# 定量社会科学的因果推断

## Causal Inference in Quantitative Social Sciences

江　艇
中国人民大学经济学院

Last updated: May 10, 2020

## Lecture 3　潜在结果分析框架

- 潜在结果分析框架 (potential outcomes framework) 对因果推断问题进行规范化表述的语言。可以追溯到统计学先驱 Neyman, Pearson, Fisher，但主要由统计学家 Donald Rubin 在 1974-1980 年的一系列研究所奠定的，因此也称为鲁宾因果模型 (Rubin causal model)。

- 定义处理效应（因果效应）。

$$
\begin{aligned}
Y_i^0 &= \text{不对个体 } i \text{ 实施某种处理的潜在结果} \\
Y_i^1 &= \text{对个体 } i \text{ 实施某种处理的潜在结果}
\end{aligned}
$$

则个体处理效应为

$$\tau_i \triangleq Y_i^1 - Y_i^0$$

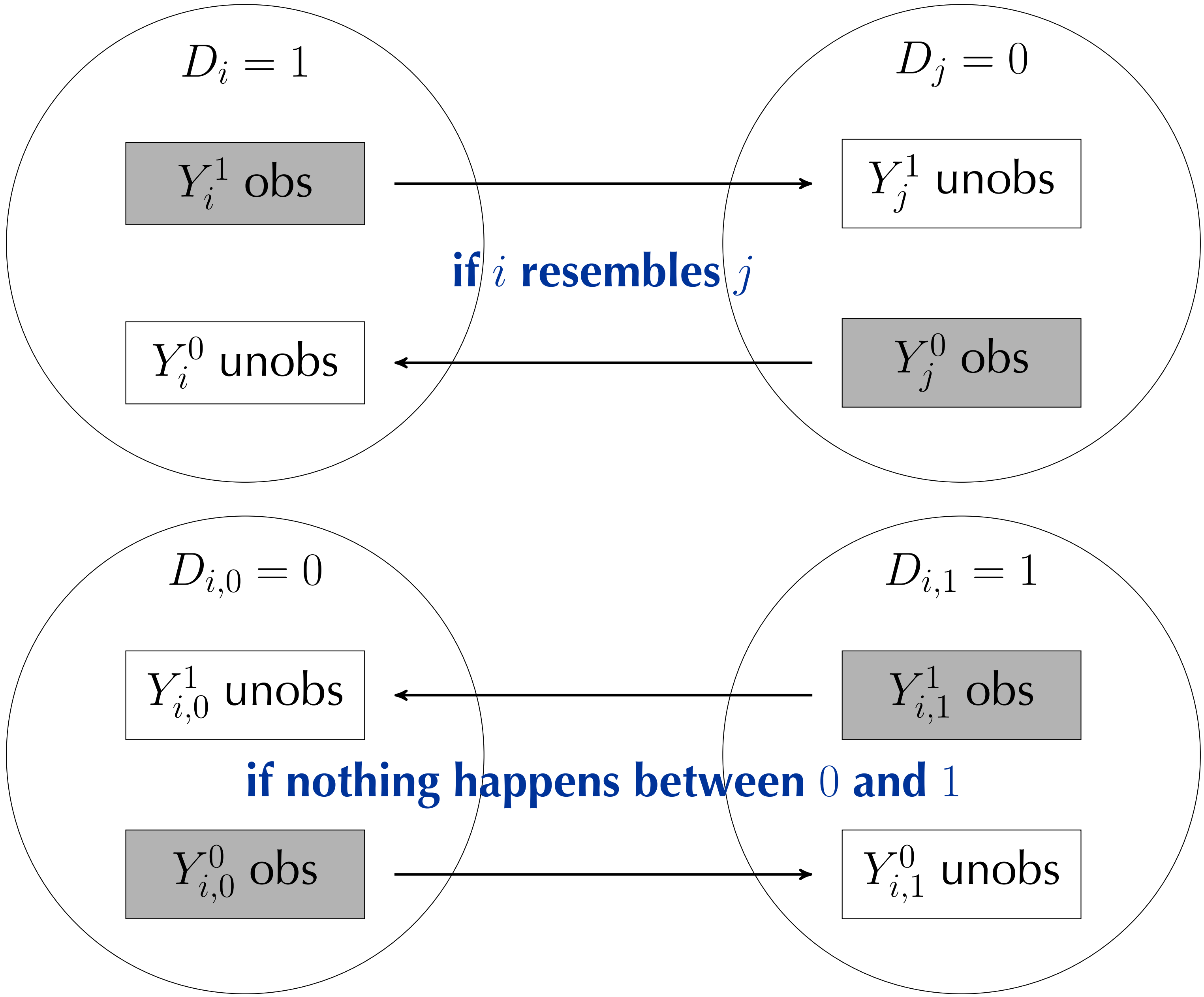
- 观测到的处理状态

$$
D_i = \begin{cases} 1 & \text{个体 } i \text{ 接受处理，进入处理组} \\ 0 & \text{个体 } i \text{ 不接受处理，进入控制组} \end{cases}
$$

观测到的结果

$$
\begin{aligned}
Y_i &= \begin{cases} Y_i^1 & D_i = 1 \\ Y_i^0 & D_i = 0 \end{cases} \\
&= (1 - D_i) \cdot Y_i^0 + D_i \cdot Y_i^1
\end{aligned}
$$

- 关于这一定义的几点评论：
  - 因果效应的定义只取决于潜在结果，而不取决于究竟哪个潜在结果被实际观测到（实际发生）。
  - 因果效应永远是同一个体在同一时点上的不同结果之间的比较。
  - 因果推断的基本难题就是数据缺失：我们永远只能至多观察到一个潜在结果。没有观察到的那个潜在结果叫做反事实 (counterfactual)。
  - 因果分析的关键就是构造反事实，也就是个插值问题。而且反事实只能从实际观测到的、处理状态相反的其他结果中去寻找，要么是其他个体的潜在结果，要么是同一个体其他时期的潜在结果。
  - 我们实际衡量因果效应时必须涉及多个个体或同一个体的不同时期，但无论是横截面比较还是事前事后比较，都不是因果效应的定义，尽管它们对于衡量因果效应很重要。



$D_i = 1$
$Y_i^1$ obs
$Y_i^0$ unobs
$D_j = 0$
$Y_j^1$ unobs
$Y_j^0$ obs
if $i$ resembles $j$
$D_{i,0} = 0$
$Y_{i,0}^1$ unobs
$Y_{i,0}^0$ obs
$D_{i,1} = 1$
$Y_{i,1}^1$ obs
$Y_{i,1}^0$ unobs
if nothing happens between $0$ and $1$

**Stable unit treatment value assumption (SUTVA).** 每个个体的潜在结果不会因为其他个体接受的处理水平的不同而不同；每个个体所接受到的每个处理都只会产生一种潜在结果。

这代表两层含义：

- No interference. 不存在同侪效应 (peer effects)、溢出效应 (spill-over effects)、一般均衡效应 (general equilibrium effects).
  解决办法：重新定义观测单位。
- No hidden variations in treatment. 实施 treatment 的方式不重要（不存在 Hawthorne effect, carry-over effect 等）。
  解决办法：重新定义处理。

- SUTVA 很关键，但很难检验。

**示例 7.** 排污权交易项目的减排效应 (Fowlie *et al*, 2012, *AER*)

| Control group | Levels | Logs | RECLAIM facilities | Controls |
|---|---|---|---|---|
| *Panel A. Change in $NO_x$ emissions between periods 1 and 4* | | | | |
| Base specification | –20.59*** | –0.25*** | 212 | 1,222 |
| | (7.63) | (0.09) | | |
| Exclude L.A. facilities | –23.50*** | –0.34*** | 210 | 778 |
| | (7.96) | (0.09) | | |
| Exclude northern CA | –26.60*** | –0.23** | 210 | 767 |
| | (7.58) | (0.11) | | |
| Severe nonattainment only | –21.65** | –0.29** | 208 | 475 |
| | (7.89) | (0.11) | | |
| Single facility only | –19.92** | –0.23** | 210 | 781 |
| | (7.60) | (0.10) | | |

