清华大学数据库技术与应用

数据统计 II

授课教师：计算机系王健楠

授课学期：2026年（春季）

课程大纲

清華大學 Tsinghua University

01
估计与自助法

02
假设检验

03
因果推断

课程大纲

01
估计与自助法

- 点估计
- 区间估计
- 自助法（Bootstrap）

02
假设检验

03
因果推断

估计（Estimation）

清華大學 Tsinghua University

问题定义

- 估计与总体相关的某个数值

示例

- 估计美国将投票给Trump的人口比例
- 估计美国所有家庭的年收入中位数

例题：家庭年收入中位数

如何估计美国家庭的年收入中位数?

- 从美国随机抽取 10,000 个家庭
- 报告其年收入中位数：50,000 美元

但是，我们需要给出类似这样的结论：

50,000 ± 500 美元

暴力解法（代价昂贵）

- 从美国随机抽取 10,000 个家庭
- 报告其年收入中位数

重复此过程 100 次

50,000 49,200 50,200 … 49,200

需要总计调查 1,000,000 个家庭!

自助法（Bootstrapping）

清華大學 Tsinghua University

核心思想：重采样（Resampling）

- 有放回地从原始样本中重新抽样

总体： 1, 1, 8, 2, … 3, 3

样本： 3, 8, 1, 8, 3

重采样： 8, 3, 3, 3, 1

自助法流程

- 从美国随机抽取 10,000 个家庭
- 从这 10,000 个家庭中抽取重采样
- 报告重采样的年收入中位数

重复此过程 100 次

无需调查任何新家庭！

自助法注意事项

- 从足够大的随机样本开始（至少 30 个样本）
- 尽可能多次重复重采样过程（超过 1000 次）
- 不适用于估计最大值 / 最小值

课程大纲

清華大學 Tsinghua University

01

估计与自助法

- 点估计
- 区间估计
- 自助法（Bootstrap）

02

假设检验

- 零假设与备择假设
- P 值与 P 值作假
- A/B 测试

03

因果推断

- 基本概念
- 平均处理效应（ATE）
- 实践案例

为什么需要假设检验?

清華大學 Tsinghua University

问题背景

- 我们想从数据中得出一个论断
- 但数据只是一个样本
- 在这种情况下，如何证明我们的论断?

示例

- 论断：数据科学家的收入高于数据工程师
- 数据：各50名数据科学家和数据工程师的样本
- 结果：100K vs. 70K

我们能用这个结果证明论断正确吗?

假设检验

清華大學 Tsinghua University

等价术语

- 假设　　= 论断
- 假设检验 = 论断证明

核心思想

- 反证法

类比

- 如何证明：不存在最小的正有理数?
- 假设存在最小正有理数 a/b，则 a/(2b) 也是正有理数且更小，矛盾

备择假设与零假设

清華大學 Tsinghua University

备择假设 H_a

- 这是你想要证明正确的论断

零假设 H_0

- H_a 的对立面

可能的结论

- 拒绝 H_0（发现矛盾） → 接受 H_a
- 未能拒绝 H_0（未发现矛盾） → 无法接受 H_a

假设检验示例

备择假设 H_a

- 清华毕业生的收入高于北大毕业生的收入

零假设 H_0

- 清华毕业生的收入不高于（等于或低于）北大毕业生的收入

若 H_0为真，观察到以下结果的概率是多少?

