流行病学与因果推断

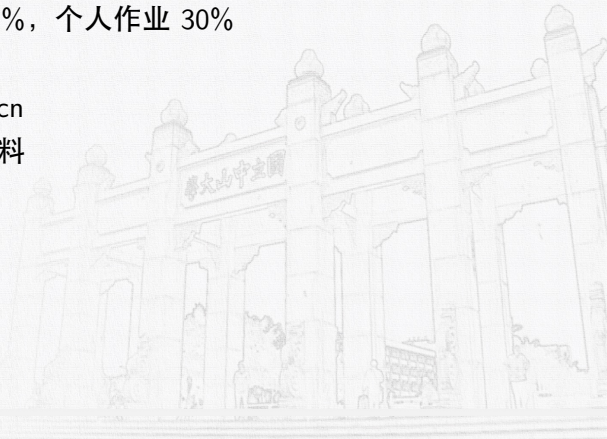
战义强

公共卫生学院（深圳）
中山大学

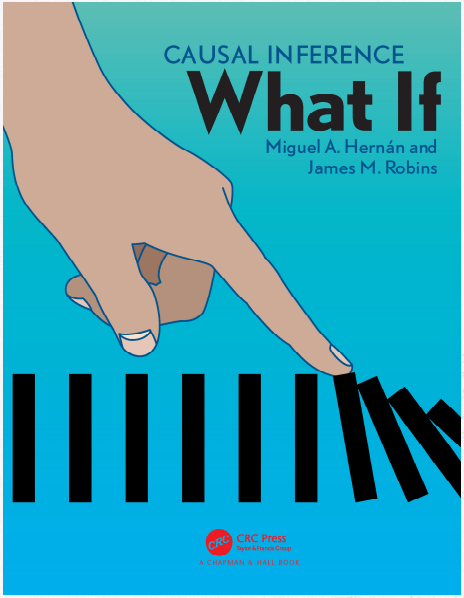
2025 年春

课程简介

- 36 学时，4 学时/周
- 考察：出勤 50%，2-3 人小组汇报 20%，个人作业 30%
- 理论 + 实践（R 软件）
- 助教：刘琦，liuq567@mail2.sysu.edu.cn
- 教材：无，每节课课后给大家参考资料



参考教材



pdf 免费下载

课程内容

- 流行病学定义、目的、因果推断的定义
- 混杂的定义和识别、RCT、工具变量、孟德尔随机化、倾向性评分
- 回归、双重差分、断点回归、固定效应模型、标准化方法
- 选择偏倚的识别、交互作用与效应修饰、测量误差
- 逆概率加权、生存分析的因果推断方法

- 什么是流行病学 Epidemiology

背景

- 什么是流行病学 Epidemiology
- 流行病学是研究人群中疾病与健康状况的分布及其影响因素，并研究防制疾病和促进健康的策略和措施的科学（詹思延，流行病学，第八版）

背景

- 什么是流行病学 Epidemiology
- 流行病学是研究人群中疾病与健康状况的分布及其影响因素，并研究防制疾病和促进健康的策略和措施的科学（詹思延，流行病学，第八版）
- 流行病学是研究人群中疾病与健康状态的分布及其影响因素，防制疾病的发生，促进人群健康的一门医学学科（徐飚，流行病学原理）

背景

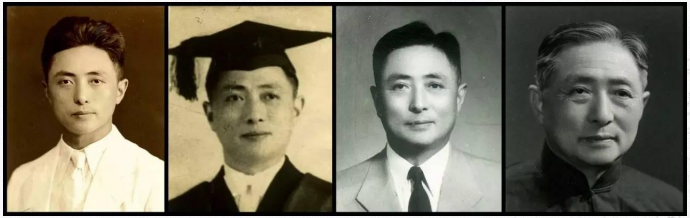
- 什么是流行病学 Epidemiology
- 流行病学是研究人群中疾病与健康状况的分布及其影响因素，并研究防制疾病和促进健康的策略和措施的科学（詹思延，流行病学，第八版）
- 流行病学是研究人群中疾病与健康状态的分布及其影响因素，防制疾病的发生，促进人群健康的一门医学学科（徐飚，流行病学原理）
- 流行病学是医学中的一门学科，它研究疾病的分布、生态学及防制对策。（苏德隆）