- 逻辑似乎是：

因为：SUTVA 不成立 $\Rightarrow$ 事件 A 发生
所以：事件 A 未发生 $\Rightarrow$ SUTVA 成立

- 实际上是：

因为：事件 A 发生 $\Rightarrow$ SUTVA 不成立
所以：事件 A 未发生 $\Rightarrow$ SUTVA 成立

- 因为我们并不能穷尽 SUTVA 不成立时可能发生的所有事件，当 SUTVA 不成立时，可能是事件 A 发生，也可能事件 A 不发生但事件 B 发生。所以对 SUTVA 的检验是一种“间接检验 (indirect test)”。

## **示例 8.** 治疗传染病提高教育参与率 (Miguel and Kremer, 2004, *ECMA*)

- 肠道蠕虫病是一种影响了至少四分之一世界人口的消化道传染病。作者在 75 家肯尼亚小学按随机先后顺序发放治疗肠道蠕虫的药物，降低旷课至少一个季度。
- 以往研究都在个体层面进行随机处理，潜在的外部性倾向于低估处理效应。该项研究在学校层面进行随机处理，并且识别了校际外部性。
- 第 1 组 1998、1999 年接受处理，第 2 组 1999 年接受处理，第 3 组 2001 年接受处理。因此，1998 年第 1 组为处理组，第 2、3 组为控制组；1999 年第 1、2 组为处理组，第 3 组为控制组。

- 识别策略：[1]

$$
\begin{aligned}
Y_{ijt} = \alpha + \beta_1 \cdot T_{1it} + \beta_2 \cdot T_{2it} + \mathbf{X}'_{ijt}\boldsymbol{\delta} \\
+ \sum_d \left(\gamma_d \cdot N^T_{dit}\right) + \sum_d \left(\phi_d \cdot N_{dit}\right) + u_i + e_{ijt}
\end{aligned}
$$

其中 $i$ 表示学校，$j$ 表示学生，$N_{dit}$ 表示距离 $d$ 之内的所有学校学生总数，$N^T_{dit}$ 表示距离 $d$ 之内的所有接受处理学校学生总数。

- 给定 $N_{dit}$，$N^T_{dit}$ 是外生的。$\phi_d$ 为邻近学校密度的直接效应，$\gamma_d$ 为校际外部性。第 1 年的平均处理效应为 $\beta_1 + \sum_d \left(\gamma_d \bar{N}^T_{dit}\right)$，第 2 年的平均处理效应为 $\beta_2 + \sum_d \left(\gamma_d \bar{N}^T_{dit}\right)$。$\beta_1$ 和 $\beta_2$ 为直接处理效应和校内外部性之和。

[1] 随机处理之下为何还要控制 **X**？此处加入控制变量，并不是为了矫正选择性，纯粹是因为 **X** 中包含的学生特征和学校特征是因变量的显然的影响因素，因此将其从扰动项中剥离，明确控制起来，可以提高回归的整体拟合程度，提高估计效率，减少核心解释变量系数估计的标准误，提高统计显著性。因此，即使是随机实验研究，通常也会尽可能控制实验被试的个体特征，尽管这些特征在处理组和控制组是平衡的，不控制也不会影响因果效应的识别。

# 被估计量

- 平均处理效应 (average treatment effect, ATE)

$$\tau \triangleq \mathbb{E}\left(Y_i^1 - Y_i^0\right)$$

- 处理组的平均处理效应 (average treatment effect on the treated, ATT)

$$\tau_1 \triangleq \mathbb{E}\left(Y_i^1 - Y_i^0 | D_i = 1\right)$$

- 控制组的平均处理效应 (average treatment effect on the untreated, ATU)

$$\tau_0 \triangleq \mathbb{E}\left(Y_i^1 - Y_i^0 | D_i = 0\right)$$

- 除非随机分组，一般来说以上三者并不相等。
- 三者存在如下关系：

$$\tau = \tau_1 \cdot \mathsf{Pr}(D_i = 1) + \tau_0 \cdot \mathsf{Pr}(D_i = 0)$$

- 定义在协变量上的条件平均处理效应 (Conditional ATE, CATE)

$$\tau(x) \triangleq \mathbb{E}\big(Y_i^1 - Y_i^0 | X_i = x\big)$$
$$\tau_1(x) \triangleq \mathbb{E}\big(Y_i^1 - Y_i^0 | X_i = x, D_i = 1\big)$$
$$\tau_0(x) \triangleq \mathbb{E}\big(Y_i^1 - Y_i^0 | X_i = x, D_i = 0\big)$$

根据期望迭代定律,

$$\tau = \mathbb{E}\left\{\mathbb{E}\big(Y_i^1 - Y_i^0 | X_i = x\big)\right\} = \int \tau(x) dF_X$$
$$\tau_1 = \mathbb{E}\left\{\mathbb{E}\big(Y_i^1 - Y_i^0 | X_i = x, D_i = 1\big)\right\} = \int \tau_1(x) dF_{X|D=1}$$
$$\tau_0 = \mathbb{E}\left\{\mathbb{E}\big(Y_i^1 - Y_i^0 | X_i = x, D_i = 0\big)\right\} = \int \tau_0(x) dF_{X|D=0}$$

# 第一类识别假设

- 显然，分配机制对于平均处理效应的识别至关重要。下面这个例子说明，不对分配机制进行正式建模，可能使得平均处理效应无法被正确识别。

| Unit | Treatment | Observed Outcome |
|---|---|---|
| Patient #1 | Surgery | 7 |
| Patient #2 | Drug | 6 |
| Patient #3 | Surgery | 5 |
| Patient #4 | Drug | 8 |
| Average Difference | | -1 |

| Unit | Potential Outcomes | | Causal Effects |
|---|---|---|---|
| | Surgery | Drug | S-D |
| Patient #1 | 7 | 1 | 6 |
| Patient #2 | 5 | 6 | -1 |
| Patient #3 | 5 | 1 | 4 |
| Patient #4 | 7 | 8 | -1 |
| Average Effect | | | 2 |

- 因此，因果识别的关键假设往往就是关于分配机制的假设。
- 在潜在结果分析框架下，分配机制（倾向得分）可以表示为

$$\pi \triangleq \Pr\left(D = 1 | X, Y^0, Y^1\right)$$

比如在前面的例子中,

$$\pi = \mathbb{1}\left(Y^1 - Y^0 > 0\right)$$

- 第一类识别假设：分配机制不取决于潜在结果和可观测变量。

$$\Pr(D = 1 | X, Y^0, Y^1) = \Pr(D = 1)$$

- 第一类识别假设（重新表述）：潜在结果均值独立于处理状态。

**Assumption ID.1:** $\mathbb{E}\left(Y^d | D\right) = \mathbb{E}\left(Y^d\right),\ d = 0, 1$

等价地,

$$\mathbb{E}(Y^0 | D = 1) = \mathbb{E}(Y^0 | D = 0)$$
$$\mathbb{E}(Y^1 | D = 1) = \mathbb{E}(Y^1 | D = 0)$$

- 识别：此时**组间均值差异**能够直接识别 ATE ($\tau$), ATT ($\tau_1$) 和 ATU ($\tau_0$).

$$
\begin{aligned}
&\mathbb{E}(Y|D=1)-\mathbb{E}(Y|D=0)\\
&=\mathbb{E}\big(Y^1|D=1\big)-\mathbb{E}\big(Y^0|D=0\big)\\
&=\mathbb{E}\big(Y^1\big)-\mathbb{E}\big(Y^0\big)\\
&=\mathbb{E}\big(Y^1-Y^0\big)\\
&=\tau
\end{aligned}
$$

