因果推断——现代统计的思想飞跃

丁 鹏

1 引言

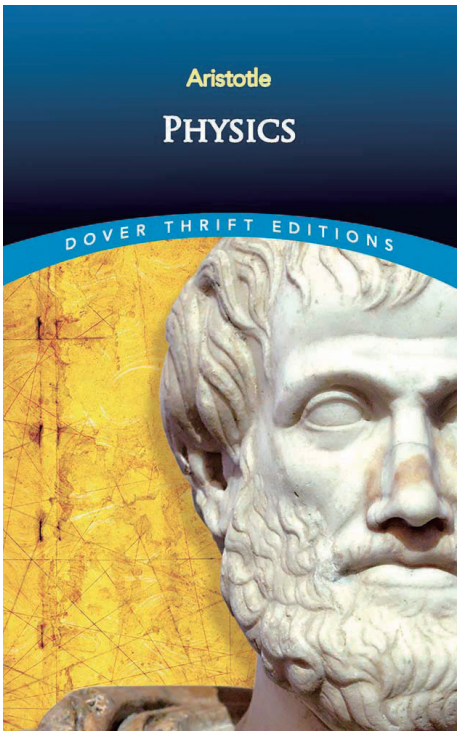
探求事物的原因，是人类永恒的精神活动之一。从古希腊的哲学到中国先秦的诗歌，都充满了对原因的追问和对因果关系的思考。比如，亚里士多德就在《物理学》（*Physics*）和《形而上学》（*Metaphysics*）两书中反复强调，我们只有知道了事物的原因，才能算真正理解这个事物。又如，屈原在《天问》开篇，就追问日月星辰运行的原因。

长期以来，人们一方面好奇地追问原因和结果的关系，一方面又苦于这些概念的模糊性。于是，这些话题在很长一段时间都仅仅局限在哲学和文学的范围内。精确地描述因果关系，尤其是用数学的语言来描述因果关系，则是非常近代的事情了。这一项思想飞跃，得益于现代统计学的发展。统计学家称之为“因果推断”（causal inference）。虽然因果推断在现代统计学的萌芽阶段就已经产生，但是它的发展并非一帆风顺：它长期被主流忽视、怀疑甚至攻击。直至最近四十年，尤其是最近十年，它才得到了广泛的认可和大力的研究，成为当今主流的研究方向之一。在最近的一篇文章中，Andrew Gelman 和 Aki Vehtari 评选了过去五十年中，统计学最重要的八个想法，排名第一的就是因果推断[1]。当今世界，很多年轻的学者加入了因果推断的研究，他们来自统计学、经济学、社会学、政治科学、教育学、流行病学、计算机科学、哲学等等领域。毫不夸张地说，统计因果推断的研究迎来了它发展的黄金时代。

本文将回顾统计因果推断的历史背景，评述中国因果推断研究的现状，并且大胆推测它未来的发展前景。

[1] A. Gelman and A. Vehtari, What are the most important statistical ideas of the past 50 years? 见 https://arxiv.org/abs/2012.00174。第一作者曾获得年轻统计学家的最高奖 COPSS 奖章。

2 哲学基础：因果推断何以成为可能？



亚里士多德《物理学》的一个英译本。这本书的 Book II 3 的开篇写道："Knowledge is the object of our inquiry, and men do not think they know a thing till they have grasped the 'why' of it (which is to grasp its primary cause)"，翻译成中文就是，我们探索的目标是知识，只有掌握了"为什么"，才算真正理解一个事物，即，掌握该事物的根本原因。

人们常常问关于原因和结果的问题。比如，某人死于肺癌，是不是因为他常常吸烟导致的？比如，我感冒症状减轻了，是不是因为服用了维生素 C 片导致的？比如，大学教育是否能够提高收入水平？类似的问题，充满了我们的日常生活。

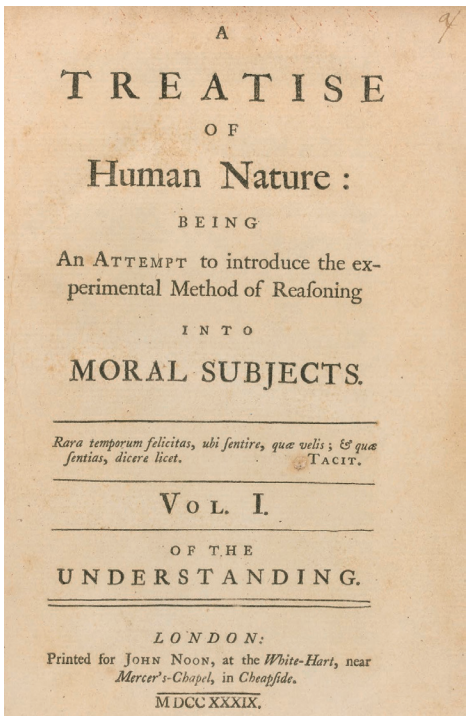
但是，这些看似直接了当的问题，却不容易回答。比如，有人吸烟，却没有得肺癌；有人不吸烟，却得了肺癌。比如，我可能仅仅喝白开水，感冒也会自己消失。比如，有人没有上大学，却做生意发了大财。当然，有点概率论常识的人很容易意识到，这些事件都带有随机性。从经验中，我们可能观察到吸烟的人更可能得肺癌；服用维生素 C 的人，平均来说，自我感觉感冒恢复得更快；上过大学的人平均收入更高。但是，这些统计上的"相关关系"是否就是"因果关系"呢？

