随机化实验：最后一次

Chapter 7 MPE

Matched-pairs Experiment

原因：① SRE 特例
有独特性

② 观察性研究：匹配估计

MPE = SRE
每一层两个个体
— 处理
— 对照

独特方差估计

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

$n_{[k]1} \geqslant 2$

$n_{[k]0} \geqslant 2$

公式不适用

我们记号 $(i, j) \longrightarrow ij$

i → 对 $i=1 \cdots n$

j → 个体 $j=1,2$

$Y_{ij}(1)$, $Y_{ij}(0)$

对么何而来？

试验设计

优化问题

$$Z_i = \begin{cases} 1 & (i,1) \text{ 处理} \\ 0 & (i,2) \text{ 处理} \end{cases}$$

$$\text{MPE} = Z_i\text{'s} \underset{\sim}{\text{iid}} \text{Bern}\left(\frac{1}{2}\right)$$

观察到
$$\begin{cases} Y_{i1} = Z_i Y_{i1}(1) + (1-Z_i) Y_{i1}(0) \\ Y_{i2} = Z_i Y_{i2}(0) + (1-Z_i) Y_{i2}(1) \end{cases}$$

数据：$\left(Z_i, Y_{i1}, Y_{i2}\right)_{i=1}^{n}$

一般 MPE:

	Y_i(处理)	Y_i(对照)
$i=1$		
2		
$\vdots$	$\vdots$	
n		

```
> library("HistData")
> ZeaMays
   pair pot  cross   self  diff
1     1   1 23.500 17.375 6.125
2     2   1 12.000 20.375 -8.375
3     3   1 21.000 20.000 1.000
4     4   2 22.000 20.000 2.000
5     5   2 19.125 18.375 0.750
6     6   2 21.500 18.625 2.875
7     7   3 22.125 18.625 3.500
8     8   3 20.375 15.250 5.125
9     9   3 18.250 16.500 1.750
10   10   3 21.625 18.000 3.625
11   11   3 23.250 16.250 7.000
12   12   4 21.000 18.000 3.000
13   13   4 22.125 12.750 9.375
14   14   4 23.000 15.500 7.500
15   15   4 12.000 18.000 -6.000
```

Fisher (1935)

DoE

FRT 用于

Darwin 数据

$\tau_i = Y_i(\text{处理}) - Y_i(\text{对照})$

MPE — FRT : H_{0F}: $Y_{ij}(1) = Y_{ij}(0)$ $\forall i\ \forall j$

$T = T(\hat{\tau}_i\text{'s})$

permutation test

= permute $\hat{\tau}_i$'s 符号

2^n 种

Neymanian:

$E(\hat{\tau}_i) = \tau_i$ SRE特例

$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\hat{\tau}_i$, $E(\hat{\tau}) = \tau$

$Var(\hat{\tau}) =$ HW

估计方差? SRE公式不可用

记号:

$$\tau_i = \frac{1}{2}\sum_{j=1}^{2}\left(Y_{ij}(1) - Y_{ij}(0)\right)$$

$$\tau = \frac{1}{n}\sum_{i=1}^{n} \tau_i$$

$$\hat{\tau}_i = (2Z_i - 1)(Y_{i1} - Y_{i2})$$

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \hat{\tau}_i \xrightarrow{\text{无偏}} \tau$$

$$\mathrm{Var}(\hat{\tau}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}(\hat{\tau}_i)$$

定义 $$\hat{V} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\hat{\tau}_i - \hat{\tau}\right)^2$$

<u>定理</u> MPE 下 $\hat{V}$ 保守估计了 $\mathrm{var}(\hat{\tau})$:

$$E(\hat{V}) - \mathrm{var}(\hat{\tau}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\tau_i - \tau)^2 \geq 0$$