背景

- 什么是流行病学 Epidemiology
- 流行病学是研究人群中疾病与健康状况的分布及其影响因素，并研究防制疾病和促进健康的策略和措施的科学（詹思延，流行病学，第八版）
- 流行病学是研究人群中疾病与健康状态的分布及其影响因素，防制疾病的发生，促进人群健康的一门医学学科（徐飚，流行病学原理）
- 流行病学是医学中的一门学科，它研究疾病的分布、生态学及防制对策。（苏德隆）

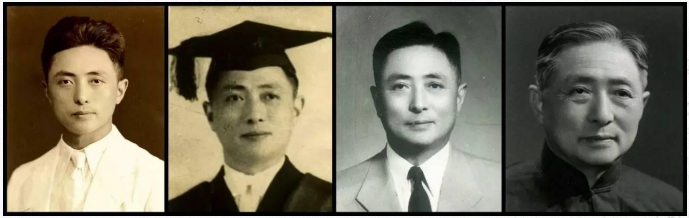
背景

- 什么是流行病学 Epidemiology
- 流行病学是研究人群中疾病与健康状况的分布及其影响因素，并研究防制疾病和促进健康的策略和措施的科学（詹思延，流行病学，第八版）
- 流行病学是研究人群中疾病与健康状态的分布及其影响因素，防制疾病的发生，促进人群健康的一门医学学科（徐飚，流行病学原理）
- 流行病学是医学中的一门学科，它研究疾病的分布、生态学及防制对策。（苏德隆）



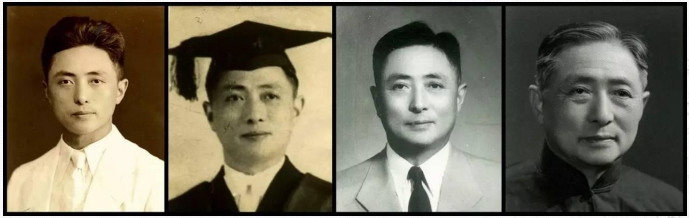
苏德隆 (1906 - 1985)

- 什么是流行病学 Epidemiology
- 流行病学是研究人群中疾病与健康状况的分布及其影响因素，并研究防制疾病和促进健康的策略和措施的科学（詹思延，流行病学，第八版）
- 流行病学是研究人群中疾病与健康状态的分布及其影响因素，防制疾病的发生，促进人群健康的一门医学学科（徐飚，流行病学原理）
- 流行病学是医学中的一门学科，它研究疾病的分布、生态学及防制对策。（苏德隆）



苏德隆 (1906 - 1985)
毛主席《送瘟神》：借问瘟君欲何往，纸船明烛照天烧

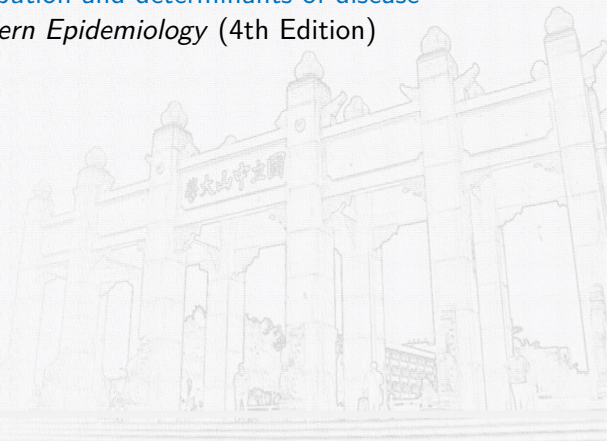
- 什么是流行病学 Epidemiology
- 流行病学是研究人群中疾病与健康状况的分布及其影响因素，并研究防制疾病和促进健康的策略和措施的科学（詹思延，流行病学，第八版）
- 流行病学是研究人群中疾病与健康状态的分布及其影响因素，防制疾病的发生，促进人群健康的一门医学学科（徐飚，流行病学原理）
- 流行病学是医学中的一门学科，它研究疾病的分布、生态学及防制对策。（苏德隆）