大部分西方哲学家都认为因果关系是一条本质的、似乎毋庸置疑的定律。但是，苏格兰哲学家大卫·休谟（David Hume, 1711-1776）曾经抛出了一条惊人的论点。简言之，他认为人类仅仅凭经验，只能认识事物之间恒定的前后相继关系（constant conjunction），并不能认识任何因果关系。很多哲学家都努力回应休谟的质疑，因为若是承认休谟是对的，那么知识何以成为可能？若人类的知识仅仅是经验性的前后相继关系，那么人类似乎没有拥有任何"心智的荣耀"[2]。

[2] "[T]he sole end of science is the honor of the human mind." —— Carl Jacobi（卡尔·雅可比）

哲学家们对休谟的回应似乎都是徒劳的。我在学生时代曾经上过邓晓芒教授“康德哲学”的课，他就直言，休谟是驳不倒的。的确，休谟这样的彻底的怀疑论者，是无法驳倒的。我回顾休谟的高论，并非想卖弄哲学史，因为休谟是绕不开的：无论何时何地，只要谈及因果推断，就可能有人引用休谟的论点质疑你问题的合理性。也正是因为休谟这种近乎诅咒似的言论，使得因果推断的数学化步履维艰。

然而，上个世纪统计学的几项辉煌成果改写了思想史。如今人们已经不再羞于讨论因果关系，统计因果推断的语言，深入到了几乎所有的应用领域。这些成果也许并没有完全解决休谟的问题，但是它们给出了因果关系新的思考方式和推理框架。下面，我将分三部分回顾历史。


A

TREATISE

OF

Human Nature:

BEING

An ATTEMPT to introduce the experimental Method of Reasoning

INTO

MORAL SUBJECTS.

Rara temporum felicitas, ubi sentire, quæ velis; & quæ sentias, dicere licet. TACIT.

VOL. I.

OF THE

UNDERSTANDING.

LONDON:
Printed for JOHN NOON, at the *White-Hart*, near *Mercer's-Chapel*, in *Cheapside*.
MDCCXXXIX.

休谟的名著《人性论》对哲学史产生了深远的影响，他指出了归纳推理的缺陷，认为我们对因果关系的信念仅仅来自于习惯（habit）和传统（custom）。

3 统计学中“哥白尼式的革命”：内曼的“潜在结果”模型

1923 年，耶日 • 内曼（Jerzy Neyman, 1894-1981）还是波兰华沙大学的博士生，他的毕业论文是“概率论在农业实验中的应用”[3]。在这篇论文中，他提出了用于因果推断的“潜在结果”（potential outcomes）的数学模型，并将它和统计推断结合起来。他的想法非常自然，数学结构也很简单。下面简单地回顾一下。

以农业实验为例，考虑 n 块田作为实验的对象，实验者想检测两种肥料对于产量的影响。用 i 表示第 i 块田，$Y_i(1)$ 和 $Y_i(0)$ 表示如果用肥料 1 和肥料 0 分别对应的第 i 块田的产量，那么 $\tau_i = Y_i(1) - Y_i(0)$ 就是肥料 1 相对于肥料 0 对第 i 块田产量的因果作用。实验者随机地分配肥料 1 或者肥料 0 到第 i 块田，

[3] 内曼的论文是用波兰语写成的。1990 年，D. M. Dabrowska 和 T. P. Speed 将论文翻译成英文，题目是 On the Applications of the Theory of Probability to Agricultural Experiments，发表于 *Statistical Science*。潜在结果的基本想法也许在历史中早就产生了，但是将它数学化、且正式地用于统计学，内曼的文章是首次。内曼是现代统计学的奠基人之一，他对假设检验、置信区间、抽样调查和实验设计等领域的研究，成为现代统计学的标准范式。我国概率论和数理统计学的先驱许宝騄教授是内曼在英国指导的学生之一。

