[k]}(1)}{n_{[k]1}} + \frac{\hat{S}^2_{[k]}(0)}{n_{[k]0}}$$

总体: $$\hat{\tau}_S = \sum_{k=1}^{K} \underbrace{\pi_{[k]}}_{= \frac{n_{[k]}}{n}} \hat{\tau}_{[k]}$$

$$\hat{V}_S = \sum_{k=1}^{K} \pi_{[k]}^2 \hat{V}_{[k]}$$

CI: $\hat{\tau}_S \pm 1.96\sqrt{\hat{V}_S}$

$\hat{\tau}_S$: CLT ?

1. K 小, $n_{[k]}$ 大

$$\underbrace{\hat{\tau}_{[k]}\ \text{CLT}}$$

$$\Downarrow$$

$$\hat{\tau}_S\ \text{CLT}$$

2. K 大, $n_{[k]}$ 小,

$$\underbrace{\hat{\tau}_{[k]}\ \cancel{\text{CLT}}}$$

$$\Downarrow$$

$$\hat{\tau}_S\ \text{CLT}$$

因为 $\hat{\tau}_{[k]}$'s 独立