这是P值（P-value）

- ~~清华毕业生(100K) vs. 北大毕业生 (70K)~~
- Salary(清华毕业生）- Salary（北大毕业生） = 30K

基于 P-value 做决策

我们希望：

- p-value 越低越好，这样我们就可以拒绝原假设$\mathbf{H_0}$（即：接受备择假设 $\mathbf{H_a}$）

设定显著性水平 α（例如：α = 0.05）

- 若 P 值 < α，则拒绝 $\mathbf{H_0}$
- 若 P 值 ≥ α，则未能拒绝 $\mathbf{H_0}$

置信水平（例如：c=1-α = 95%）

- 我们对所做的决定有多大把握（信心）？

P 值作假（P-Hacking）

常见错误

- 持续收集数据，直到假设检验通过为止
- 对同一组数据不断进行分析，直到发现显著的结果为止

- 这是一种严重的科学不诚信行为
- 可重复性危机的主要原因之一

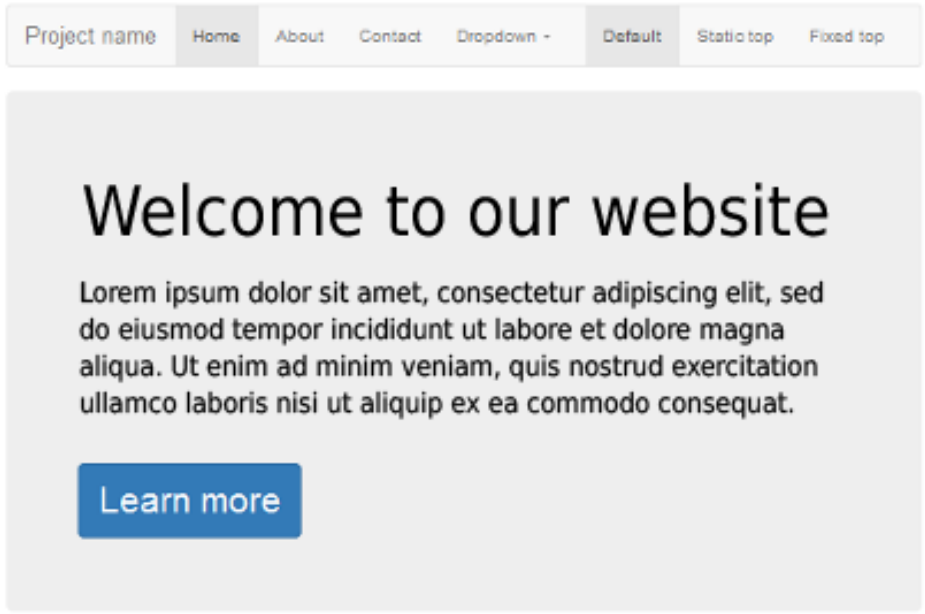
解决方案

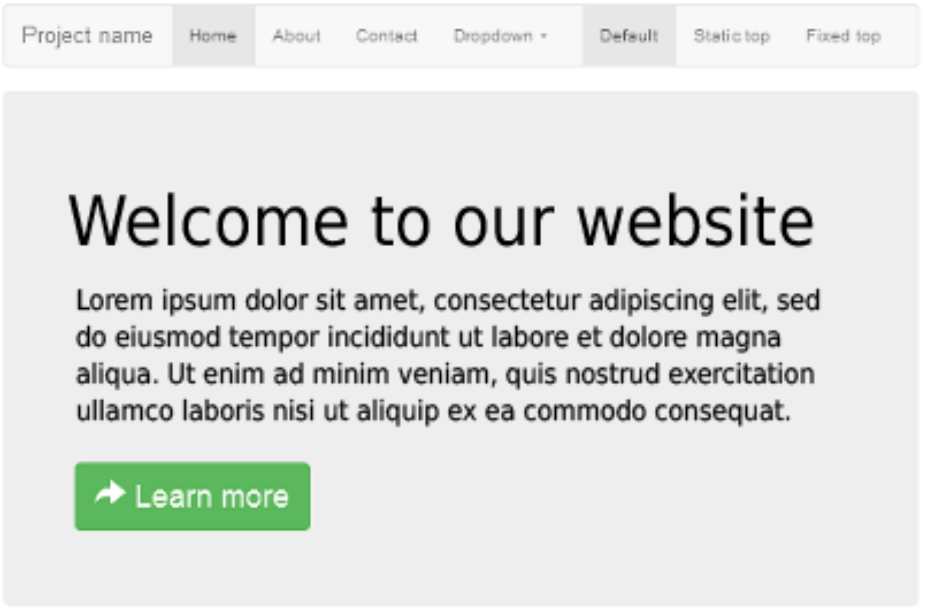
- 开始前就明确原假设 (H_0) 和备择假设 (H_a)，避免“射箭后再画靶子”。
- 降低显著性水平（例如：如果对同一组数据进行两次假设检验，则将显著性水平设为 $\alpha/2$）。

A/B 测试

清華大學 Tsinghua University

哪种界面更好?