年轻时的内曼。内曼是加州大学伯克利分校统计系的创始人（照片由该系提供）。

所以最终我们要么观测到 $Y_i(1)$，要么观测到 $Y_i(0)$，不可能同时观测两者。显而易见，在这个模型下，因果推断的本质困难就是无法同时观测 $Y_i(1)$ 和 $Y_i(0)$，也就无法直接观测到 τ_i。观测单个的 τ_i 太困难，退而求其次，我们可以考虑研究它的平均数：

$$\tau = \frac{1}{n}\sum_{i=1}^{n}\tau_i.$$

这个 τ 通常被称为平均因果作用（average causal effect）。这可能是因果作用最简单的定义了。到此为止，内曼引入了一些数学记号来定义“因果作用”。也许读者会觉得这平平无奇，无非就是 $Y_i(1)$ 和 $Y_i(0)$。但是，这些记号将开启一扇窗，迎接新思想的曙光。

潜在结果 $Y_i(1)$ 和 $Y_i(0)$，以及平均因果作用 τ，在某种意义上，都是假想的数字。仅有这些定义，还不能说明这个模型的现实意义。问题的关键是：我们能否根据观测到的数据推断 τ？内曼给出了肯定的回答。在随机化实验下，第 i 块田接受肥料 1 或者肥料 0 是完全随机的。用 $Z_i = 1$ 表示第 i 块田接受肥料 1，用 $Z_i = 0$ 表示第 i 块田接受肥料 0。随机化实验固定接受肥料 1 和肥料 0 的田的总数，分别是 n_1 和 n_0，对应的 $(Z_1, \ldots, Z_n)$ 这个向量是 n_1 个 1 和 n_0 个 0 的随机置换（random permutation）。如果第 i 块田接受了肥料 Z_i，那么我们观测到的产量就是

$$Y_i = Y_i(Z_i) = Z_iY_i(1) + (1 - Z_i)Y_i(0).$$

这个恒等式似乎显而易见：从数学上讲，它无非说明，当 $Z_i = 1$ 时，$Y_i = Y_i(1)$；当 $Z_i = 0$ 时，$Y_i = Y_i(0)$。但是，我在和朱迪亚·珀尔（Judea Pearl）交流时，他认为这是因果推断最重要的恒等式，因为它联系了左边我们能够观测到的结果和右边的潜在结果。

最终能够被观测的数据是 $\{(Z_i, Y_i) : i = 1, \ldots, n\}$。一个显而易见的估计量是

$$\hat{\tau} = \frac{1}{n_1}\sum_{i:Z_i=1} Y_i - \frac{1}{n_0}\sum_{i:Z_i=0} Y_i.$$

它是接受肥料 1 和肥料 0 下，平均结果的差值。内曼证明了 $\hat{\tau}$ 是平均因果作用 τ 的无偏估计（即 $\hat{\tau}$ 的期望是 τ），计算了这个估计量的方差，讨论了如何估计这个方差，还提出了一个基于 $\hat{\tau}$ 的中心极限定理的置信区间（即这个区间以指定的概率盖住真值 τ）。最后一步的中心极限定理在内曼的原文仅仅是一个直

觉的证明，一直到了 Paul Erdős，Alfréd Rényi 和 Jaroslav Hájek 工作的出现，这类中心极限定理的证明才被严格化[4]。

上面仅仅讨论了一个最简单的数学结构：两个组的随机化实验中的因果推断。现实中的随机化实验丰富多彩，如何在各种随机化实验中做因果推断取决于具体的实验设计方案。内曼本人于 1935 年在英国皇家统计学会宣读的论文，讨论了随机区组设计（randomized block design）和拉丁方设计（Latin squares design）的因果推断，引发了包括罗纳德 • 费希尔（Ronald Fisher）在内的统计学家的激烈争论。同时期，费希尔对随机化实验进行了深入的研究，虽然他没有使用内曼潜在结果的记号，但是因果推断始终是他思考的对象。随后的几十年，随机对照实验（randomized controlled trial；RCT）成为美国食品药品监督管理局批准新药的黄金标准。最近二十年，大量的随机化实验出现在社会科学中，用来研究复杂社会问题中的因果关系。比如，麻省理工学院和哈佛大学的三位经济学家，Abhijit Banerjee，Esther Duflo 和 Michael Kremer，因为用实验的方法研究发展经济学，获得了 2019 年的诺贝尔经济学奖。

内曼生前对自己在统计假设检验方面的奠基性工作颇为自豪，认为那是统计学中“哥白尼式的革命”（Copernican Revolution）[5]。他并未预料他在因果推断的奠基性工作，也将产生深远的影响。这个影响则是由唐纳德 • 鲁宾（Donald Rubin）开启的。