要想识别 $\tau_1$，需要用到 $\mathbb{E}\big(Y^0|D\big)=\mathbb{E}\big(Y^0\big)$.
要想识别 $\tau_0$，需要用到 $\mathbb{E}\big(Y^1|D\big)=\mathbb{E}\big(Y^1\big)$.
证明过程类似。因此有

$$\tau=\tau_1=\tau_0$$

- 估计：样本组间均值差异是总体组间均值差异的一致估计。

$$\bar{Y}_{D=1}-\bar{Y}_{D=0}\to_p \mathbb{E}(Y|D=1)-\mathbb{E}(Y|D=0)$$

- 第一类识别假设就是随机实验 (也称随机控制试验, randomized controlled trial, RCT) 的识别假设。

- 第一类识别假设应用举例
  - 实验室实验。

    **示例 9.** 禀赋效应是否存在 (Kahneman *et al*, 1990, *JPE*)？

    一旦一个人拥有一件物品，这件物品就会突然变得更值钱。将实验者随机分成两组，其中一组每人得到一个杯子，让两组人对杯子进行估值，得到杯子的人平均而言愿意以 7 元的价格卖出杯子，而没有杯子的人平均而言只愿意以 3.5 元的价格买入杯子。
  - 实地实验：前面提到的 Carter *et al* (2017) 和 Miguel and Kremer (2004) 均属此列。

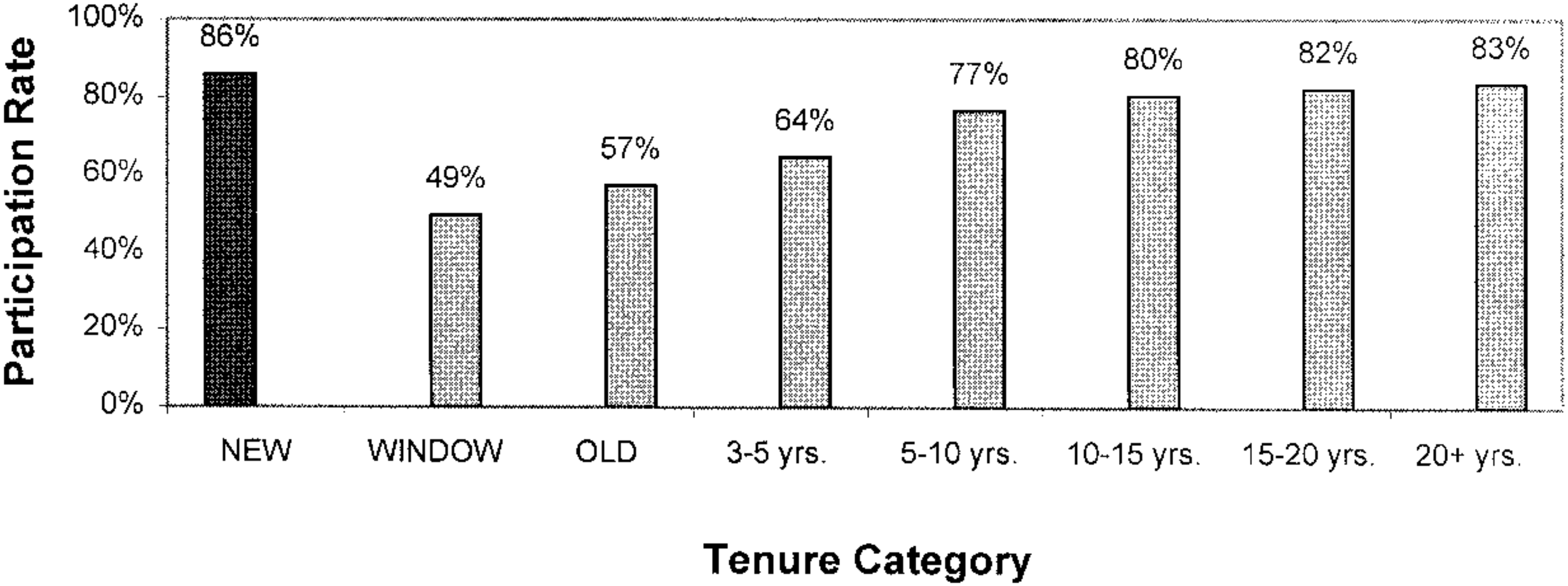
    **示例 10.** 在家工作是否更有效率 (Bloom *et al*, 2015, *QJE*)？

    携程将客服中心的一部分职员随机分配在家工作，相对于在办公室工作，在家工作提升了 13% 工作绩效，其中 9% 来自更长的工作时长（更少的休假和病假），4% 来自更高的工作效率（每分钟接听更多电话，因为在家的工作环境更安静）。

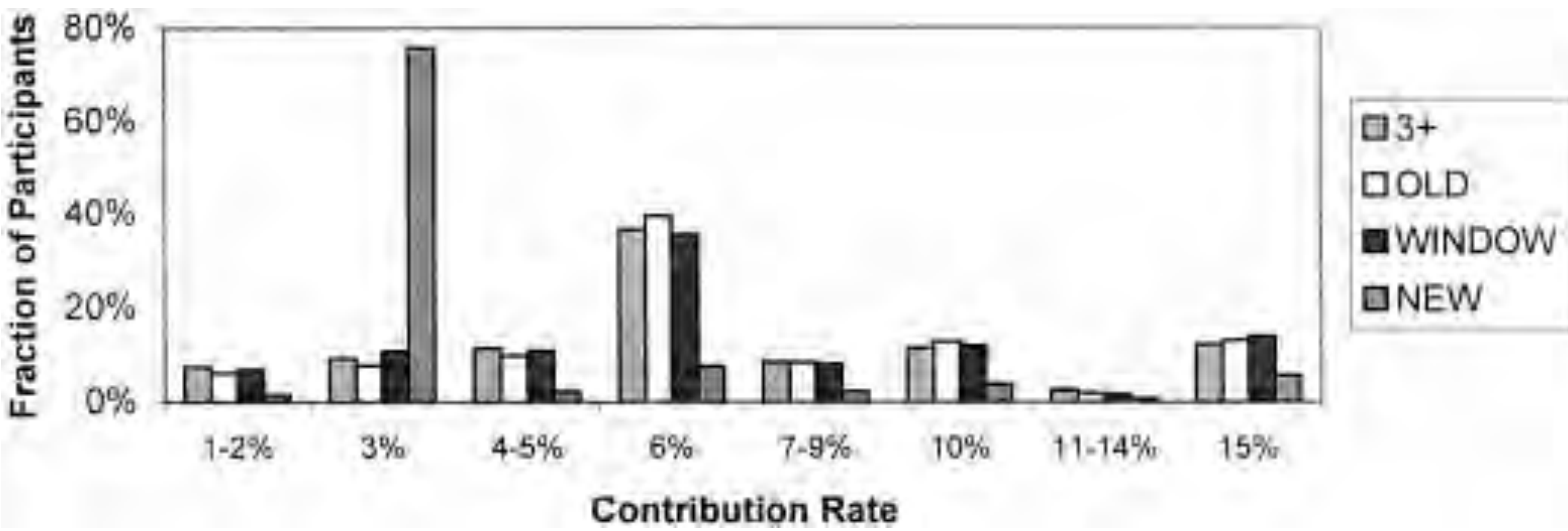
- 自然实验 (natural experiment)：在观测性数据情境中，某种特定的社会或政治过程使得某项原因或某种干预以随机或近似随机的方式分配在某些观测单位上（这里的观测单位可能是个体、企业、地区甚至国家），从而这一观测情境很像真正的实验，受到原因影响的观测单位和未受到原因影响的观测单位之间的简单比较就能揭示因果效应。

**示例 11.** 退休金计划表格设计的奥妙 (Madrian and Shea, 2001, *QJE*)

美国一家大企业在 1998 年 4 月 1 日修改了 401(k) 计划申请表的格式，此前的默认选项为"不加入"，此后的默认选项为"加入"且默认缴纳率设定为 3%。此举大幅提高了退休金计划的参与率，且大多数人的缴纳率为默认值。



Participation Rate
100%
80%
60%
40%
20%
0%
86%
49%
57%
64%
77%
80%
82%
83%
NEW
WINDOW
OLD
3-5 yrs.
5-10 yrs.
10-15 yrs.
15-20 yrs.
20+ yrs.
Tenure Category