令人惊讶的 A/B 测试

- 即使是微小的视觉差异也能产生显著的转化率差异
- A/B 测试常常得出反直觉的结论

A. Get $10 off the first purchase. Book online now!

B. Get an additional $10 off. Book online now.

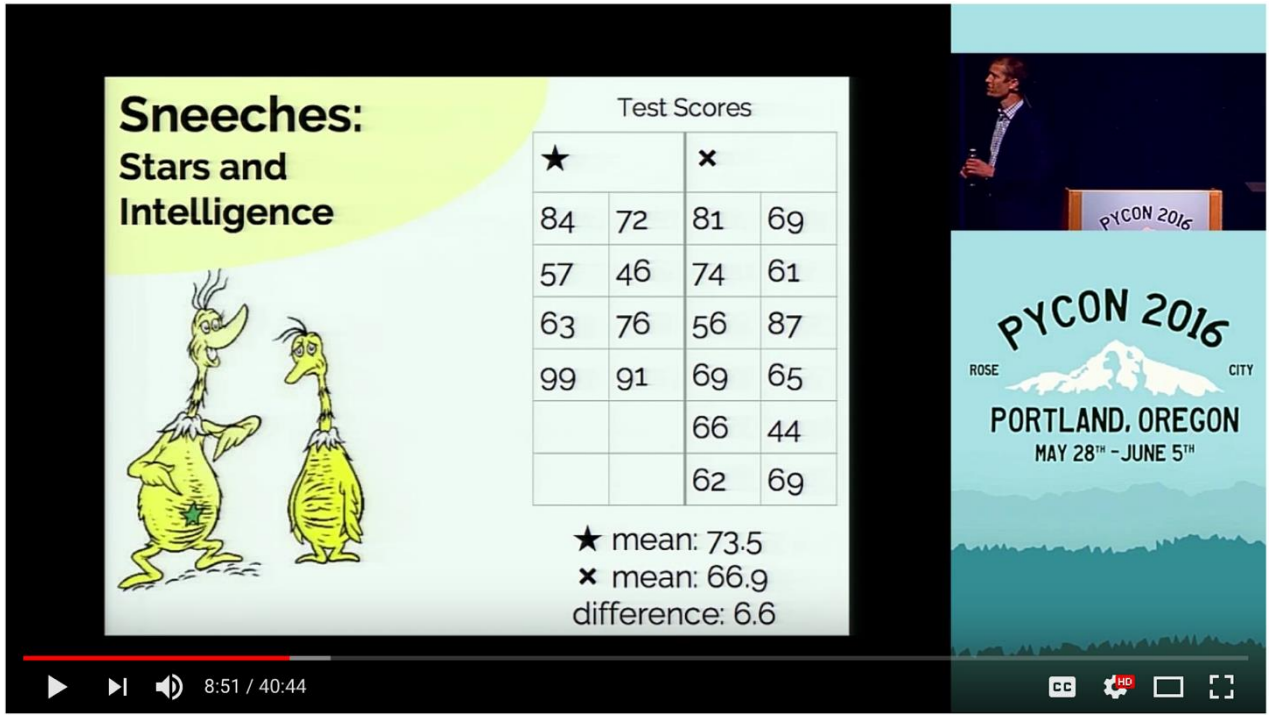
参考：https://www.wordstream.com/blog/ws/2012/09/25/a-b-testing

