第一章　相关与辛普森悖论

罗姗姗

北京工商大学 数学与统计学院

因果推断课题组

目录

关于因果的思考

- 古希腊哲学家 Democritus (约公元前 400 年) 认为:

“发现一个因果关系胜过做国王。”

- 贝叶斯网络的创始人、2011 年图灵奖获得者 Judea Pearl 教授:

“传统的机器学习方法难以突破“弱”人工智能的瓶颈，因此我们期待通过因果推理的方法，从因果关系的角度而不仅仅是数据拟合的角度来进行人工智能研究。”

因果推断已成为国际研究热点


中华人民共和国中央人民政府
www.gov.cn
首页 | 繁体 | 英文EN | 登录 |

首页 > 信息公开 > 国务院文件 > 科技、教育 > 科技　　字号：默认 大 超大 | 打印 收藏 留言 |

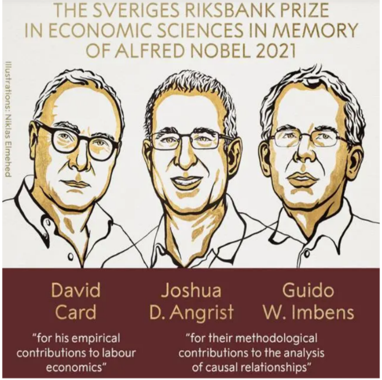
索 引 号：	000014349/2017-00142	主题分类：	科技、教育\科技
发文机关：	国务院	成文日期：	2017年07月08日
标　　题：	国务院关于印发新一代人工智能发展规划的通知		
发文字号：	国发〔2017〕35号	发布日期：	2017年07月20日

国务院关于印发
新一代人工智能发展规划的通知
国发〔2017〕35号

各省、自治区、直辖市人民政府，国务院各部委、各直属机构：
现将《新一代人工智能发展规划》印发给你们，请认真贯彻执行。

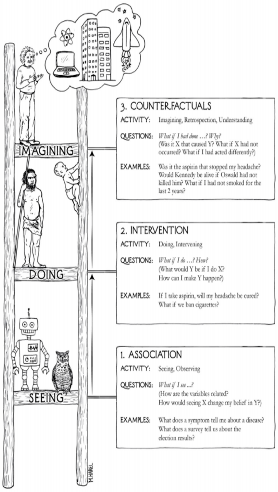
国务院
2017年7月8日

(a) 重点突破直觉推理与因果模型等基础理论瓶颈



(b) 现代经济学的因果推断革命 2021 年诺贝尔经济学奖

因果推断的三个层级



1. 第一层级为：预测；第二层级为：干预；第三层级为：反事实；
2. 目前绝大部分因果推断研究集中于第二层级，研究因果效应评价的问题；
3. 第三层级的因果推断与归因问题有关；

因果问题与工具

- 关于因果关系的相关问题：
  - 因果关系的哲学意义？
  - 因果作用，原因的发生对结果有何影响？(Effect of Cause).
  - 归因，一个结果发生的原因是什么？(Cause of Effect).

  对于评价一个原因的因果作用，目前的理论与实践都比较成熟，已有一套行之有效的统计推断方法．在本课程中，我们将重点放在评估因果作用 (Effect of Cause).

- 需要的工具：数理统计基础，编程基础

一些应用

Z: 原因变量 (如干预，处理，暴露) 为了说明，我们将主要关注二值原因

Y: 结果变量 (例如疾病状况)

X: 可观测的协变量或可观测的混杂

U: 未观测的协变量或未观测的混杂

在随机化或观察试验中，我们将有以下应用：

1. 生物医学领域：研究暴露 Z 对疾病 Y 的因果影响
2. 流行病学领域：评估接种疫苗 Z 对流感 Y 的影响
3. 经济学和政策领域：项目评估和政策影响分析

因果与相关

- 探索因果关系的研究推动着统计科学的发展.
- 统计学家提出了各种相关关系的形式化度量，并根据这些相关关系进行一些预测分析
- 在大多数情况下，相关关系并不能表示因果关系（Yule-Simpson Paradox）.
- 因果分析是更进一步的评价，它是一个关于反事实结果的预测，即：

  如果同一个体/受试者暴露在不同（反事实）条件下会发生什么？
- 因果推断的本质困难：

  任何人不可能同时两次踏入相同的河。

Outline

相关性与回归 I

- 两个随机变量 Z 和 Y 之间的 Pearson 相关系数为：

$$\rho_{ZY} = \frac{\operatorname{cov}(Z, Y)}{\sqrt{\operatorname{var}(Z)\operatorname{var}(Y)}}$$

该系数用于衡量 Z 和 Y 之间的线性依赖关系。

- 线性回归模型是关于 Y 对 Z 的模型：

$$Y = \alpha + \beta Z + \varepsilon$$

其中 $E(\varepsilon) = 0$ 且 $E(\varepsilon Z) = 0$。我们可以证明回归系数 β 等于

$$\beta = \frac{\operatorname{cov}(Z, Y)}{\operatorname{var}(Z)} = \rho_{ZY}\sqrt{\frac{\operatorname{var}(Y)}{\operatorname{var}(Z)}}$$