Fraction of Participants
80%
60%
40%
20%
0%
1-2%
3%
4-5%
6%
7-9%
10%
11-14%
15%
Contribution Rate
3+
OLD
WINDOW
NEW

**示例 12.** 文化分歧会不会带来政治分歧 (Posner, 2004, *APSR*)？
通布卡族和切瓦族是赞比亚-马拉维边境上具有鲜明文化差异的两个民族。两国的边境是在殖民时期根据山势地形划分的，创造了一种近似实验情境，将这两个民族均匀分到了两个国家，每个国家中两个民族的分布很相似，但在马拉维，这两个民族占全国人口的多数，而在赞比亚，这两个民族占全国人口的少数。问卷调查显示，在马拉维，两个民族对彼此的政治态度比较敌对，而在赞比亚，两个民族的政治态度比较友善合作。这说明文化分歧是否会带来政治分歧取决于群体规模是否足够大（文化分歧成为政治动员的工具）。

# 自然实验、准实验与准自然实验

- 学术界目前对这几个术语的用法比较混乱。一个自恰的理解框架是：
  1. **实验** 是指对实验对象随机分组；实验的研究方法是组间比较。
  2. **准实验 (quasi-experiment)** 是指随机分组存在缺陷，比如，某些协变量（出于偶然）在组间不平衡；或，是否实际接受干预存在一定的选择性；或，部分被试在实验过程中退出。
  3. **自然实验** 是指本身并不是实验，但自然力量或某些不可控制、未经计划的事件决定了干预的实施。自然实验是对实验的类比，在这种理解下，自然实验的研究方法也是组间比较。
  4. **准自然实验 (quasi-natural experiment)** 是指本身并不是实验，而且干预的实施或处理组-控制组的划分也不是随机的，而是受到某些社会政治经济因素的系统性影响。准自然实验是对准实验的类比，在这种理解下，准自然实验的研究方法包括匹配、双重差分、工具变量、断点回归等。

- 我的立场：
  - “自然”这个限定词的含义是用来说明研究者在数据生成过程中所扮演的角色，究竟是主动参与了数据的生成，还是作为数据的观测者。
  - 在上述 **1-4** 的理解框架下,“准实验”的范畴是很狭窄的，或者说，大家一般不在 **2** 的意义下理解“准实验”，不管是完全随机分组，还是分层随机分组，还是存在不完美遵守/服从 (imperfect compliance) 或实验个体中途退出 (attrition) 的实验，大家都倾向于称之为“实验”而非“准实验”，只要实验数据的生成是研究者主动参与的结果。
  - 因此我认为可以在 **3** 的意义下使用“自然实验”这个词，而在 **4** 的意义下使用“准实验”这个词，**1-2** 都可以被归作“实验”。

- 在这种理解下，一方面,“准自然实验”有重复累赘之嫌，既然是“准”，那么天然就是“自然”的。而另一方面，目前对“自然实验”的使用又过于宽泛了，适用匹配、双重差分、工具变量、断点回归等方法的观测性研究应该叫做“准实验”而非“自然实验”。
- 另一种折衷的理解方式是，所有观测性研究都可以叫做“自然实验”，其内部又可以划分成随机自然实验 (randomized natural experiment) 和准自然实验。毕竟所有的方法，究其本质来看，都是通过各种各样的调整，最后试图逼近“（自然）实验”。过于纠结术语的用法并无必要。

## 重新理解线性结构模型

- 考察如下的结构模型，其中核心解释变量 $D$ 为二元变量：

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i \tag{3.1}$$

- 关键假设：扰动项与核心解释变量**不相关**。

$$\textbf{Assumption LS.1:} \quad \mathrm{Cov}(D, \varepsilon) = \mathbb{E}(D\varepsilon) = 0$$

  在假设**LS.1**下，OLS 可以得到结构参数 $\beta_1$ 的一致估计。

$$b_1 = \bar{Y}_{D=1} - \bar{Y}_{D=0} \to_{\mathrm{p}} \beta_1$$

- 在假设**ID.1**下，$b_1$ 是 $\tau$ 的一致估计；在假设**LS.1**下，$b_1$ 是 $\beta_1$ 的一致估计。一个自然的问题是：$\tau$ 和 $\beta_1$ 是什么关系？以及，假设**ID.1**和假设**LS.1**是什么关系？

- 定义潜在结果：

$$Y_i^d = \mathbb{E}(Y_i^d) + U_i^d \triangleq \alpha_d + U_i^d,\ d = 0, 1$$

- 斜率参数就是平均处理效应。

$$\begin{aligned} Y_i =& (1 - D_i)Y_i^0 + D_iY_i^1 \\ =& (1 - D_i)(\alpha_0 + U_i^0) + D_i(\alpha_1 + U_i^1) \\ =& \alpha_0 + \underbrace{(\alpha_1 - \alpha_0)}_{\text{ATE}}D_i + \left[D_iU_i^1 + (1 - D_i)U_i^0\right] \end{aligned}$$

- 回忆一下，我们之前认为诸如 (3.1) 这样的线性结构模型有一个限制性假设 $\beta_{1i} \equiv \beta_1$，这其实是一种误解！事实上，线性模型 (3.1) 允许处理效应异质性的存在，斜率参数衡量的是其平均值。

- 假设**LS.1**要求

$$Cov(D, DU^1 + (1-D)U^0) = 0$$

$$\begin{aligned}
&LHS \\
&= \mathbb{E}(DU^1) - \mathbb{E}(D)\mathbb{E}(DU^1 + (1-D)U^0) \\
&= (1-\mathbb{E}(D))\mathbb{E}(DU^1) - \mathbb{E}(D)\mathbb{E}((1-D)U^0) \\
&= (1-\mathbb{E}(D))\mathbb{E}(U^1|D=1)P(D=1) - \mathbb{E}(D)\mathbb{E}(U^0|D=0)P(D=0) \\
&= \mathbb{E}(D)\big(1-\mathbb{E}(D)\big)\Big[\mathbb{E}\big(U^1|D=1\big) - \mathbb{E}\big(U^0|D=0\big)\Big]
\end{aligned}$$

- 如果假设**ID.1**成立,

$$\mathbb{E}\big(Y^d|D\big) = \mathbb{E}\big(Y^d\big)$$

则

$$\mathbb{E}(U^d|D) = 0$$

$$\mathbb{E}\big(U^1|D=1\big) - \mathbb{E}\big(U^0|D=0\big) = 0$$

因此，**假设LS.1等价于假设ID.1**。

- 再强调一遍，下面的说法是错误的 :“当假设**LS.1**成立时，线性结构模型 (3.1) 是正确的；当假设**LS.1**不成立时，线性结构模型 (3.1) 是错误的。”
- 正确的说法是：
  - 不论假设**LS.1**是否成立，线性结构模型 (3.1) 都是正确的；这一模型甚至并没有同质处理效应这一限制性假定，因为结构参数 $\beta_1$ 具有平均处理效应的含义。
  - 当假设**LS.1**成立时，可以使用线性回归得到对线性结构模型中的总体因果参数的好的估计；当假设**LS.1**不成立时，线性回归无法得到对总体因果参数的好的估计。等价地，当假设**ID.1**成立时，可以使用线性回归得到对平均处理效应的好的估计；当假设**ID.1**不成立时，线性回归无法得到对平均处理效应的好的估计。

## 随机推断 (randomization inference)

- 关于随机实验的统计推断，Athey and Imbens (2017, *Handbook*) 建议采用基于随机化的方法，而非基于抽样的传统方法。
  - 基于抽样的方法假定处理的分配是固定的，而结果是随机的，统计推断基于这样的想法：被试样本是来自于总体的随机样本。
  - 基于随机化的方法假定被试的潜在结果是固定的，而处理的分配是随机的。在这种视角下，样本就是感兴趣的总体本身。此时平均处理效应的定义是所有被试接受处理的平均结果和所有被试不接受处理的平均结果的差异。但我们无法确定性地得到这个平均处理效应，此时不确定性的来源是，对于每个被试，只观测到两个潜在结果中的一个。
- 费雪准确检验 (Fisher's exact test)。检验精确原假说 (sharp null hypothesis).