置换检验（Permutation Test）

置换检验（Permutation Test）

- 通过随机置换标签来构建零假设分布
- 计算观察统计量在零分布中的位置（即 P 值）
- 不依赖参数假设，适用于小样本

视频：https://youtu.be/Iq9DzN6mvYA?t=8m9s



小结：假设检验

假设检验小结

- 零假设 H_0 与备择假设 H_a
- P 值与 P 值作假（P-Hacking）
- A/B 测试

记住：反证法是假设检验的核心逻辑

课程大纲

清華大學 Tsinghua University

01

估计与自助法

- 点估计
- 区间估计
- 自助法（Bootstrap）

02

假设检验

- 零假设与备择假设
- P 值与 P 值作假
- A/B 测试

03

因果推断

- **为什么重要**
- **基本概念**
- **因果推断方法**

数据科学家能回答的问题

问题类型	机器学习任务
A 还是 B?	分类
有多少?	回归
这异常吗?	异常检测
如何分组?	聚类
What if?	因果推断（Causal Inference）

从预测到因果

清華大學 Tsinghua University

淘宝用户活跃度预测

- Y = 下个月的登录次数
- X = 过去登录次数、好友数量等

如果我们增加好友数量会怎样?

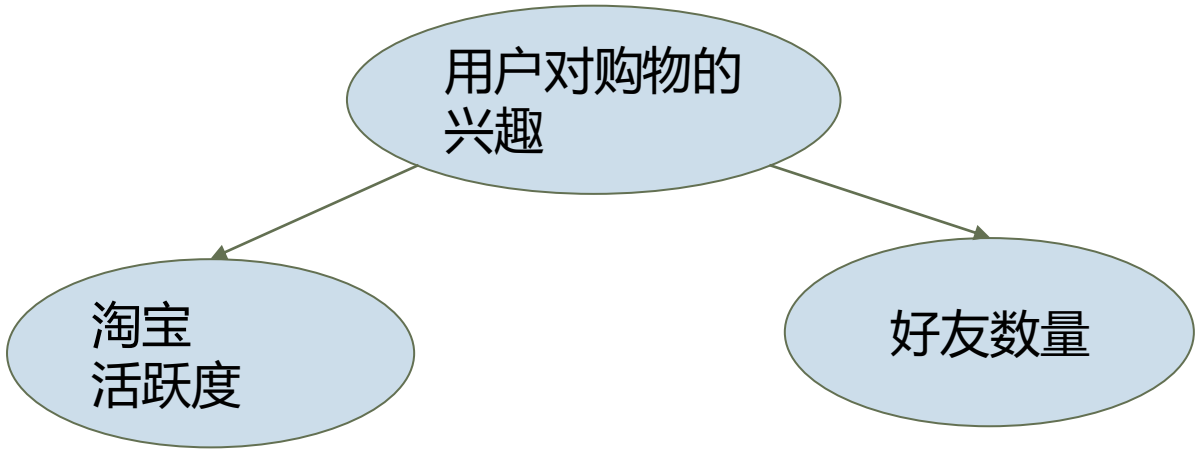
- 好友数量增加是否会提升用户活跃度?

可能是，也可能不是

可能性 1：A 导致 B

可能性 2：B 导致 A

可能性 3：C 同时导致 A 和 B



A/B 测试大有帮助！

实验组（Treatment Group）

- 随机抽样：从用户中随机选出一部分。
- 施加干预：上线一个旨在“增加好友数”的营销活动。
- 观测结果：统计这组用户在下个月的平均活跃度。

对照组（Control Group）

- 随机抽样：从用户中随机选出一部分。
- 保持现状：不开展增加好友的活动。
- 观测结果：统计这组用户在下个月的平均活跃度。

假设检验

A/B 测试行不通的情况

清華大學 Tsinghua University

情形一：A/B 测试不可行

- 如果你去的是 PKU 而不是 THU，求职会更好吗?
  - 无法对同一人做两种不同的大学选择实验

情形二：A/B 测试不道德

- 如果订阅价格定为 69 美元而非 99 美元，会增加收入吗?
  - 对不同用户收取不同价格可能涉及歧视

这时候，我们需要因果推断方法!