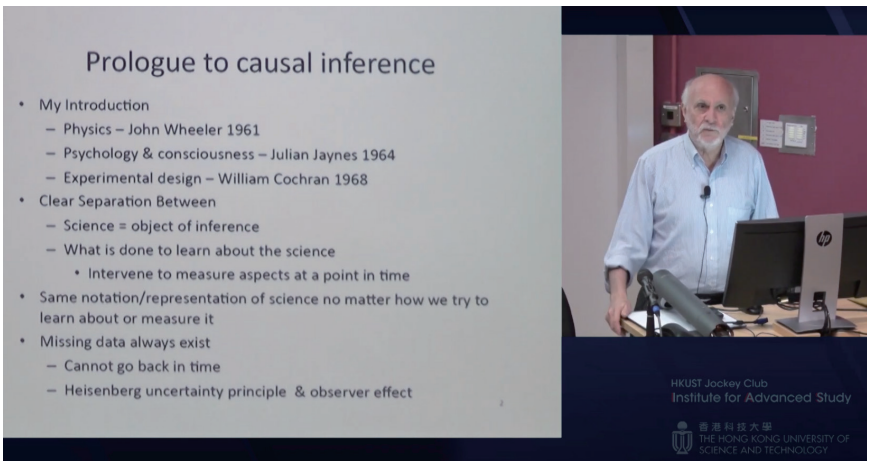
4 统计学的拓荒者：鲁宾关于观察性研究中的因果推断的研究

从直觉上，也许大家不会对随机化实验中的因果推断感到惊奇。毕竟随机化实验保证了两个组在平均意义下是相似的，那么他们之间的区别就可以归因于不同肥料对产量的因果作用。但是，现实的统计问题，很多数据收集并非源自随机化实验——这类研究通常被称为观察性研究（observational study）。比如，如果要研究吸烟和肺癌的因果关系，基本的伦理不允许我们随机地让一部分人抽烟、让一部分人不抽烟。再如，研究大学教育对收入的影响，我们不能随机地让一部分人上大学、让一部分人不上大学。很多流行病学和社会科学的问题，本质上一定是观察性研究，人们也迫切地想从这些观察性研究中获得关于因果关系的知识。

虽然潜在结果模型成功地数学化了随机化实验中的因果推断，但是它长期并未用于观察性研究——内曼本人是持怀疑态度的，因为缺乏随机化，观察性研究有太多复杂性，比如抽烟的人和不抽烟的人，可能就是两群完全不同的人，不具有可比性。虽然他从未尝试用他的潜在结果模型分析观察性数据，但是他

[4] 这方面的文献综述是：Li, X. and Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112, 1759-1769。

[5] 见内曼的传记：C. Reid (1982), Neyman - From Life。注意，哥白尼和内曼都是波兰人。



鲁宾教授正在作报告（截屏自 https://www.youtube.com/watch?v=N4tQC3elGK4）

间接地启发了一些更加有冒险精神的学者。其中一人就是鲁宾[6]。

鲁宾认为，观察性研究也对应着一个假想的随机化实验，因此内曼的潜在结果模型可以用来定义一般的因果作用。这里我们考虑一般的问题，不再局限于农田、肥料和产量。用 i 表示个体 i，它的观测结果 Y_i 有两个潜在结果 $Y_i(1)$ 和 $Y_i(0)$，分别对应两个处理水平，一般来说 1 被称为“处理”（treatment），而 0 被称为“对照”（control）。每个个体 i 有一个二值的处理水平 Z_i 和一些处理前的协变量 X_i。一个具体的例子是：

- Z_i: 个体 i 吸烟与否的指示变量；
- Y_i: 个体 i 是否得肺癌的指示变量；
- X_i: 个体 i 的年龄、性别、教育、收入、家庭病史等等，统计学中称它们为协变量（covariates）。

假设 $\{(Z_i, X_i, Y_i(1), Y_i(0)) : i = 1, \ldots, n\}$ 是独立同分布的随机采样而来，我们关心的参数是如下的总体平均因果作用：

$$\tau = E\{Y_i(1) - Y_i(0)\}.$$

鲁宾给了一个关于 τ 的因果推断的充分条件：

给定协变量 X_i，潜在结果 $(Y_i(1), Y_i(0))$ 和处理变量 Z_i 条件独立。

鲁宾称这个条件为“可忽略性”（ignorability）。这个条件还有很多其他名字：流行病学家常常称之为“无混杂性”（unconfoundedness）；经济学家常常称之为“可观测的选择机制”（selection on observables）。在可忽略性下，我们可

[6] 另外一位受内曼影响的是计量经济学家 Trygve Haavelmo。他是在计量经济学中讨论因果推断的先驱。他曾在 1989 年诺贝尔经济学奖的获奖感言中谈及内曼对他的影响：https://www.nobelprize.org/prizes/economic-sciences/1989/haavelmo/facts/。

以通过简单的数学推导得到下面的结果：

$$\begin{aligned}\tau &= \sum_x \{E(Y_i(1)\mid X_i=x) - E(Y_i(0)\mid X_i=x)\}P(X_i=x)\\ &= \sum_x \{E(Y_i(1)\mid Z_i=1, X_i=x) - E(Y_i(0)\mid Z_i=0, X_i=x)\}P(X_i=x)\\ &= \sum_x \{E(Y_i\mid Z_i=1, X_i=x) - E(Y_i\mid Z_i=0, X_i=x)\}P(X_i=x).\end{aligned}$$

