1 实验与现代科学

伽利略铁球实验：对照 、可重复

实验开启了现代科学之门，而中国由于“阴阳五行”学说对几乎所有问题都可以进行解释，因此阻碍了中国进行实验。

科学的本质是“可重复”，不可重复的不是科学

史上第一个对照试验： 林德与败血症，12个船员分为6组，给予不同的治疗方法，只有吃了柠檬和橘子的人痊愈。**但这个实验一个很重要的问题在于他没有随机抽样，不同组的病人症状并不一致。**

1.1 幸存者偏差

1. 用于分析的数据使用了不能保证随机化的方法获得，因此从样本中分析的结果不能代表总体特征。
2. 统计科目和样本是否能被纳入统计有相关性，所以样本呈现的结果和实际结果存在偏差。（如为什么科研领域英年早逝的现象越来越多，其实所有领域都有）
3. **解决方法**
    1. 统计学推导
    2. 样本覆盖的范围足够大、足够多、足够随机

1.2 史上第一个随机对照试验： 马歇尔用链霉素治疗肺结核的实验

研究目的：评价链霉素治疗肺结核的有效性和安全性

研究对象：入组经细菌学确诊的，年龄为15–30岁的双侧急性进展性原发型肺结核患者，排除陈旧性肺结核，结核空洞患者

研究方法：多中心、随机、空白对照设计；实验组接受链霉素治疗并卧床休息，对照组仅卧床休息

随机对照实验三原则：随机、对照、盲法(双盲：参与者和研究者都不知道具体的干预措施，从而避免数据收集和评价过程中带来的可能的偏差)

选择性偏差都能用随机对照试验解决吗？可以。

读书无用论的观点完全来源于幸存者偏差。

很多时候随机对照试验不可行「伦理问题」，因此用准实验方法代替

2 准实验方法（自然实验）：观察，不会人为给对照组

2.1 回归分析

作为随机对照实验的替代，回归进行因果推断的前提假设是：当处理组与控制组在可观察的关键变量上都一样时，我们看不见的因素造成的选择性偏差几乎都能得到消除

高尔顿与回归：他发现孩子的身高是父亲身高和当地平均身高的加权平均值

回归的实质就是求条件期望函数

回归在严格控制其他变量的情况下，是**可以进行因果推断**的

2.2 准实验的定义

利用**观测数据和外生冲击**，找到类似于随机对照实验的环境，在不对实验对象进行干预的情况下得出需要的结论，作用是尽可能消除选择性偏差

常用方法： 匹配[1]、双重差分[2]、工具变量[3]、断点回归[4]