$$H_0: \; Y_i^0 = Y_i^1, \; \forall \, i = 1, \ldots, n$$

- 当原假说为真时，我们能够知晓所有的潜在结果。

$$Y_i^0 = Y_i^1 = Y_i$$

从而可以知道任何统计量 $S(\mathbf{D}, \mathbf{Y}^{obs}, \mathbf{X})$ 的准确分布（其中 **D** 为处理状态向量）。

- 计算准确 $p$ 值 (exact $p$-value).

$$\begin{aligned} p &= \Pr\left(\left|S(\mathbf{D}, \mathbf{Y}^{obs}, \mathbf{X})\right| > \left|S(\mathbf{D}^{obs}, \mathbf{Y}^{obs}, \mathbf{X})\right|\right) \\ &= \frac{\sum_{\tilde{\mathbf{D}} \in \Omega_{\mathbf{D}}} \mathbb{1}\left\{\left|S\left(\tilde{\mathbf{D}}, \mathbf{Y}^{obs}(\tilde{\mathbf{D}}), \mathbf{X}\right)\right| > \left|S\left(\mathbf{D}^{obs}, \mathbf{Y}^{obs}, \mathbf{X}\right)\right|\right\}}{|\Omega_{\mathbf{D}}|} \end{aligned}$$

- 检验统计量举例
  - 均值差异

$$S_{\text{avg}} = \bar{Y}_t - \bar{Y}_c$$

  - 分位数差异

$$S_{\text{med}} = \text{med}(Y_t) - \text{med}(Y_c)$$

$$S_q = q_\delta(Y_t) - q_\delta(Y_c)$$

- $t$ 统计量 (`help ttest`)

$$S_{t\text{-stat}} = \frac{\bar{Y}_t - \bar{Y}_c}{\sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}}}$$

其中

$$s_t^2 = \frac{1}{n_t - 1} \sum_{i:D_i=1} (Y_i - \bar{Y}_t)^2,\ s_c^2 = \frac{1}{n_c - 1} \sum_{i:D_i=0} (Y_i - \bar{Y}_c)^2$$

- 秩统计量 (rank statistic)，将所有观测到的结果按升序排列。

$$S_{\text{rank}} = \bar{R}_t - \bar{R}_c$$

- 基于回归的统计量

$$\left(\hat{\beta}_0, \hat{\beta}_X, \hat{\beta}_D\right) = \text{arg} \min_{\{\beta_0, \beta_X, \beta_D\}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_X X_i - \beta_D D_i)^2$$

$$S_{\text{coef}} = \hat{\beta}_D$$

回归模型是否正确设定不影响检验的有效性，但影响统计功效。

- 实现：可以穷举处理状态的所有分配方式，或采用模拟方法近似。(`help permute`)
  1. 基于对分配机制的理解抽取 $\tilde{\mathbf{D}}$.
  2. 根据 $\tilde{\mathbf{D}}$ 构造 $\mathbf{Y}^{obs}(\tilde{\mathbf{D}})$.
  3. 计算 $S(\tilde{\mathbf{D}}, \mathbf{Y}^{obs}(\tilde{\mathbf{D}}), \mathbf{X})$.
  4. 重复以上步骤。

- 例：假定之前的治疗数据来自随机实验。

| | | | | | |
|---|---|---|---|---|---|
| $Y_i$ | 7 | 6 | 5 | 8 | |
| $D_i$ | 1 | 0 | 1 | 0 | $\hat{S}=-1$ |
| | | | | | $\hat{S}(\tilde{\mathbf{D}})$ |
| $\tilde{\mathbf{D}}_1$ | 1 | 1 | 0 | 0 | 0 |
| $\tilde{\mathbf{D}}_2$ | 1 | 0 | 1 | 0 | $-1$ |
| $\tilde{\mathbf{D}}_3$ | 1 | 0 | 0 | 1 | 2 |
| $\tilde{\mathbf{D}}_4$ | 0 | 1 | 1 | 0 | $-2$ |
| $\tilde{\mathbf{D}}_5$ | 0 | 1 | 0 | 1 | 1 |
| $\tilde{\mathbf{D}}_6$ | 0 | 0 | 1 | 1 | 0 |
| | | | | | $\Pr\left(\lvert\hat{S}(\tilde{\mathbf{D}})\rvert > \lvert\hat{S}\rvert\right) = 0.333$ |

- 练习：下表是关于某种药物杀灭致病菌效果的小型随机试验中搜集的数据。共有 12 名被试参加了该项试验，随机选择其中 6 名服用该药物（处理组），另外 6 名服用安慰剂（控制组）。结果是关于被试体内致病菌寄生程度的指标。

| 控制组 | 处理组 |
|---|---|
| 8.62 | 0.06 |
| 1.48 | 1.72 |
| 8.93 | 2.19 |
| 9.57 | 7.32 |
| 2.65 | 7.53 |
| 7.30 | 7.62 |

我们希望检验该药物对这 12 名被试无效的精确原假说。

1. 当检验统计量为组间均值差异时，计算双侧检验的准确 $p$ 值。
2. 进行 $1,000$ 次模拟分配来计算近似 $p$ 值。
3. 计算双样本 $t$ 检验的 $p$ 值。
4. 当检验统计量为 $t$ 统计量时，计算准确 $p$ 值及其模拟近似值（注意与上题的区别）。

- 注意：随机推断是一种“推断”方法, 它不能帮助“识别”。“较小的 $p$ 值”这一发现本身并不能增强我们对因果关系成立的信心。
- 随机推断的 `Stata` 实现请见 Heß (2017, *SJ 17-3*)，更多应用请参考 Ho and Imai (2006, *JASA*), Fujiwara and Wantchekon (2013, *AEJ: Applied*) 等。

**示例 13.** 殖民制度的长期经济后果 (Dell and Olken, 2020, *RES*)

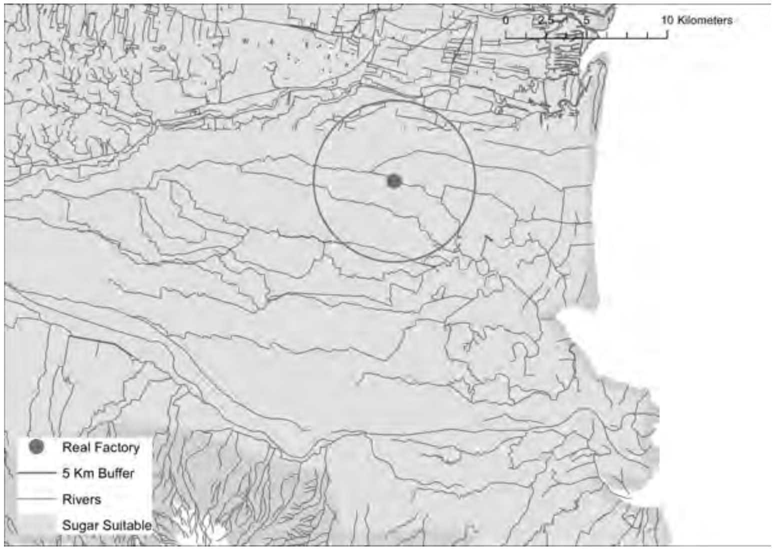
- 1830s-1870s 荷兰殖民者在印尼爪哇岛北部强制征用土地和劳动力制糖。尽管这一制糖工业体系在大萧条时期崩溃，但对当今的爪哇经济产生了持久影响。居住在旧糖厂附近的人更多地从事制造业和零售业，更少地从事农业。研究发现，影响的渠道是产业上下游的联系、基础设施建设和人力资本积累。
- 作者首先论证 94 间旧糖厂的选址除了受到一些系统性的制约外（周围必须有足够的适宜种植甘蔗的土地、必须靠近河流为糖厂提供动力、相邻糖厂不能间隔太近），由于荷兰殖民者对当地经济地理条件并不熟悉，实际选址布局在很大程度上是随机的（各糖厂生产力差异很大，未充分考虑人口因素）。