为了简单起见，上面的公式假设 X 是离散的随机变量；一般化的公式可以同理得到。上面的推导仅仅用到了最基本的概率法则：第一步是全概率公式；第二步由可忽略性要求的条件独立性得到；第三步根据 Z_i 将 Y_i 替换成 Y_i (1) 或者 Y_i (0)。这个公式的意义在于，最左边的平均因果作用 τ 的定义依赖于不可以完全被观测的潜在结果，最右边的量仅仅依赖于可以观测的变量 (Z_i, X_i, Y_i) 的联合分布。用一个技术性的术语来描述上面的公式，就是，基于观测数据，平均因果作用是可识别的（identifiable）。直观上，我们可以用观测数据构造平均因果作用的估计量。比如，我们可以拟合 Y_i 关于 (Z_i, X_i) 的统计模型，则可以进一步根据上面的公式估计 τ .

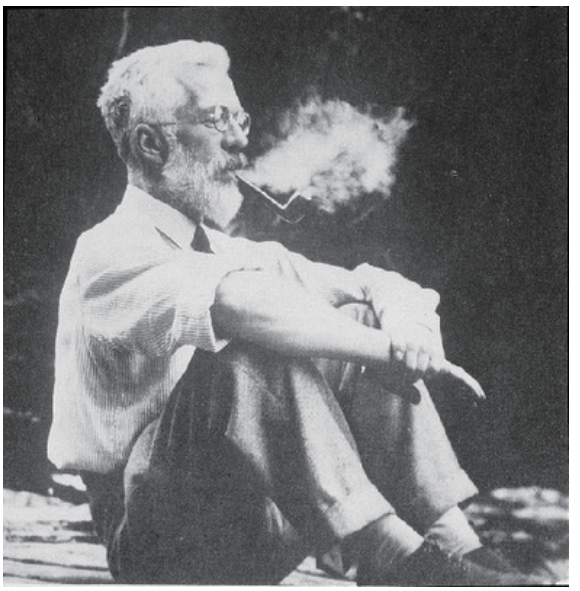
我们还可以证明如下的公式：

$$\tau = E\left\{\frac{Z_iY_i}{e(X_i)}\right\} - E\left\{\frac{(1-Z_i)Y_i}{1-e(X_i)}\right\},$$

其中 $e(X_i) = P\ (Z_i = 1 \mid X_i)$ 是处理的指示变量给定协变量的条件概率。这个公式也有比较直观的解释：处理组和对照组的个体并非完全随机选择的，我们需要根据他们入组的概率进行调整。Paul Rosenbaum 和鲁宾在他们 1983 年的 *Biometrika* 文章中指出，$e(X_i)$ 在观察性研究的因果推断中，发挥着至关重要的作用，他们把这个条件概率称为“倾向得分”（propensity score）。这个公式有类似的、不平凡的意义：右边的量仅仅依赖于可以观测的变量 (Z_i, X_i, Y_i) 的联合分布。一旦拟合了 Z_i 关于 X_i 的统计模型，我们可以得到 $e(X_i)$ 的估计，则可以进一步估计 τ。这个估计方法涉及到了用条件概率的逆进行加权，所以在文献中它也被称为“逆概加权”（inverse probability weighting；IPW）。

Rosenbaum 和鲁宾的这篇文章是 *Biometrika* 这个杂志创刊以来引用率最高的两篇文章之一[7]。在它发表后的三十多年里，引起了很多理论统计学家和

[7] 文章是 Rosenbaum and Rubin (1983) The central role of the propensity score in observational studies for causal effects, *Biometrika*, 70, 41-55。在纪念 *Biometrika* 第一百期的时候，这篇文章的引用数在该杂志排名第二；参看 Titterington (2013) *Biometrika* highlights from volume 28 onwards, *Biometrika*, 100, 17-73。截至写作本文的时候，Google Scholar 显示这篇文章已经被引用了 28392 次，已经超越了之前引用最高的文章 Liang and Zeger (1986) Longitudinal data analysis using generalized linear models, *Biometrika*, 73, 13–22（Google Scholar 显示引用了 18345 次）。这种改变，反映了近十年来，因果推断的研究在学术界的极端活跃性。另外，*Biometrika* 创刊于 1901 年，是最早的理论统计杂志之一。