证明: $E\left(n(n-1)\hat{V}\right)$

$$\overset{\text{定义}}{=} E\left(\sum_{i=1}^{n}\left(\hat{\tau}_i-\hat{\tau}\right)^2\right)$$

$$\overset{\text{代数}}{=} E\left(\sum_{i=1}^{n}\hat{\tau}_i^2 - n\hat{\tau}^2\right)$$

$$\overset{E(\cdot)\text{线性}}{=} \sum_{i=1}^{n}E\left(\hat{\tau}_i^2\right) - nE\left(\hat{\tau}^2\right)$$

$$\overset{\text{方差定义}}{=} \sum_{i=1}^{n}\left(\mathrm{var}(\hat{\tau}_i)+\tau_i^2\right) - n\left(\mathrm{var}(\hat{\tau})+\tau^2\right)$$

$$\overset{\text{重组}}{=} \underbrace{\sum_{i=1}^{n}\mathrm{var}(\hat{\tau}_i)}_{\because\, n^2\,\mathrm{var}(\hat{\tau})} - n\,\mathrm{var}(\hat{\tau}) + \left(\sum_{i=1}^{n}\tau_i^2 - n\tau^2\right)$$

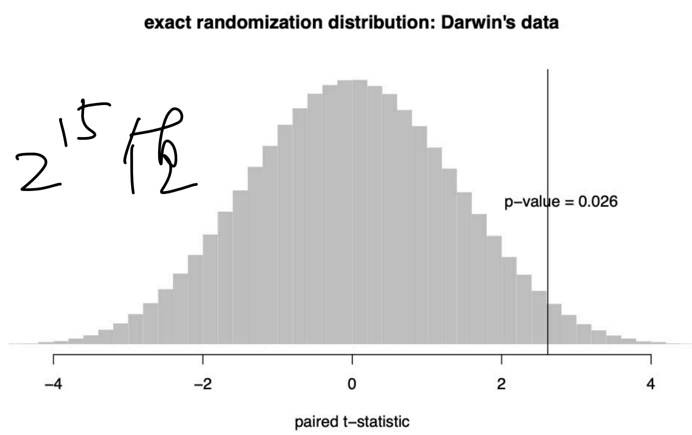
$$\overset{\text{代数}}{=} n(n-1)\,\mathrm{var}(\hat{\tau}) + \sum_{i=1}^{n}\left(\tau_i-\tau\right)^2$$

$\square$

MPE:
$$\begin{cases} \hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \hat{\tau}_i \\ \hat{V} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\hat{\tau}_i - \hat{\tau}\right)^2 \end{cases}$$

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

```
> library("HistData")
> ZeaMays
   pair pot  cross   self   diff
1     1   1 23.500 17.375  6.125
2     2   1 12.000 20.375 -8.375
3     3   1 21.000 20.000  1.000
4     4   2 22.000 20.000  2.000
5     5   2 19.125 18.375  0.750
6     6   2 21.500 18.625  2.875
7     7   3 22.125 18.625  3.500
8     8   3 20.375 15.250  5.125
9     9   3 18.250 16.500  1.750
10   10   3 21.625 18.000  3.625
11   11   3 23.250 16.250  7.000
12   12   4 21.000 18.000  3.000
13   13   4 22.125 12.750  9.375
14   14   4 23.000 15.500  7.500
15   15   4 12.000 18.000 -6.000
```



2^{15} 個

FIGURE 7.1: Randomization distribution of $\hat{\tau}$ using Darwin's data

Darwin

玉米数据

FRT $\hat{\tau}$

HW 7.6 $(\hat{\tau}, \hat{V})$

Chapter 8

Fisherian
Neymanian
统一

H_{0F} : $Y_i(1) = Y_i(0) \quad \forall i$

估计 H_{0N} : $\tau = 0$

Fisher: H_{0F}, 任何统计量都对

p_{FRT} 精确

无需渐近理论

Neyman: H_{0N}

估计 τ

渐近理论 $n \to \infty$

CLT

P_{FRT} 是然依赖 H_{0F}

但是在 H_{0F} 有精确 p 值

也是最低要求

$\Rightarrow$ 只考虑 FRT

但是 H_{0F} 似乎很无聊

万一也想检验 $H_{0N}: \tau = 0$?