课程大纲

清華大學 Tsinghua University

01
估计与自助法

- 点估计
- 区间估计
- 自助法（Bootstrap）

02
假设检验

- 零假设与备择假设
- P 值与 P 值作假
- A/B 测试

03
因果推断

- **为什么重要**
- **基本概念**
- **因果推断方法**

基本概念

- 结果变量与处理变量
- 干预与 Do 算子
- 反事实
- 因果图

结果变量与处理变量

如果参加**课外班**，是否会提高**成绩**？

干预变量

结果变量

学生	性别	班级	课外班（处理变量）	成绩（结果变量）
Jacky	男	1	0	78
Terry	男	1	1	82
Mary	女	1	0	86
Sarah	女	2	1	83

干预与 Do 算子

Do 算子（Do Operator）

- 由 Judea Pearl 提出，用于表示实验干预

$$P(Grade \mid do(StudyProgram = 1))$$

上述概率表示若强制让某人参加课外班，其成绩的分布

干预与 Do 算子（续）

清華大學 Tsinghua University

两者的关键区别

P(A | B = b)：

- 在观察到 B = b 的条件下，A 为真的概率

P(A | do(B) = b)：

- 在通过干预将 B 设置为 b 的条件下，A 为真的概率

观察到某人参加课外班 ≠ 强制让某人参加课外班

反事实（Counterfactual）

如果改变了处理变量，结果会怎样?

注意：反事实成绩是现实中无法观测到的假设结果

学生	性别	班级	课外班	实际成绩	反事实成绩
Jacky	男	1	0	78	82
Terry	男	1	1	82	82
Mary	女	1	0	86	90
Sarah	女	2	1	83	85