费希尔否定吸烟导致肺癌

应用统计学家的兴趣，他们提出了很多推广的、更加精致的理论和方法，这些理论和方法被用在流行病学、经济学、政治科学等诸多学科的研究中。

虽然内曼的因果推断的文章为老一辈的统计学家所熟知，但是在很长一段时间它几乎销声匿迹了。它不仅仅不在观察性研究中被使用，也不在随机化实验中被使用。从上个世纪七十年代开始，鲁宾写了一系列文章告诉大家，潜在结果是思考统计因果推断的有力武器，但是他的文章起初并不被统计杂志所接受。多年以后，他这些在当时看来离经叛道的文章使他成为名副其实的统计学的拓荒者。

鲁宾还有很多其他关于因果推断的研究，这里就不再深入叙述；更多精彩的细节，可以在他的专著中找到[8]。为了引入下一部分的内容，我需要对鲁宾的工作进行恰当的批判。上面介绍的理论有两个致命的问题。第一个问题是，处理 Z_i 和结果 Y_i 之间的先后顺序是固定的，一前一后。但是，很多实际问题可能存在 Z_i 和 Y_i 同时产生，或者两者之间有动态关系的情况。鲁宾的这个简单模型，无法讨论这个问题。在计量经济学中，这被称为“联立方程模型”（simultaneous equation model）。第二个问题是，可忽略性假定的合理性如何判定？这个条件独立性不可能被观测数据验证，那么我们如何能相信由它导出的数学结果呢？费希尔曾经质疑吸烟导致肺癌的研究，他认为，可能存在一个基因，它既导致人更容易吸烟，也导致人更容易得肺癌，所以我们看到的吸烟和肺癌之前的相关性可能是虚假的因果作用。如果我们遗漏掉了关于这个基因的信息，那么鲁宾要求的可忽略性就不成立。

第一个问题不太容易有简单的解答。珀尔试图回答第二个问题。简言之，回答第二个问题，需要更多的关于数据生成机制的知识，而图模型是描述数据生成机制的一种有力工具。他提出了新的因果推断的范式，在某些条件下重新推导出了鲁宾的结果，并且得到了新的结果。

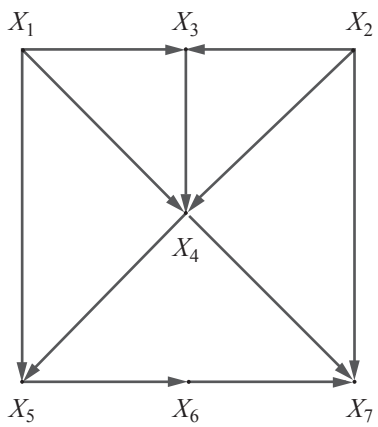
[8] 第一本是 Rubin (2006) *Matched Sampling for Causal Effects*。第二本是 Imbens and Rubin (2016) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*。两书均由剑桥大学出版社出版。

5 人工智能的“因果革命”：珀尔对图模型的因果解释

珀尔工作的雏形是图模型。直观上，这种模型用图来刻画条件分布，尤其是变量之间的条件独立性[9]。很多统计学家非常习惯用一个有向无环图 (directed acyclic graph；DAG) 来表示数据的生成机制。珀尔创造性地赋予了它因果关系的解释，并给了一系列运算法则。

为了描述珀尔的因果图理论，我们需要一些图的基本语言。一组随机变量 $X = (X_1, \ldots, X_p)$ 形成一个 DAG，每个节点对应着一个随机变量。我们用 pa_j 表示和节点 X_j 紧邻且处于箭头上游的变量集合（parent node），这个集合可能为空集。DAG 中变量的联合分布可以分解成

$$P(X) = \prod_{j=1}^{p} P(X_j \mid \mathrm{pa}_j).$$



一个 DAG 的例子

考虑上图中的 DAG。上面的联合分布的公式具体化成：

$$P(X) = P(X_1) \cdot P(X_2) \cdot P(X_3 \mid X_1, X_2) \cdot P(X_4 \mid X_1, X_2, X_3)$$
$$\cdot P(X_5 \mid X_1, X_4) \cdot P(X_6 \mid X_5) \cdot P(X_7 \mid X_2, X_4, X_6).$$

用上面的图，如何思考因果关系的问题呢？珀尔引入了 do 算子，表示干预某个随机变量到某个值，这类似我们在实验中控制某个变量。我先给一般的公式，再给具体的例子。一般地，

$$P(X_1 = x_1, \ldots, X_p = x_p \mid \mathrm{do}(X_j = x_j')) = \frac{P(X_1 = x_1, \ldots, X_p = x_p)}{P(X_j = x_j \mid \mathrm{pa}_j)} 1(x_j = x_j').$$

[9] 比如 A. P. Dempster 就用一个无向图来表示联合正态分布中的条件独立性：给定其他变量，如果两个变量条件独立，那么他们之间的边不存在。他的文章是：Dempster, A.P. (1972) Covariance selection. *Biometrics,* 157-175。