- 然后作者随机抽取满足一定条件的 1000 次反事实糖厂选址布局。这些条件包括：必须在实际糖厂上下游 5-20 公里范围内；方圆 5 公里内的适宜土地面积不低于实际糖厂周围适宜土地面积分布的 10 分位；反事实糖厂的间距不低于实际糖厂间距分布的 10 分位；整个反事实布局的平均经纬度和实际布局的平均经纬度相近。
- 对于实际布局和随机抽取的 1000 次反事实布局，估计如下方程：

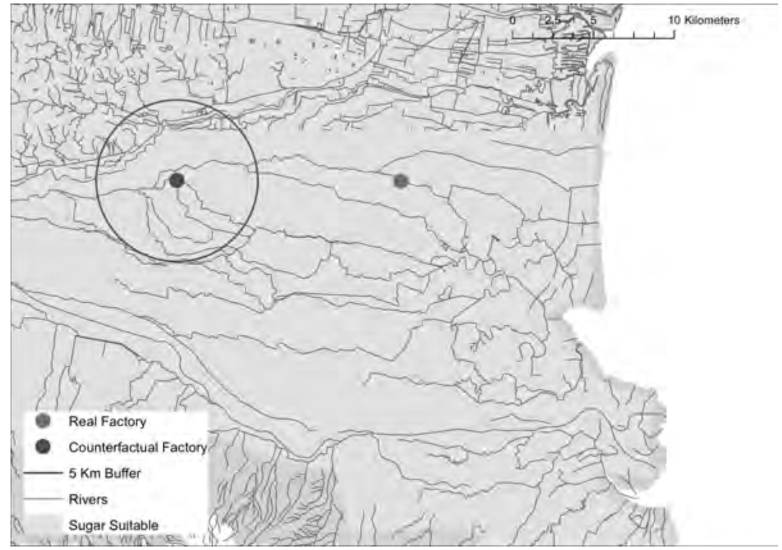
$$\texttt{outcome}_v = \alpha + \sum_{i=1}^{20} \gamma_i \texttt{dfact}_v^i + \mathbf{X}_v' \boldsymbol{\beta} + \sum_{j=1}^{n} \texttt{fact}_v^j + \varepsilon_v$$

  其中 $\texttt{dfact}_v^i$ 是虚拟变量，表示村庄 $v$ 与最近糖厂的距离是否在 $(i-1, i)$ 公里之内；fact 是最邻近糖厂固定效应。
- 糖厂效应的点估计是实际布局的 $\hat{\gamma}_i$ 和反事实布局的均值 $\bar{\hat{\gamma}}_i$ 之差。
- 根据实际布局的 $\hat{\gamma}_i$ 在反事实估计值分布中的位置计算 $p$ 值。