所以 β 和 ρ_{ZY} 总是具有相同的符号。

相关性与回归 II

- 我们还可以定义多元回归，例如 Y 对 Z 和 X 的回归模型：

$$Y = \alpha + \beta Z + \gamma X + \varepsilon$$

其中 $E(\varepsilon) = 0$，$E(\varepsilon Z) = 0$ 以及 $E(\varepsilon X) = 0$。

- 通常我们将 β 解释为在给定 X 下、在 X 条件下、或在 X 控制下，Z 对 Y 的影响。

列联表与相关性 I

- 我们可以通过一个 2×2 的列联表来表示两个二元随机变量 Z 和 Y 的联合分布。假设 $p_{zy}=\mathrm{pr}(Z=z,Y=y)$，我们可以总结联合分布如下表所示：

	$Y=1$	$Y=0$
$Z=1$	p_{11}	p_{10}
$Z=0$	p_{01}	p_{00}

- 将 Z 视为处理或原因变量，Y 视为结果变量，我们可以定义风险差异（Risk Difference）为

$$\begin{aligned}\mathrm{RD} &= \mathrm{pr}(Y=1\mid Z=1)-\mathrm{pr}(Y=1\mid Z=0)\\ &= \frac{p_{11}}{p_{11}+p_{10}}-\frac{p_{01}}{p_{01}+p_{00}}\end{aligned}$$

列联表与相关性 II

- 危险比（Risk Ratio）定义为

$$\begin{aligned}\mathrm{RR} &= \frac{\mathrm{pr}(Y=1 \mid Z=1)}{\mathrm{pr}(Y=1 \mid Z=0)} \\ &= \frac{p_{11}}{p_{11}+p_{10}} \Big/ \frac{p_{01}}{p_{01}+p_{00}}\end{aligned}$$

- 优势比（Odds Ratio）定义为

$$\begin{aligned}\mathsf{OR} &= \frac{\mathrm{pr}(Y=1 \mid Z=1)/\,\mathrm{pr}(Y=0 \mid Z=1)}{\mathrm{pr}(Y=1 \mid Z=0)/\,\mathrm{pr}(Y=0 \mid Z=0)} \\ &= \frac{\frac{p_{11}}{p_{11}+p_{10}} \big/ \frac{p_{10}}{p_{11}+p_{10}}}{\frac{p_{01}}{p_{01}+p_{00}} \big/ \frac{p_{00}}{p_{01}+p_{00}}} \\ &= \frac{p_{11}p_{00}}{p_{10}p_{01}}.\end{aligned}$$

列联表与相关性 III

- 风险差异、危险比和优势比这些术语来自流行病学。因为流行病学中的结果通常是疾病，因此将“风险”这个名称用于患病的概率是自然的。

- 关于这些度量的一些简单事实, 以下陈述都是等价的::
  1. $Z \perp\!\!\!\perp Y$，$\mathrm{RD} = 0$，$\mathrm{RR} = 1$ 和 $\mathrm{OR} = 1$。
  2. 如果 p_{zy} 都是正数，那么 $\mathrm{RD} > 0$ 等价于 $\mathrm{RR} > 1$，也等价于 $\mathrm{OR} > 1$。
  3. 如果 $\mathrm{pr}(Y = 1 \mid Z = 1)$ 和 $\mathrm{pr}(Y = 1 \mid Z = 0)$ 都很小，那么 $\mathrm{OR} \approx \mathrm{RR}$。

Outline

Outline

何为混杂？

- 因果关系与相关关系的本质区别：混杂因素
- 与处理和结果变量都相关（原因和结果的共同原因）的背景变量称为混杂因素 (confounder or confounding factor).
- 观察性研究中不可避免地存在一些混杂变量未被观测，即未观测混杂 (unobserved confounder).
- 观察性研究中可观测的混杂变量.

$$Z \longrightarrow Y$$

(a) 因果关系

$$\begin{tikzcd} & \text{混杂} \arrow[dl] \arrow[dr] & \\ Z \arrow[rr] & & Y \end{tikzcd}$$

(b) 存在混杂

经典的肾结石例子 I

- 本示例源自Charig et al. (1986)，其中 Z 是治疗方法，1 表示开放手术，0 表示小刺穿；Y 是结果，1 表示成功，0 表示失败。原因和结果数据可以总结如下 2×2 表格：

	$Y=1$	$Y=0$
$Z=1$	273	77
$Z=0$	289	61

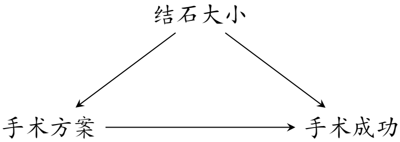
- 估计的风险差异为：

$$
\begin{aligned}
\widehat{\mathrm{RD}} &= \widehat{\mathrm{pr}}(Y=1 \mid Z=1) - \widehat{\mathrm{pr}}(Y=1 \mid Z=0) \\
&= \frac{273}{273+77} - \frac{289}{289+61} = 78\% - 83\% = -5\% < 0.
\end{aligned}
$$