上面等式的左边定义的联合分布对应着一个新的 DAG：在原来的 DAG 上强制 X_j 取 x_j'，并且删除所有指向 X_j 的边（由于我们强制 X_j 取 x_j'，那么 pa_j 指向 X_j 的边不再起作用）。等式的右边展示了这个新 DAG 的联合分布和原始 DAG 联合分布的关系。从左边的联合分布，我们可以推出边缘分布，比如

$$P\,(X_7 = x_7 \mid \mathrm{do}(X_5 = 1)) \text{ 和 } P\,(X_7 = x_7 \mid \mathrm{do}(X_5 = 0)),$$

他们两者的差，度量了干预 X_5 在两个不同的值，X_7 分布的变化。我们可以用这两个边缘分布计算出对应的期望

$$E(X_7 \mid \mathrm{do}(X_5 = 1)) \text{ 和 } E(X_7 \mid \mathrm{do}(X_5 = 0)),$$

他们之间的差，就是 X_5 对 X_7 的平均因果作用。这就是在因果图下，用 do 算子定义的 X_5 对 X_7 的平均因果作用。一个至关重要的点是

$$P\,(X_7 = x_7 \mid \mathrm{do}(X_5 = 1)) \neq P\,(X_7 = x_7 \mid X_5 = 1),$$

即 do 算子和通常的条件概率在一般情况下是不同的。这也说明了，仅仅用传统概率论的语言，不足以定义因果作用。内曼和鲁宾用潜在结果，珀尔则用 do 算子。

来看一个具体的例子。从上面的 DAG 我们可以得到

$$P\,(X \mid \mathrm{do}(X_5 = 1)) = P\,(X_1) \cdot P\,(X_2) \cdot P\,(X_3 \mid X_1, X_2) \cdot P\,(X_4 \mid X_1, X_2, X_3)$$
$$\cdot P\,(X_6 \mid X_5) \cdot P\,(X_7 \mid X_2, X_4, X_6) \cdot 1(X_5 = 1).$$

从这个联合分布积分，我们可以得到边缘分布 $P\,(X_7 = x_7 \mid \mathrm{do}(X_5 = 1))$。类似可得 $P\,(X_7 = x_7 \mid \mathrm{do}(X_5 = 0))$。进一步可以计算 X_5 对 X_7 的平均因果作用。但是这个例子的趣味性还不够，因为上面的计算公式要求我们观测到所有变量的联合分布。

珀尔给出了一些更加有趣的结果：某些情况下，我们并不需要观测到所有的变量，也可以识别因果作用。下面用上面的 DAG 作为例子，解释他提出的“后门准则”（backdoor criterion）和“前门准则”（frontdoor criterion）。更一般的数学结果需要更多的术语和技术细节；感兴趣的读者可以参见珀尔的文章和专著[10]。

5.1 后门准则

根据珀尔的理论，要研究 X_5 对 X_7 的因果作用，我们无需观测所有的变量，仅仅观测 (X_4, X_5, X_7) 即可。直观上，X_4 阻断了从 X_5 到 X_7 的所有“后门路径”：

$$X_5 \leftarrow X_4 \rightarrow X_7,$$
$$X_5 \leftarrow X_4 \leftarrow X_2 \rightarrow X_7,$$
$$X_5 \leftarrow X_1 \leftarrow X_4 \rightarrow X_7,$$
$$X_5 \leftarrow X_1 \leftarrow X_3 \leftarrow X_4 \rightarrow X_7.$$

[10] 珀尔的开创性文章是：Pearl (1995) Causal diagrams for empirical research. *Biometrika*, 82, 669-688. 他的专著是：Pearl (2009) *Causality: Models, Reasoning and Inference*, 剑桥大学出版社。

那些指向 X_5 的、看似后门路径但是有“$\rightarrow \bullet \leftarrow$”这种结构的路径，并不算成真正的后门路径。珀尔证明，仅仅用 (X_4, X_5, X_7) 的联合分布，我们就可以表示

$$P(X_7 = x_7 \mid \mathrm{do}(X_5 = 1)) = \sum_{x_4} P(X_7 = x_7 \mid X_4 = x_4, X_5 = 1) P(X_4 = x_4),$$