苏德隆 (1906 - 1985)
毛主席《送瘟神》：借问瘟君欲何往，纸船明烛照天烧

- Epidemiology is the study and analysis of the distribution (who, when, and where), patterns and determinants of health and disease conditions in defined populations. (Wikipedia)

- Epidemiology is the study and analysis of the distribution (who, when, and where), patterns and determinants of health and disease conditions in defined populations. (Wikipedia)
- Epidemiology is the study of the distribution and determinants of disease frequency in human populations. *Modern Epidemiology* (4th Edition)



- Epidemiology is the study and analysis of the distribution (who, when, and where), patterns and determinants of health and disease conditions in defined populations. (Wikipedia)
- Epidemiology is the study of the distribution and determinants of disease frequency in human populations. *Modern Epidemiology* (4th Edition)
- Epidemiology is the study of how disease is distributed in populations and the factors that influence or determine this distribution. *Gordis Epidemiology* (6th)

Prof. Leon Gordis (1934 - 2015)

流行病学的研究目的

流行病学的研究目的

- 描述 Description
  - 连续性变量：均数、标准差、众数等
  - 分类变量：患病率、发病率、死亡率、病死率等
- 预测 Prediction
  - 诊断
  - 预后
- 因果 Causation/Association (Counterfactual Prediction 反事实预测)
  - 治疗效果/副作用
  - 政策干预效果
  - 病因（危险因素）探索

$$X \qquad C \qquad Y$$

科学推理的逻辑

- 演绎推理 Deductive reasoning (上 $\rightarrow$ 下)

- 演绎推理 Deductive reasoning (上 → 下)
  - 所有的人都会死亡. (第一个前提假设)

科学推理的逻辑

- 演绎推理 Deductive reasoning (上 → 下)
  - 所有的人都会死亡. (第一个前提假设)
  - 苏格拉底是个人. (第二个前提假设)

科学推理的逻辑

- 演绎推理 Deductive reasoning (上 → 下)
  - 所有的人都会死亡. (第一个前提假设)
  - 苏格拉底是个人. (第二个前提假设)
  - 所以, 苏格拉底是会死亡的. (结论)

科学推理的逻辑

- 演绎推理 Deductive reasoning (上 $\rightarrow$ 下)
  - 所有的人都会死亡. (第一个前提假设)
  - 苏格拉底是个人. (第二个前提假设)
  - 所以, 苏格拉底是会死亡的. (结论)
- 归纳推理 (下 $\rightarrow$ 上)

科学推理的逻辑

- 演绎推理 Deductive reasoning (上 $\rightarrow$ 下)
  - 所有的人都会死亡. (第一个前提假设)
  - 苏格拉底是个人. (第二个前提假设)
  - 所以, 苏格拉底是会死亡的. (结论)
- 归纳推理 (下 $\rightarrow$ 上)
  - 这个公园里的乌鸦是黑色的

科学推理的逻辑

- 演绎推理 Deductive reasoning (上 $\rightarrow$ 下)
  - 所有的人都会死亡. (第一个前提假设)
  - 苏格拉底是个人. (第二个前提假设)
  - 所以, 苏格拉底是会死亡的. (结论)
- 归纳推理 (下 $\rightarrow$ 上)
  - 这个公园里的乌鸦是黑色的
  - 那个公园里的乌鸦是黑色的

科学推理的逻辑

- 演绎推理 Deductive reasoning (上 $\rightarrow$ 下)
  - 所有的人都会死亡. (第一个前提假设)
  - 苏格拉底是个人. (第二个前提假设)
  - 所以, 苏格拉底是会死亡的. (结论)
- 归纳推理 (下 $\rightarrow$ 上)
  - 这个公园里的乌鸦是黑色的
  - 那个公园里的乌鸦是黑色的
  - 所有乌鸦都是黑色的

科学推理的逻辑

- 演绎推理 Deductive reasoning (上 $\rightarrow$ 下)
  - 所有的人都会死亡. (第一个前提假设)
  - 苏格拉底是个人. (第二个前提假设)
  - 所以, 苏格拉底是会死亡的. (结论)
- 归纳推理 (下 $\rightarrow$ 上)
  - 这个公园里的乌鸦是黑色的
  - 那个公园里的乌鸦是黑色的
  - 所有乌鸦都是黑色的
- 流行病学是那种推理?