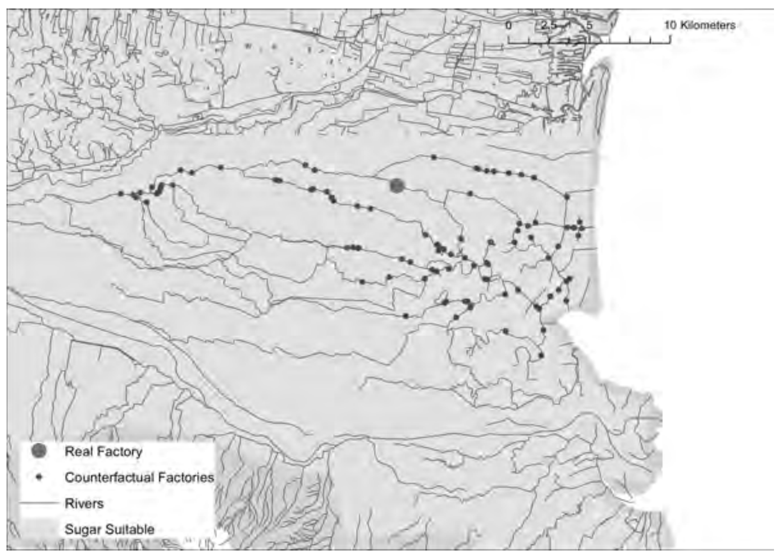
(a) Real Factory




(b) Counterfactual Factory Suitability




(c) Counterfactual Factories



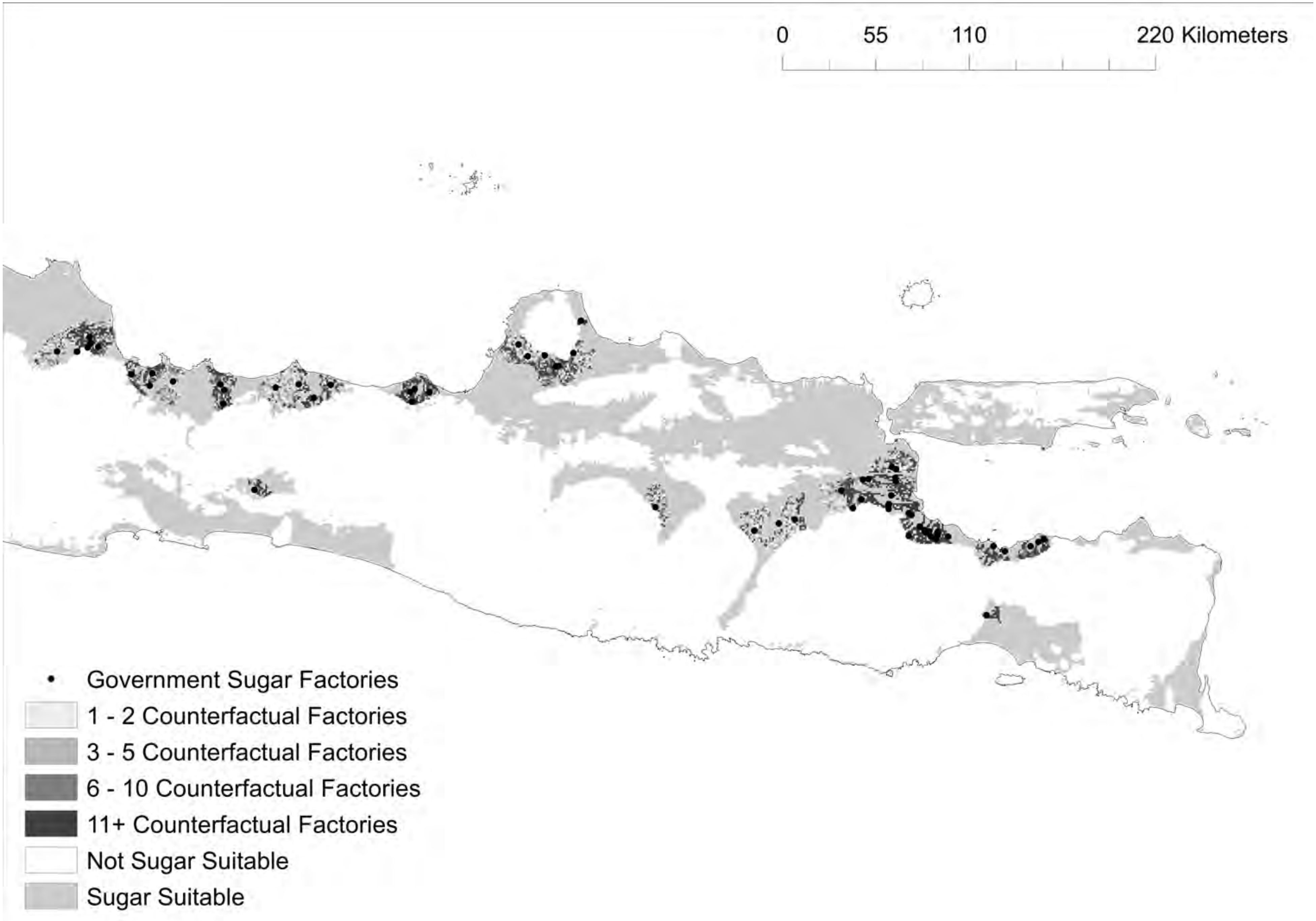
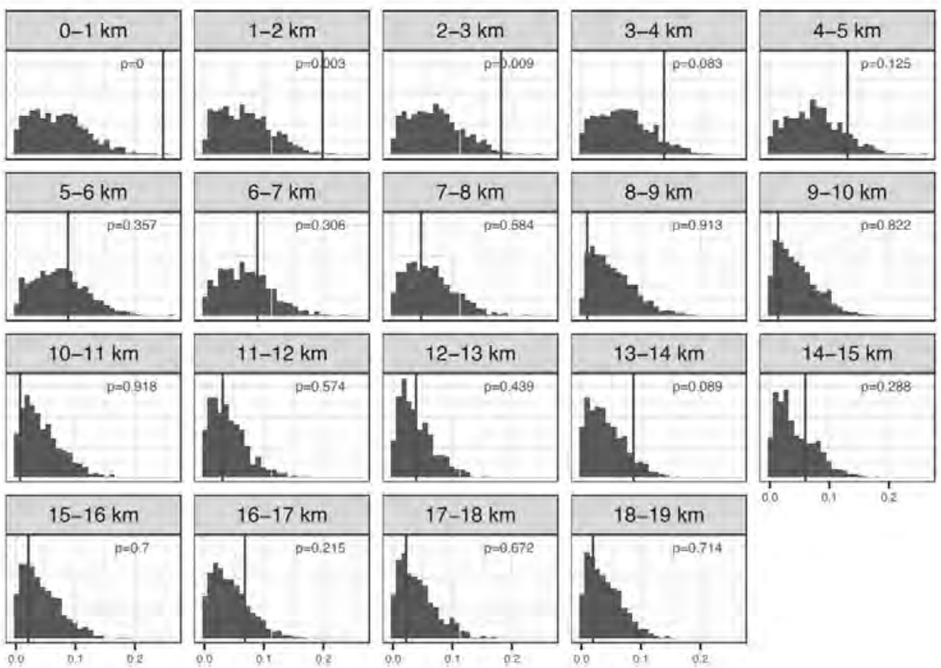
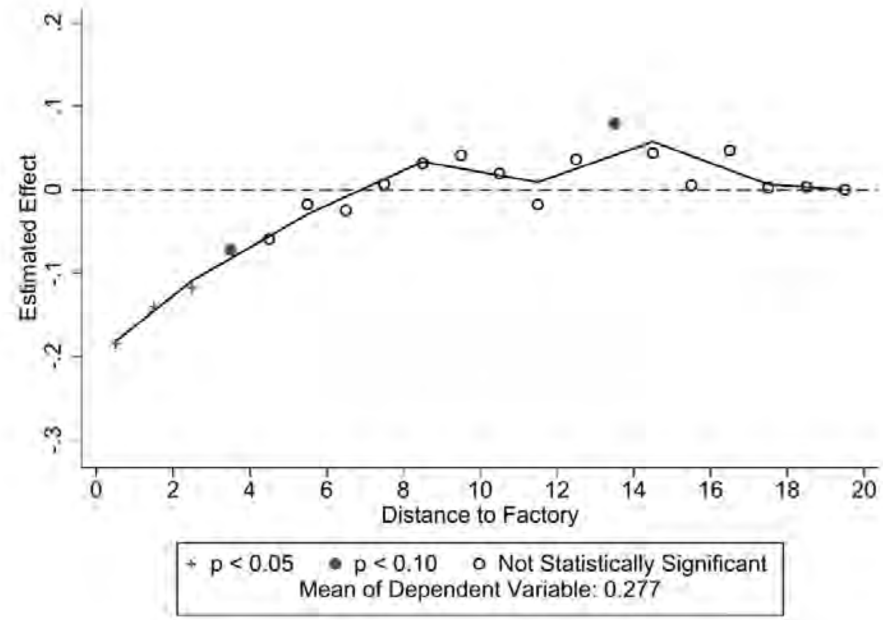

0
55
110
220 Kilometers
Government Sugar Factories
1 - 2 Counterfactual Factories
3 - 5 Counterfactual Factories
6 - 10 Counterfactual Factories
11+ Counterfactual Factories
Not Sugar Suitable
Sugar Suitable

Figure 6: Share in Agriculture (2001-11): Illustration of Methodology

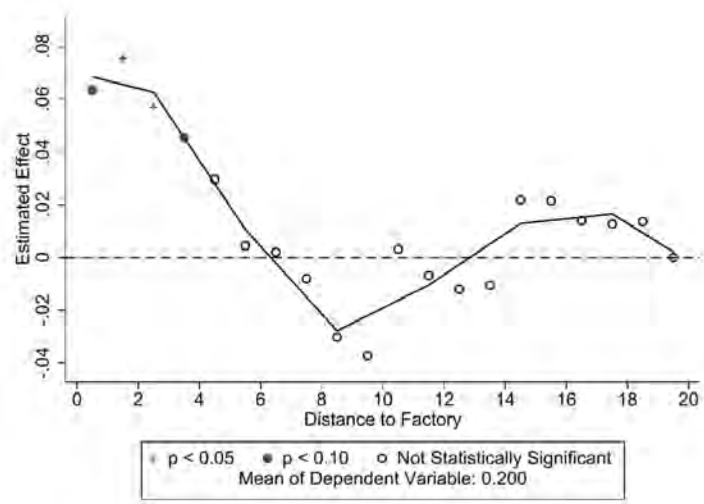
(a) Independent Shifts: Counterfactuals



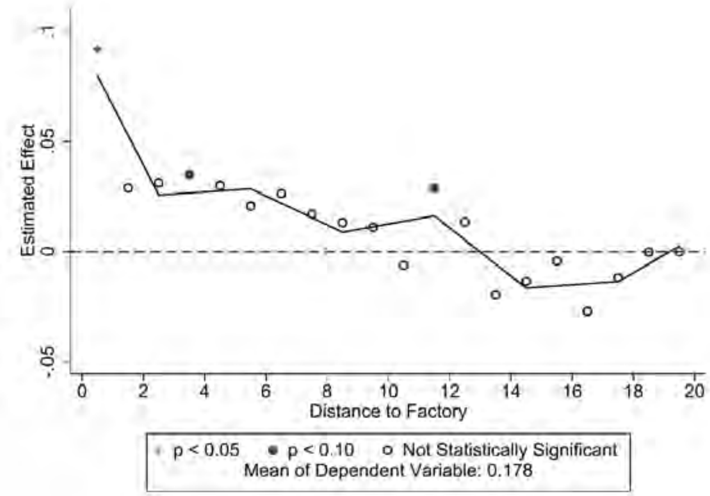

(b) Independent Shifts: Plotted Coefficients



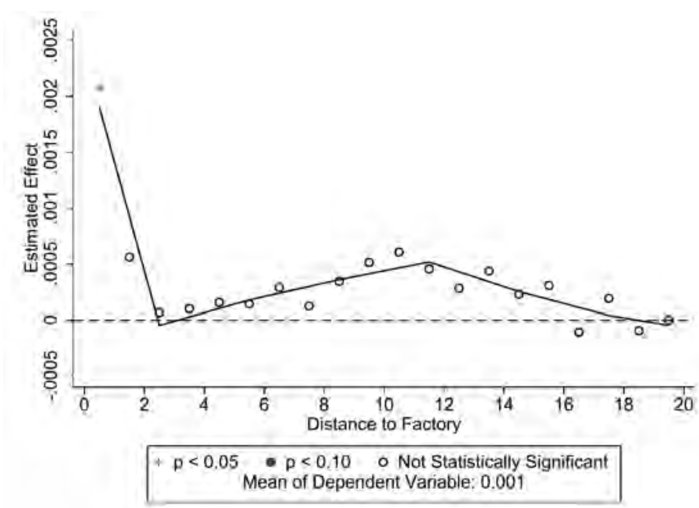
(c) Manufacturing (Susenas 2001-11)