因果关系（Causality）

因果关系的定义

- 实际结果与反事实结果之间的差值

因果效应 = (现实中你做了某事的结果) - (如果你当初没做某事的结果)。

药效 = （你吃药后的感冒痊愈时间） - （减去你不吃药也会痊愈的时间）

因果推断的根本性难题

- 我们无法观测到反事实结果

因果图（Causal Graph）

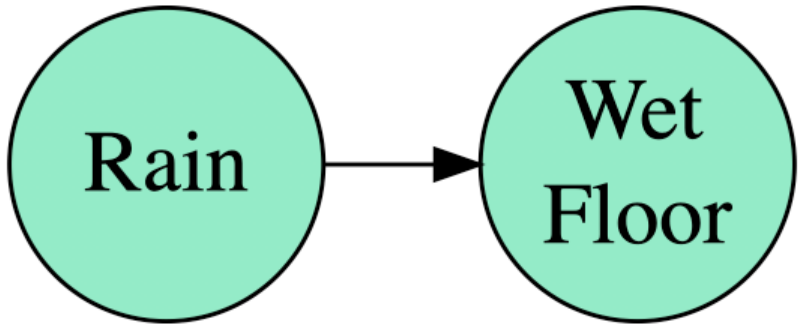
清華大學 Tsinghua University

因果图是有向图

- 节点：变量
- 有向边 X → Y：X 影响 Y

示例

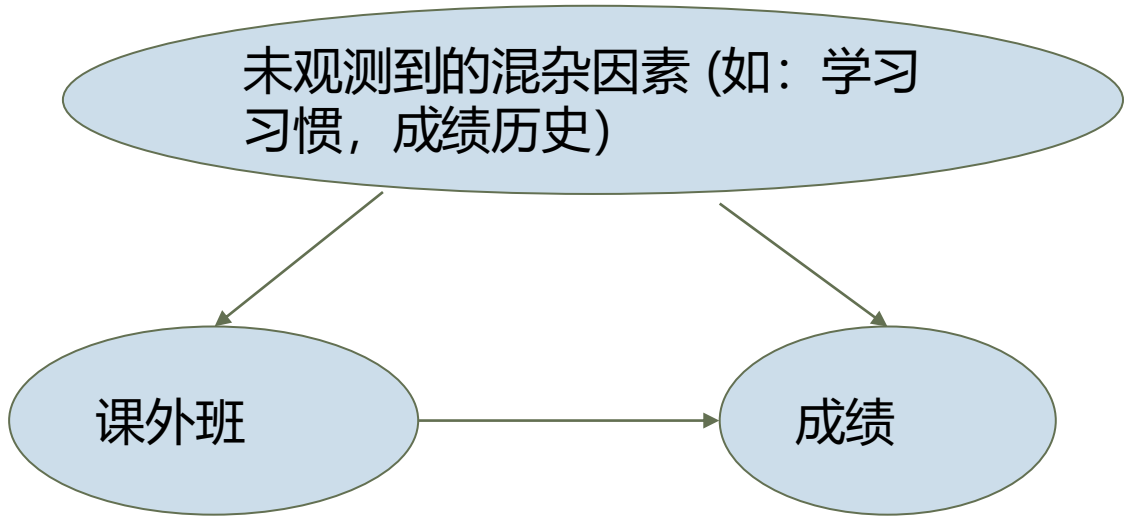
- 上课外班 → 成绩好
- 下雨 → 地面湿



因果图（续）

混杂因素（Confounding Variable）

- 同时是处理变量和结果变量的共同原因



为什么要用因果图?

- 有助于识别哪些变量需要被控制 (Control)。
- 使分析中的假设更加明确化。

练习1

研究人员发现：青少年溺水死亡案例增加的同时，冰淇淋销量也在上升。

练习1

练习2

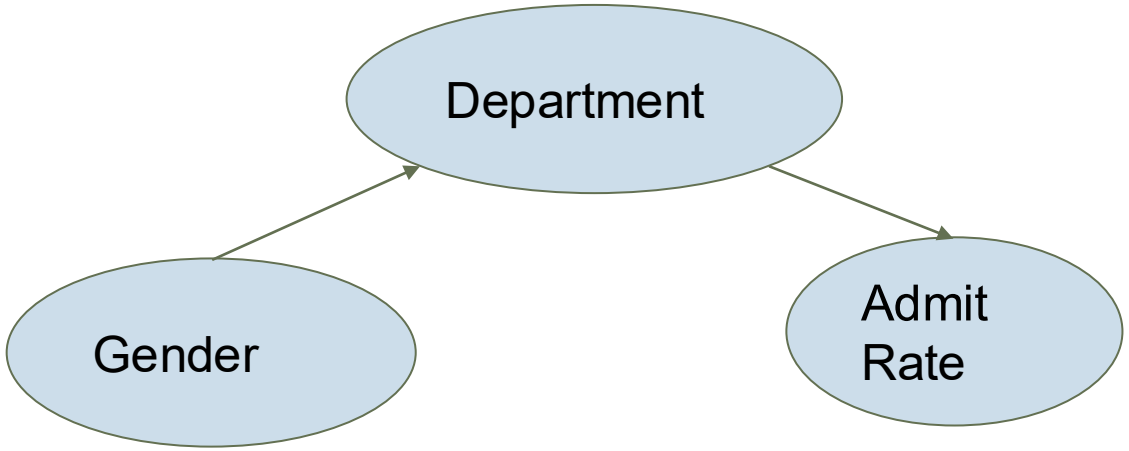
清華大學 Tsinghua University

UC Berkeley 存在性别偏见吗?