推荐：用 FRT + 好统计量

$$\Longrightarrow \begin{cases} H_{0F}: & P_{FRT} \text{ 精确} \\ H_{0N}: & P_{FRT} \text{ 渐近 } \checkmark \end{cases}$$

FRT + t

CRE : $t = \dfrac{\hat{\tau}}{\sqrt{\hat{V}}}$

$$= \underbrace{\frac{\hat{\tau}}{\sqrt{Var(\hat{\tau})}}}_{\substack{\text{CLT} \ \downarrow d \\ N(0,1)}} \underbrace{\sqrt{\frac{Var(\hat{\tau})}{\hat{V}}}}_{\substack{\downarrow P \\ C \le 1}}$$

H_{0N}: $\tau = 0$

$$\xrightarrow[d]{H_{0N}} C \cdot N(0,1)$$

理论分布

FRT 假设 $(Y_i , Y_i)_{i=1}^{n}$ 为潜在结果 (H_{0F})

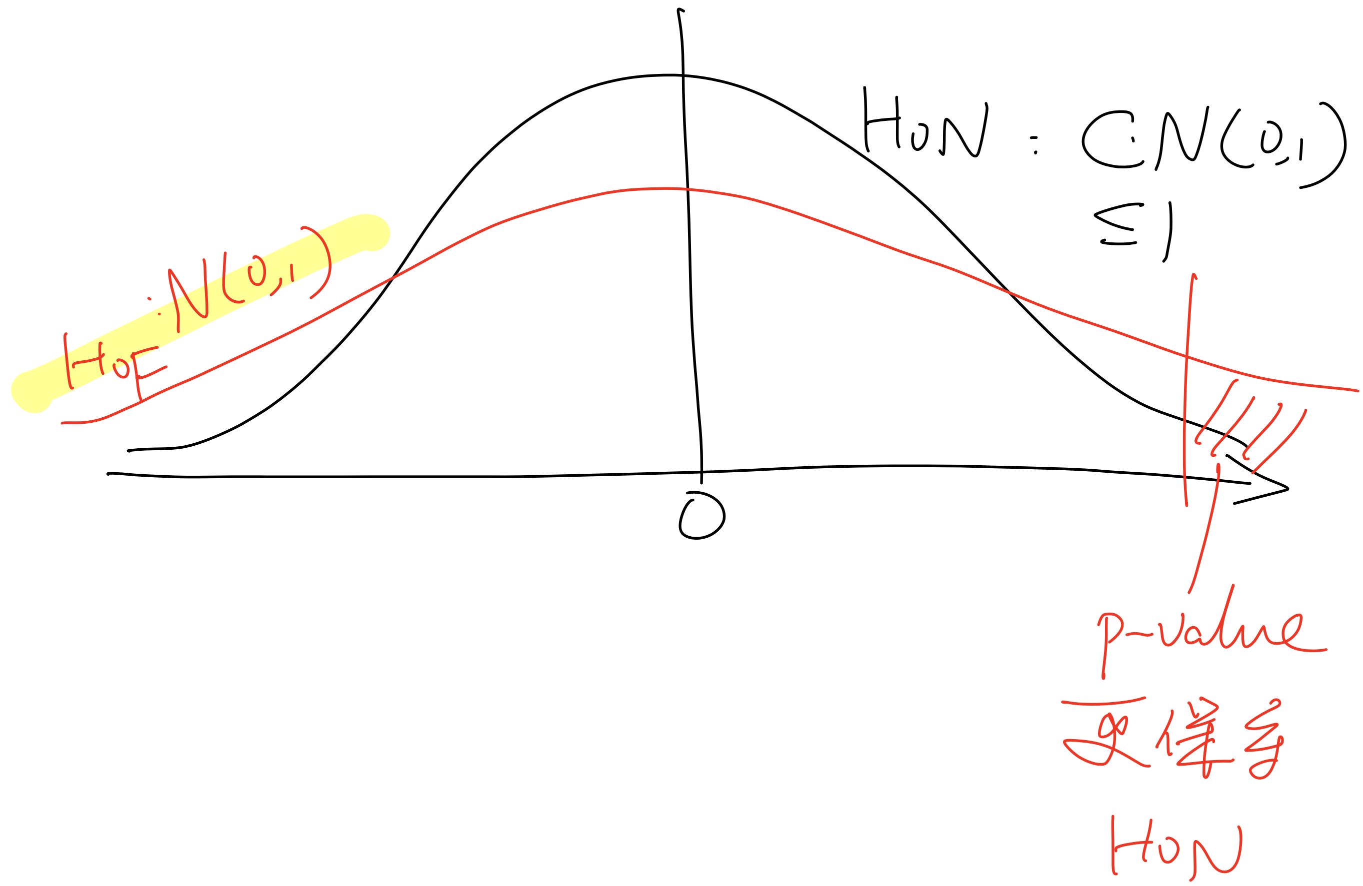
FRT 产生了一个可计算分布

$$(t)^{\pi} = \left(\frac{\hat{\tau}}{\sqrt{\hat{V}}}\right)^{\pi}$$

$$\xrightarrow{d} N(0,1)$$

$$\downarrow$$

$C=1$ 因为不保守



推广：

① 协变量

Lin: $Y_i \sim 1 + Z_i + Z_i \cdot X_i$

$\downarrow$

t-stat

EHW

$$t_L = \frac{\hat{\tau}_L}{\sqrt{\hat{V}_L}}$$

② SRE

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

实验统计术语: studentization

t-distribution

bootstrap 理论

高阶性质 P. Hall

Chapter 9

Bridging finite and super population causal inference

桥梁性

Part II: 实验

finite population

Part III: 观测性研究

及之后

super population

我的选择: Chapters 2—8

无法抛弃这套古典内容:

① 历史：想法来自 Neyman
Fisher

古典统计

抽样调查 —— 依赖 design-based

On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection
Author(s): Jerzy Neyman
Reviewed work(s):
Source: *Journal of the Royal Statistical Society*, Vol. 97, No. 4 (1934), pp. 558-625
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2342192

Cochran 书
Fuller

实验设计 —— 没有广泛继承

Statistical Problems in Agricultural Experimentation