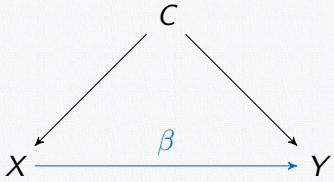
如何进行因果推断呢?

如何进行因果推断呢?

做实验/试验，最好是随机对照试验

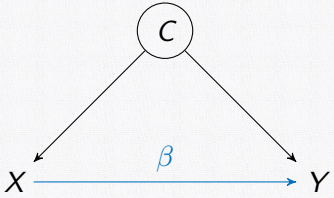
- 细胞层面：基因编辑等
- 动物层面：给药、食物等
- 人：临床试验、随机对照临床试验

观察性研究中的因果推断



- 我们希望，在 C 存在的情况下，能够估计 X 对 Y 的因果效应
- 我们假设 C 是足够的
- 然后我们使用可以考虑/控制 C 的统计方法：
  - 调整、控制、标准化
  - 倾向性评分
  - ...

然而



- 然而，通常我们不知道 C 是否是足够的
- 或者我们不愿意假设 C 是足够的
- 所以，存在一些尚未被测量的 C
- 然而，我们仍然希望能够估计 X 对 Y 的效应
- 怎么办?

在回答这个问题之前

统计学的关联

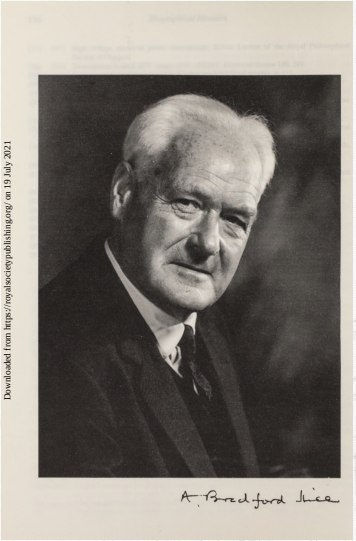
- 要得到因果效应，首先需得到统计学的关联
- 统计学上的关联意味着：两个因素倾向于一起出现
- 比如：吸烟的人通常会出现肺癌（与一般人相比）
- 在医院里面的人通常是患病的人（与在医院外面的人相比）

然而，相关并不意味着因果关系

- 除了因果关系之外，还有其他可以解释以下关系的原因吗?
- 吸烟和肺癌
- 在医院里面和患病

The Bradford Hill 因果推断准则 (viewpoints & perspectives) [1] (1965)

- Temporality 关联的时序性
- Strength 强度
- Consistency 一致性
- Specificity 特异性
- Dose-response relationship 剂量反应关系
- Plausibility 合理性
- Coherence 一致性
- Experimental evidence 实验证据
- Analogy 类推性



Sir Austin Bradford Hill
(8 July 1897 - 18 April 1991)



Sir Austin Bradford Hill (8 July 1897 - 18 April 1991)

因果效应的定义

- Hill 准则只是关于病因的一些可能的共同的特征（定性描述）
- 没有对因果效应进行定义
- 除了第一个“时序性”，病因可能不满足其余的任何条件
- 因此，我们需要对因果效应 causal effect，进行准确的定义