	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

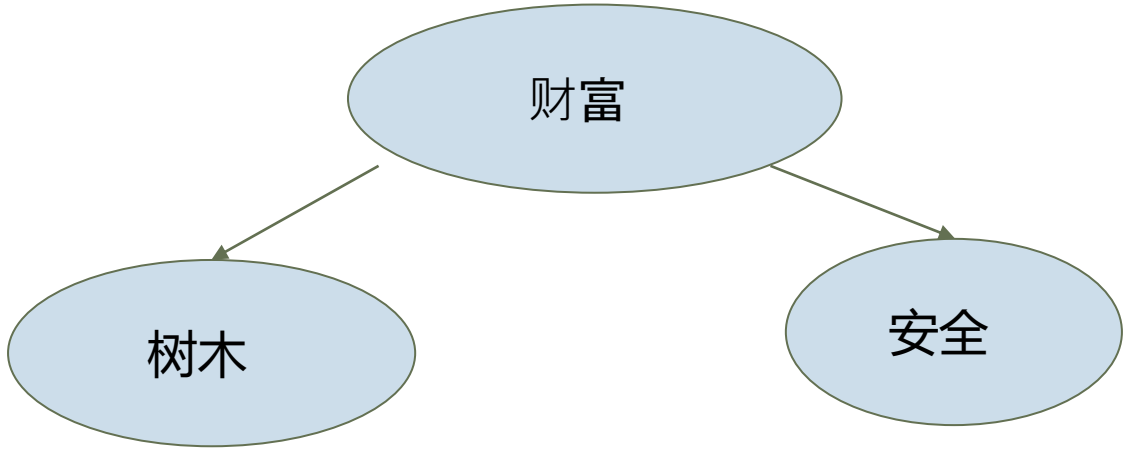
练习2

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**



练习3


The Morton Arboretum
Did you know?
Trees make our cities safer.
Learn more:
MortonArb.org/EveryTreeMatters


财富
树木
安全

课程大纲

清華大學 Tsinghua University

01

估计与自助法

- 点估计
- 区间估计
- 自助法（Bootstrap）

02

假设检验

- 零假设与备择假设
- P 值与 P 值作假
- A/B 测试

03

因果推断

- **为什么重要**
- **基本概念**
- **因果推断方法**

统计推断 vs. 因果推断

清華大學 Tsinghua University

统计推断

- 数据是样本
- 目标是推断总体
- 核心思路：思考如何从局部样本“反向推导”出整体总体的特征。

因果推断

- 从数据中划分出实验组
- 从数据中划分出对照组
- 核心思路：思考如何通过实验组与对照组的对比，推断出干预措施带来的实际效果。

个体干预效应

参加 vs. 不参加学习项目，成绩差异是多少?

学生	性别	班级	上课外班（成绩）	未上课外班（成绩）	个体干预效应
Jacky	男	1	82	78	4
Terry	男	1	82	82	0
Mary	女	1	90	86	4
Sarah	女	2	83	85	-2

平均干预效应（ATE）

清華大學 Tsinghua University

ATE = 所有个体干预效应的平均值

$$ATE = (4 + 0 + 4 + (-2)) / 4 = 1.5$$

学生	性别	班级	上课外班（成绩）	未上课外班（成绩）	个体干预效应
Jacky	男	1	82	78	4
Terry	男	1	82	82	0
Mary	女	1	90	86	4
Sarah	女	2	83	85	-2