2.2.1 匹配

- 通过那些可观测特征人为地构造出一个对照组
- 通过匹配构造出的对照组与实验组拥有相同的随机分布。（例如：找一对双胞胎，找不到就找两个条件近似双胞胎的人）

`TWINS DATA（双胞胎数据）：`解决不可观察的能力和家庭背景造成的偏误

> 双胞胎数据下的分析发现遗传因素和家庭背景同时影响着受教育年限和收入，受教育年限对收入的影响大大降低了

2.2.2 DID 双重差分，PSM倾向值匹配：横向是基准值，纵向是变动值

DID：difference in difference

修高铁穿过一组城市，未穿过一组城市，用穿过的城市的GDP的增幅减去未穿过的城市的GDP的增幅。（自身纵向比较+不同组横向比较）

1. Δx −Δy
2. 先纵向：给定一个城市，比较时间前后给定属性的差值。（自身发展）
3. 后横向：比较两个城市之差。

这是两个不同的维度。（基准）

来到中国的科研人员是否提升？寻找原情况基本相同的来中国与没来中国的两组科学家，比较他们转移后的业绩提升。

2.2.3 断点回归

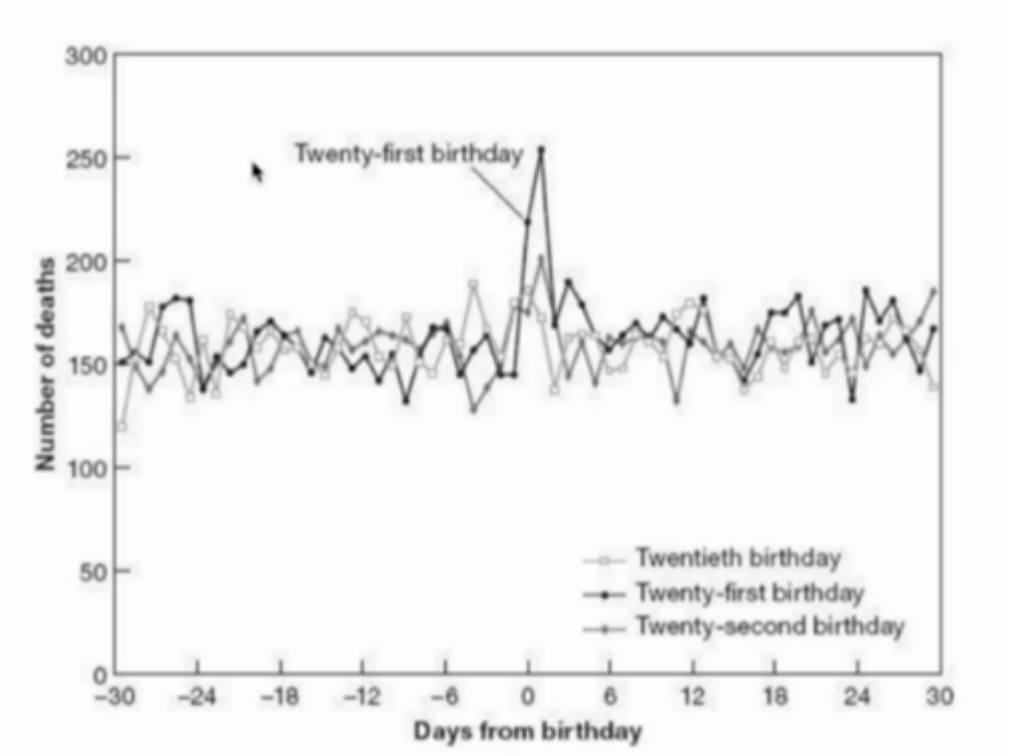
存在一个变量，如果该变量大于一个临界值时，个体接受treatment，而在该变量小于临界值时，个体不接受treatment。

一个群体因某个固定标准被分为两类，我们认为恰好满足标准的人与差一点满足标准的人并没有太大的差异，只是“运气”让一部分人被选中而另一部分没被选中。

> 例如：高考“踩分数线”与“差一分就考上”的人认为它们能力没有差异，但是两人在不同学校学习后的影响不同，从而看出不同学校培养能力的差距。

> 再例：研究雾霾对人健康的影响，由于**以淮河为界供暖政策（固定标准）**导致河南北的大气污染差异，这样就控制了气候、经济等其他的变量。

> 再例：21岁魔咒，因为禁酒令被解除



2.3.4 工具变量

既然可能会有无法掌控因素同时影响着受教育年限和未来收入，引起选择性偏差，那是否可以**找一个不受这些无法掌控因素影响的，但也能够对受教育年限进行直接影响的变量**，讨论它和未来收入的关系，从而避开选择性偏误的问题

> 例如：美国满16岁即可离校，因此年末出生的人往往比年初出生的人多学几个月，因此成为一个研究受教育程度和收入关系的绝佳变量

3 相关和因果

1. 夏天溪水蒸发量和冰棍销量：有相关性，无因果性
2. 夏天气温和冰棍销量：有相关性，有因果性

偶然谬误、虚假因果、反向因果

1. 张学友开演唱会抓罪犯、萧敬腾雨神：偶然现象
2. 住房越挤，夫妻吵架频率越高：虚假因果，本质因素是家庭经济状况
3. 企业开的工资越高，反而越缺人：反向因果，因为缺人才开高工资

因果推断是建立在随机对照试验或准试验的基础上的

1. 逻辑推理：“人会死，苏格拉底是人，苏格拉底会死”
2. 人工智能：贝叶斯网络推断因果关系

相关性是因果性的必要不充分条件，因果性包含相关

在社会问题中，用单纯的相关性来指导人们的行为、政策的制定是不可靠、甚至危险的

破案时需要进行因果推断。那么刑侦警察是怎么做因果推断的？其实目前有两套结论可以被接受的推理方法：

1. 来自计量经济学的因果推断：是定量分析方法
2. 来自哲学的逻辑推理：是定性研究方法

4 数据从哪里来

开放科学正在拆掉知识的高墙

1. 因皇家个人偏好研究科学
2. 科学研究职业化：申请书争取纳税人缴纳的税款中所拨的科研经费
3. 出版商寡头垄断地位与开放科学的抗争。

政策一刀切：断点回归

1. 受资助的资金充裕
2. 没受资助的更拼命

开放科学对于科研成果透明、提高科研效率、提高对科研的经济赞助、更快的知识传播和更高的知识参与度以及政府决策和人权民主都具有积极作用。`实例：APS、ORCID、PubMed`

---

1. **2.2.1 匹配**
2. **2.2.2 DID 双重差分，PSM倾向值匹配：横向是基准值，纵向是变动值**
3. **2.3.4 工具变量**
4. **2.2.3 断点回归**