简史：1970s

Donald Rubin 对因果效应进行了正式的定义：潜在结果模型（反事实模型、虚拟现实模型、平行空间的另外一个场景）

More

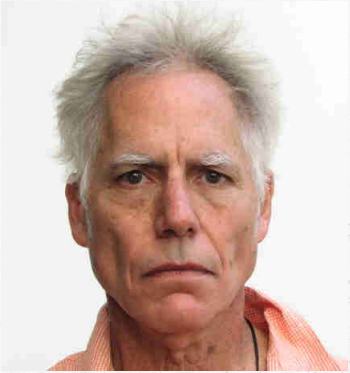

清華大學 新闻
Tsinghua University NEWS

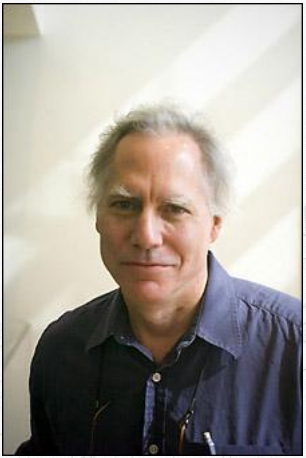

清華大學 新闻
Tsinghua University NEWS


新闻

简史：1980s

James Robins 研发了估计因果效应的模型：边际结构模型 Marginal Structural Models (MSMs)，特别适合于纵向数据/重复测量数据





简史, 1990s

Judea Pearl 研发了单向无环图 Directed Acyclic Graphs (DAGs): 极大的方便了观察性研究中因果推断的解释、变量的选择、以及与他人的交流和解释

潜在结果模型/反事实模型

- Brian MacMahon 和 Thomas Pugh: “...an association may be classed presumptively as causal when it is believed that, had the cause [exposure] been altered, the effect [outcome] would have changed.”
- 如果对暴露因素做出改变，结局也会发生变化，那么就可以把这个“关联”当作“因果关系”
- 所以，这个描述的关键点在于它的“反事实（平行空间的)”部分，如果（时间可以倒流）暴露因素不是现在我们已经观察到的这个暴露因素/水平，是其他，那么结局将会是什么?

潜在结果模型/反事实模型

- 我们要估计 X 对 Y 的效应
- 在潜在结果模型的框架下，用 Y_x 代表：如果 X 的值被设置成 x 时的量
- 如果暴露是二分类的，那么就有两个潜在的结果：Y_0, Y_1
- 对某一个个体来说，我们只能观察到其中的一个结果，这个结果就是当暴露因素是实际发生的时候，实际观察到的结果
- 如果某个个体的暴露水平是 $X = 0$，那么我们观察到了 Y_0，同时，我们没有观察到 Y_1，$Y_0 = Y$
- 如果某个个体的暴露水平是 $X = 1$，那么我们观察到了 Y_1，同时，我们没有观察到 Y_0，$Y_1 = Y$

潜在结果模型/反事实模型

- Y_0 和 Y_1 就是潜在的结果
- 理想状态下，我希望能够对所有人，都同时观察到 Y_0 and Y_1
- 例如：

Subject	Y_0	Y_1
id1	0	1
id2	0	0
id3	1	1

- 如果，对于某一个具体的个体来说，其潜在的结果不一致，Y_0 和 Y_1 不相等，那么，我们就说，X 对 Y 有因果效应
- id1, 暴露有因果效应 $Y_0 \neq Y_1$
- id2 和 id3, 暴露没有因果效应 $Y_0 = Y_1$

潜在结果模型/反事实模型

- 假设 id1 是暴露 ($X = 1$), 那么 $Y = Y_1$, Y_0 是反事实的
- 假设 id2 和 id3 是非暴露 ($X = 0$), 那么 $Y = Y_0$, Y_1 is 是反事实的

Subject	X	Y	Y_0	Y_1
id1	1	1	?	1
id2	0	0	0	?
id3	0	1	1	?

潜在结果模型/反事实模型

- 我们实际上对每一个研究对象真的观察到了“潜在结果”中的一个，这被称为“一致性假设 consistency assumption”；
- 用公式表述的话，一个实际暴露水平为 $X = x$, 那么他的实际观察到的结局与潜在的结局是相等的 $Y = Y_x$
- 我们定义个体的因果效应为两个潜在结果之间的差异 $Y_1 - Y_0$.
- 如果差值 $Y_1 - Y_0$ 不为 0，那么我们就说暴露对结局有作用
- 然而，我们通常观察不到某一个具体的个体的两个潜在结果，所以，我们在个体水平上无法计算个体的因果效应
- 然而，我们可能希望能够计算整个人群的平均的个体因果效应 $E(Y_1 - Y_0)$
- 即使我们能够计算整个人群的平均的因果效应 $E(Y_1 - Y_0)$，我们仍然需要其他的研究假设：我们可以假设“潜在结果”在不同的组别之间的分布是一致的（不同组别之间是可比的 comparable ）