看起来，手术方法 0 更好，即小刺穿相比于开放手术具有更高的成功率。

经典的肾结石例子 II

- 但是，这些数据并不是来自随机对照试验（RCT）。
- 接受手术方案 1 的患者可能与接受手术方案 0 的患者非常不同。
- 在这项研究中的一个“隐变量”是病例的严重程度：一些患者的结石较小，而一些患者的结石较大。
- 结实的大小即会影响病人接受不同的手术，也会影响病人的最终结果是否成功，从而是一个混杂因素。



经典的肾结石例子 III

- X 是二值示性变量，$X = 1$ 表示小结石，$X = 0$ 表示大结石。

	$Y = 1$	$Y = 0$
$Z = 1$	81	6
$Z = 0$	234	36

(a) 结石较小的患者,$X = 1$

	$Y = 1$	$Y = 0$
$Z = 1$	192	71
$Z = 0$	55	25

(b) 结石较大的患者,$X = 0$

	$Y = 1$	$Y = 0$
$Z = 1$	273	77
$Z = 0$	289	61

(c) 所有患者

经典的肾结石例子 IV

- 根据小结石患者的表格，估计的风险差异为

$$\begin{aligned}\widehat{\mathrm{RD}}_{\text{smaller}} &= \widehat{\mathrm{pr}}(Y=1 \mid Z=0, X=1) - \widehat{\mathrm{pr}}(Y=1 \mid Z=0, X=0) \\ &= \frac{81}{81+6} - \frac{234}{234+36} = 93\% - 87\% = 6\% > 0,\end{aligned}$$

  表明手术方案 1 更好。

- 对于大结石患者的表格，估计的风险差异为

$$\begin{aligned}\widehat{\mathrm{RD}}_{\text{larger}} &= \widehat{\mathrm{pr}}(Y=1 \mid Z=0, X=1) - \widehat{\mathrm{pr}}(Y=1 \mid Z=0, X=0) \\ &= \frac{192}{192+71} - \frac{55}{55+25} = 73\% - 69\% = 4\% > 0,\end{aligned}$$

  也表明手术方案 1 更好。

- 上述结论表明 $\widehat{\mathrm{RD}} < 0, \widehat{\mathrm{RD}}_{\text{smaller}} > 0, \widehat{\mathrm{RD}}_{\text{larger}} > 0.$

经典的肾结石例子 V

	$Y=1$	$Y=0$
$Z=1$	$n_{11\|1}$	$n_{10\|1}$
$Z=0$	$n_{01\|1}$	$n_{00\|1}$

(a) 结石较小的患者,$X=1$

	$Y=1$	$Y=0$
$Z=1$	$n_{11\|0}$	$n_{10\|0}$
$Z=0$	$n_{01\|0}$	$n_{00\|0}$

(b) 结石较大的患者,$X=0$

	$Y=1$	$Y=0$
$Z=1$	n_{11}	n_{10}
$Z=0$	n_{01}	n_{00}

(c) 所有患者

- 纯数学的角度，上面的悖论可以写成初等数学

$$\frac{n_{11|1}}{n_{11|1}+n_{10|1}} > \frac{n_{01|1}}{n_{01|1}+n_{00|1}}, \quad \frac{n_{11|0}}{n_{11|0}+n_{10|0}} > \frac{n_{01|0}}{n_{01|0}+n_{00|0}}$$

$$\frac{n_{11}}{n_{11}+n_{10}} = \frac{n_{11|1}+n_{11|0}}{n_{11|1}+n_{10|1}+n_{11|0}+n_{10|0}} < \frac{n_{01|1}+n_{01|0}}{n_{01|1}+n_{00|1}+n_{01|0}+n_{00|0}} = \frac{n_{01}}{n_{01}+n_{00}}.$$

- 在统计上，这具有重要的意义，即变量之间的相关关系可以完全的被第三个变量“扭曲”，忽略潜在的“第三个变量”可能改变已有的结论。

经典的肾结石例子 VI

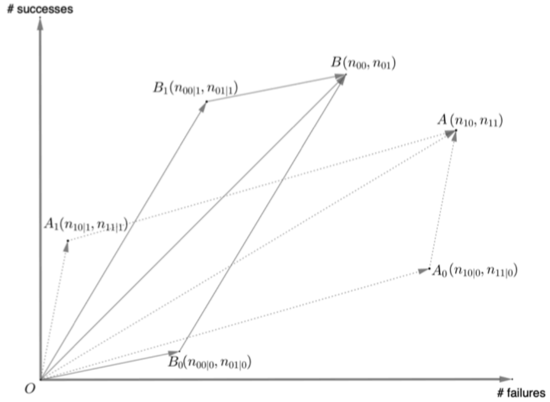


FIGURE 1.2: Geometry of the Yule–Simpson Paradox

References I

Charig, C. R., Webb, D. R., Payne, S. R., and Wickham, J. E. (1986). Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *British Medical Journal*, 292:879–882.

作业

- 阅读教材 P3-12。
- 上机作业：运行 1.4 节代码。
- 生活中存在辛普森悖论的例子。