ATE 估计方法

基于匹配：

- 精确匹配
- 最近邻匹配
- 倾向评分匹配

基于机器学习：

- 回归方法
- 表示学习

精确匹配

清華大學 Tsinghua University

寻找「反事实世界中的完美匹配」

- 对每个处理组样本，找到特征完全相同的对照组样本

学生	性别	班级	课外班	成绩
Terry	男	1	1	82
Sarah	女	2	1	83
Jacky	男	1	0	78
Mary	女	1	0	86

精确匹配（续）

Terry（男, 班级1）可以与 Jacky（男, 班级1）精确匹配，效应 = 82 − 78 = 4

- 但 Sarah（女, 班级2）无法找到精确匹配的对照组样本

学生	性别	班级	课外班	成绩
Terry	男	1	1	82
Sarah	女	2	1	83
Jacky	男	1	0	78
Mary	女	1	0	86

精确匹配的局限性

清華大學 Tsinghua University

无法计算整个总体的 ATE

- 对于 Sarah，无法在对照组中找到完美匹配（女性, 班级2）
- 实际数据中，往往很难找到完美匹配

学生	性别	班级	课外班	成绩
Terry	男	1	1	82
Sarah	女	2	1	83
Jacky	男	1	0	78
Mary	女	1	0	86

最近邻匹配

清華大學 Tsinghua University

寻找「反事实世界中最接近的匹配」

- Sarah（女, 班级2）的最近邻是 Mary（女, 班级1）
- 效应 = 83 − 86 = −3

学生	性别	班级	课外班	成绩
Terry	男	1	1	82
Sarah	女	2	1	83
Jacky	男	1	0	78
Mary	女	1	0	86

最近邻匹配：计算 ATE

- Terry ↔ Jacky（完美匹配）：效应 = 82 − 78 = 4
- Sarah ↔ Mary（最近邻匹配）：效应 = 83 − 86 = −3

ATE = ½ × (4 + (−3)) = 0.5（近似）

学生	性别	班级	课外班	成绩
Terry	男	1	1	82
Sarah	女	2	1	83
Jacky	男	1	0	78
Mary	女	1	0	86

倾向性评分匹配

清華大學 Tsinghua University

步骤

- 第一步：用逻辑回归估计倾向性评分 $e(x) = Pr[T=1 \mid X=x]$
  - 即每个个体接受处理的条件概率

- 第二步：基于倾向评分，使用匹配方法将对照组（T=0）的用户与实验组（T=1）的用户进行配对。

示例：倾向性评分匹配

- Terry（PSE=0.70）与 Jacky（PSE=0.70）匹配 → 效应 = 4
- Sarah（PSE=0.60）与 Mary（PSE=0.55）匹配 → 效应 ≈ −3

ATE ≈ 0.5，与最近邻匹配结果相近

学生	性别	班级	课外班	成绩	倾向性评分
Terry	男	1	1	82	0.70
Sarah	女	2	1	83	0.60
Jacky	男	1	0	78	0.70
Mary	女	1	0	86	0.55

回归方法（Regression Method）

清華大學 Tsinghua University

直觉

- 在不同处理条件下，Y 给定 X 的分布不同

步骤

- 分别在 T=0 和 T=1 的数据上训练两个回归模型
- 推断 p(Y | T=0, X) 和 p(Y | T=1, X)

$$\mathbf{ATE = E[p(Y|T=1,X)] - E[p(Y|T=0,X)]}$$

回归方法：示例

- 模型1：在 [Terry, Sarah]（T=1）上训练
- 模型2：在 [Jacky, Mary]（T=0）上训练
- 用模型1为 [Jacky, Mary] 预测反事实结果，模型2同理

学生	性别	班级	课外班	成绩
Terry	男	1	1	82
Sarah	女	2	1	83
Jacky	男	1	0	78
Mary	女	1	0	86

其他基于机器学习的方法

表示学习（Representation Learning）

- 直觉：将数据集变换到一个处理分配更均匀的空间
- 在该空间中，处理组和对照组的协变量分布更为接近

其他进展

- 更多前沿技术，请参阅最新学术论文
- 因果推断是机器学习领域的重要研究方向

JUDEA PEARL DL

United States – 2011

CITATION

For fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning.

A.M. TURING AWARD

因果推断案例研究

清華大學 Tsinghua University

如果你选择了 PKU 而非 THU，就业前景会更好吗?

1. 需要收集哪些数据?

2. 如何利用数据回答这个因果问题?

应用案例：无家可归者政策

无家可归者干预措施分配

- 研究人员利用因果推断，为无家可归者分配不同干预措施（紧急庇护所、快速再入住等)

- 发表于 AAAI 2019

Amanda Kube, Sanmay Das, Patrick J. Fowler: Allocating Interventions Based on Predicted Outcomes: A Case Study on Homelessness Services. AAAI 2019: 622-629

应用案例：社交媒体

社交媒体中的因果推断

- 研究 Twitter 上种族、性别与亲密度对说服效果的影响

- 发表于 NeurIPS 2019

https://cpb-us-w2.wpmucdn.com/sites.coecis.cornell.edu/dist/a/238/files/2019/12/Id_104_final.pdf

工具：DoWhy Python 库

清華大學 Tsinghua University

DoWhy Python 库

- 微软开发的因果推断 Python 库
- 支持因果假设的显式建模与检验
- 结合因果图模型与潜在结果框架

https://github.com/microsoft/dowhy

总结

估计与自助法

- 点估计与区间估计
- Bootstrap 重采样方法

假设检验

- H_0 与 H_a
- P 值
- P-value 作假
- A/B 测试

因果推断

- What if 问题
- 为什么不能只用 A/B 测试?
- 因果关系 ≠ 相关关系

- 基本概念：结果变量、Do 算子、反事实、因果图
- ATE 估计：匹配法、回归法
- 实践案例与工具（DoWhy）