Author(s): J. Neyman, K. Iwaszkiewicz and St. Kolodziejczyk
Source: *Supplement to the Journal of the Royal Statistical Society*, 1935, Vol. 2, No. 2 (1935), pp. 107-180
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.com/stable/2983637

② 个人：PhD 训练

最先看到这套

Imbens & Rubin book

D. B. Rubin —— W. G. Cochran

Fisher, Yates

Rothamsted statistics

③ 什么叫统计？

统计推断的基础是什么？

随机性从何而来？

来自可控的试验设计

对这套理论的批评：

internal validity ✓

external validity ?

一 回应：

sampling process

大总体

↓

有限样本

↓

实验

Design-based inference

两层随机性

Chapter 9 内容：也许 15 分钟

CRE: $\left(Z_i, Y_i(1), Y_i(0)\right)_{i=1}^{n} \overset{iid}{\sim} \left(Z, Y(1), Y(0)\right)$

随机化假设：

Rubin (1978 AoS)

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

: strong ignorability 可忽略性

$$\Rightarrow \tau = E\left(Y(1) - Y(0)\right)$$

记号

相同

$$= E\left(Y(1)\right) - E\left(Y(0)\right)$$

随机化 $\Rightarrow$ $= E\left(Y(1) \mid Z=1\right) - E\left(Y(0) \mid Z=0\right)$

$$= E\left(Y \mid Z=1\right) - E\left(Y \mid Z=0\right)$$

SRE: X_i 离散 $\{1 \cdots K\}$

$$\left(X_i, Z_i, Y_i(1), Y_i(0)\right)_{i=1}^{n} \overset{iid}{\sim} \left(X, Z, Y(1), Y(0)\right)$$

$$Z \perp\!\!\!\perp \left(Y(1), Y(0)\right) \mid X = k \quad \forall k$$

$$\Rightarrow \tau_k = E\left(Y(1) - Y(0) \mid X = k\right)$$

$$= E\left(Y \mid Z=1, X=k\right) - E\left(Y \mid Z=0, X=k\right)$$

$$\Rightarrow \tau = E\left(Y(1) - Y(0)\right)$$

$$= \sum_{k=1}^{K} \Pr(X=k)\, \tau_k$$

核心公式已经出现!

$$= \sum_{k=1}^{K} \Pr(X=k) \cdot \left[E(Y \mid Z=1, X=k) - E(Y \mid Z=0, X=k) \right]$$