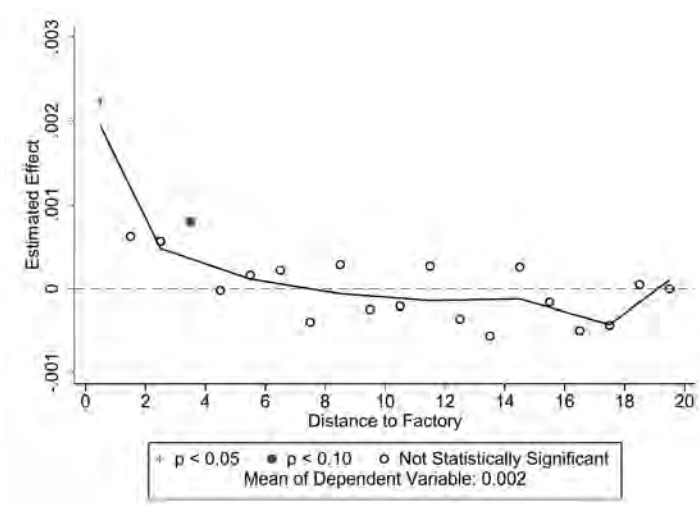

(e) Retail (Susenas 2001-11)



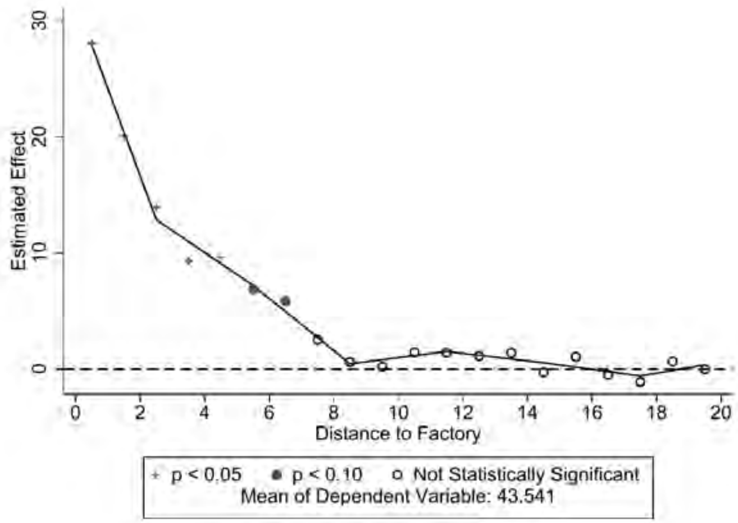

(e) Employment Share Upstream (Full Sample, Economic Census 2006)



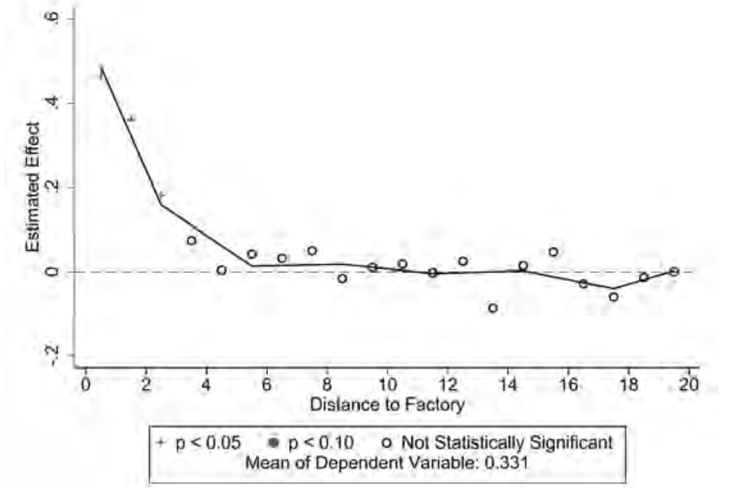

(g) Employment Share Downstream (Full Sample, Economic Census 2006)



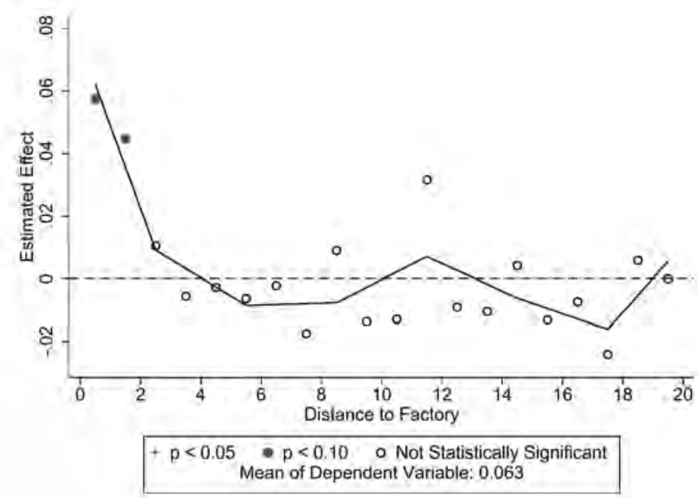
(e) Local Road Density (2017)



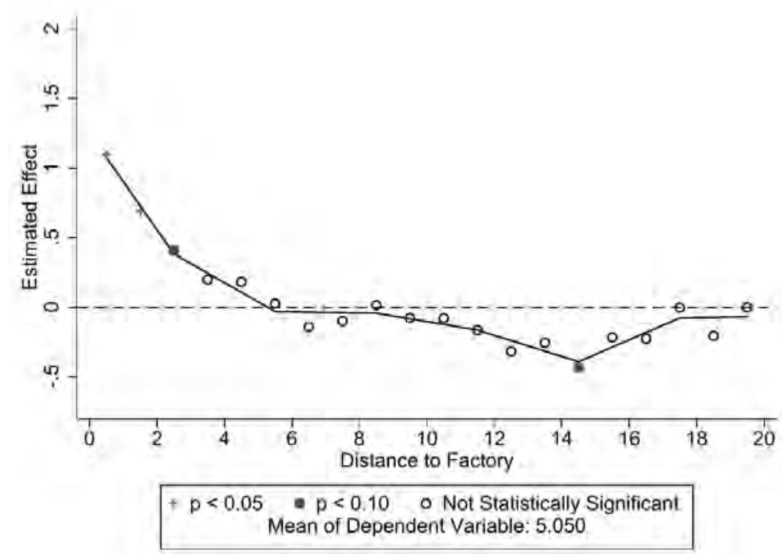

(a) Village Has Electricity (PODES 1980)




(c) High Schools (PODES 1996-2011)




(a) Years Education (2000 Census)



- 从技术上说，随机推断方法适用于任何观测性研究。Dell and Olken (2020) 为什么要使用随机推断方法，而不是传统的基于抽样的 $t$ 检验？当我们认为旧糖厂周边的村庄样本来自于一个更大的总体时，这个总体是什么？更一般地，当我们用中国省级数据做研究时，我们所关心的总体是什么？Abadie *et al* (2019) [2] 讨论了他们称之为 design-based 标准误和 sampling-based 标准误的区别。令人欣慰的结论是，一般而言前者小于后者，也就是说，传统的做法相对保守，倾向于过度拒绝。

[2] Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. 2019. "Sampling-Based vs. Design-Based Uncertainty in Regression Analysis." *Working Paper*.

- 随机推断和自助法 (bootstrapping) 的区别。
  - 自助法是指有放回的随机再抽样 (random re-sampling with replacement)，可以用来计算形式复杂的统计量的各种精确度指标（例如偏误、标准误、置信区间等）。
  - 自助法仍然是基于抽样视角，其不确定性仍然来自抽样的随机性。基本想法是把当前样本视作总体的代表性样本，因此从实际样本中进行有放回的再抽样可以视作从总体中进行再抽样。
  - 而随机推断的不确定性来自处理分配的随机性。自助法进行的是再抽样，而随机推断进行的是再分配 (re-assigning)。