潜在结果模型/反事实模型

- 如何估计人群的平均因果效应呢 $E(Y_1 - Y_0)$?
- 我们仅仅观察到了 $E(Y|X = 1) - E(Y|X = 0)$.
- 所以，我们需要其他额外的假设

潜在结果模型/反事实模型

- 在 RCT 中，所有处理/治疗之前（pre-treatment) 的变量都与处理/治疗措施 (treatment) 无关，例如：年龄、性别、遗传、环境等因素均与治疗本身无关；
- “潜在结果”(Y_0, Y_1) 是处理前变量，他们是“虚拟的/假设的”，不影响处理/治疗措施，也不受处理/治疗措施的影响；
- 因此，$(Y_0, Y_1) \perp X$
- 这意味着，

$$E(Y_0) = E(Y_0|X = 0) = E(Y_0|X = 1) \quad (1)$$
$$E(Y_1) = E(Y_1|X = 0) = E(Y_1|X = 1) \quad (2)$$

- 另外，基于“一致性的假设 consistency assumption”

$$E(Y_0) = E(Y_0|X = 0) = E(Y|X = 0) \quad (3)$$
$$E(Y_1) = E(Y_1|X = 1) = E(Y|X = 1) \quad (4)$$

潜在结果模型/反事实模型

- 这样的情况在暴露因素（包括治疗措施等）是随机分配的情况下是合适的；
- 但是如果是观察性研究或者非随机暴露，这样的情况通常是不太可能
- 在这样的情况下，我们可能会假设至少在某些已测量的层内 C，不同的暴露组之间的潜在结果是可比的/一样的
- 这样的假设被称为 “无混杂假设 unconfoundedness”
- 不同的学科有不同的名称：“no-unmeasured-confounding” assumption 或者 “exchangeability” assumption (流行病), an “ignorable treatment assignment” assumption (统计学), or an “exogeneity” assumption (经济学).

潜在结果模型/反事实模型

- 如果测量了足够的 Z, 每一个潜在结果 Y_x 都与 X 相互独立
- 这意味着，在每个 Z 的组合生成的层里，潜在结果 Y_x 与实际接受到的 X 没有关系
- 如果满足这样的假设的话，这两个组 $X = 1$ 和 $X = 0$ 的潜在结果 Y_0 是一致的；Y_0 的意思是，如果把暴露设置成 0，其所得到的潜在结果
- 与之类似，如果满足这样的假设的话，这两个组 $X = 1$ 和 $X = 0$ 的潜在结果 Y_1 是一致的；Y_1 的意思是，如果把暴露设置成 1，其所得到的潜在结果

- 我们用 $E(Y|z)$ 表示当 $Z = z$ 是 Y 的条件期望，意思是在亚组 $Z = z$ 中 Y 的平均值
- 如果没有未被测量的混杂因素，那么我们可以得到每一个层内的平均因果效应值:

$$
\begin{aligned}
E(Y_1 - Y_0|Z) &= E(Y_1|Z) - E(Y_0|Z) && (5)\\
&= E(Y_1|X = 1, Z) - E(Y_0|X = 0, Z) && (6)\\
&= E(Y|X = 1, Z) - E(Y|X = 0, Z) && (7)
\end{aligned}
$$

- 第二个等式：no-unmeasured-confounding assumption
- 第三个等式：consistency assumption

传统的方法与因果推断的方法

传统的方法与因果推断的方法

- 最近的这些年，很多复杂的因果推断技术相继出现，比如 marginal structural models, propensity scores, Instrumental variables
- 有时候，这些方法比传统的方法好 (e.g. logistic regression)
- 然而，大多数情况下，传统的统计学分析方法，从因果推断的角度来说，同样奏效