类似有 $P(X_7 = x_7 \mid \mathrm{do}(X_5 = 0))$ 的公式，从而有如下的平均因果作用的公式：

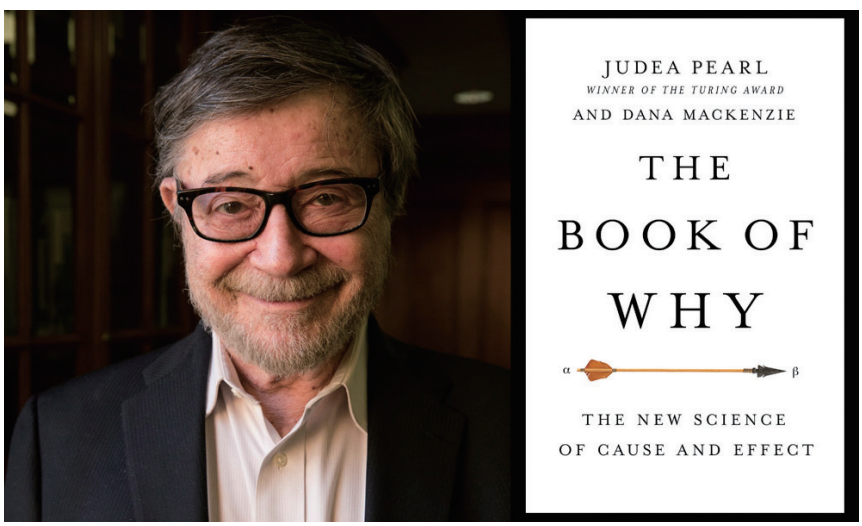
$$\begin{aligned}\tau &= E(X_7 \mid \mathrm{do}(X_5 = 1)) - E(X_7 \mid \mathrm{do}(X_5 = 0)) \\ &= \sum_{x_4} \{E(X_7 \mid X_4 = x_4, X_5 = 1) - E(X_7 \mid X_4 = x_4, X_5 = 0)\} P(X_4 = x_4).\end{aligned}$$

若将 X_4, X_5, X_7 换成 X, Z, Y，那么上面这个公式和在潜在结果下假定可忽略性推导出来的平均因果作用的公式一模一样。

鲁宾和珀尔的理论至此殊途同归。为了研究两个变量之前的因果关系，我们需要观测他们的“共同原因”（common cause），即，那些既影响原因又影响结果的变量。否则，鲁宾认为可忽略性不成立，而珀尔认为后门准则的条件不成立。

5.2 前门准则

珀尔的后门准则并没有给统计学家带来很大的惊喜，因为他给的公式在形式上并不是新的。但是，他的前门准则却让很多人吃惊。根据前门准则，我们仅仅需要观测 (X_5, X_6, X_7) 的联合分布，就可以识别 X_5 到 X_7 的因果作用。直观上，X_6 阻断了所有从 X_5 到 X_7 的“前门路径”；另外，X_5 到 X_6 没有后门路径，X_6 到 X_7 的后门路径都被 X_5 阻断了。在这些约束下，珀尔证明了下面的前门准则公式：



珀尔和他的畅销书《为什么》，图片来自：https://momentmag.com/author-interview-judea- pearl/

$$P(X_7 = x_7 \mid \mathrm{do}(X_5 = 1))$$
$$= \sum_{x_6} P(X_6 = x_6 \mid X_5 = 1) \sum_{x_5} P(X_7 = x_7 \mid X_5 = x_5, X_6 = x_6) P(X_5 = x_5).$$

这个公式乍看有些奇妙，甚至难以置信。或许下面的直观解释对理解这个公式何以成为可能有所帮助：

(a) X_5 到 X_6 的因果作用是可以识别的，因为他们之间没有后门路径；

(b) X_6 到 X_7 的因果作用是可以识别的，因为他们的后门路径被 X_5 阻断了；

(c) X_5 到 X_7 的因果作用仅仅通过 X_6 产生，因此，X_5 到 X_7 的因果作用可以理解成 X_5 到 X_6 的因果作用和 X_6 到 X_7 的因果作用的“乘积”。

珀尔在他 1995 年的 *Biometrika* 文章中给出了上面的和其他更一般的结果。他的文章引发了众多统计学家的讨论，当时大部分统计学家都保持怀疑甚至否定的态度，因为他的理论要求一个完全已知的图，这对大部分应用统计问题来说，是不切实际的。但是，珀尔的因果图，作为理论工具，对大家思考因果关系有很大的帮助。即使它不能直接用于数据分析，不少统计学家也认为他的理论有助于指导数据分析。珀尔由于这项工作于 2011 年获得了计算机科学的最高奖——图灵奖。

6 中国因果推断的研究

从古希腊开始，西方的哲学家似乎就钟情于因果关系的讨论。这种传统一直流传至今。爱因斯坦曾说，西方科学的发展以两个伟大的成就为基础：一是希腊哲学家发明的形式逻辑体系，二是通过系统的实验寻找因果关系。前者集中体现在欧几里得几何学中，后者肇始于文艺复兴时期，以伽利略为代表。

中国的文学作品，如屈原的《天问》和辛弃疾模仿而作的词《木兰花慢·可怜今夕月》，有一些对自然现象很感性的追问。佛教也有因果循环的理论。但是这些都没有和科学发生紧密联系。到了近代，中国学者受到了西方哲学的影响，也开始关注这个问题。比如，严复先生于 1902 年翻译了约翰·穆勒（John Stuart Mill）的名著《穆勒名学》[11]，其中卷下第五章是“论因果”、第七章

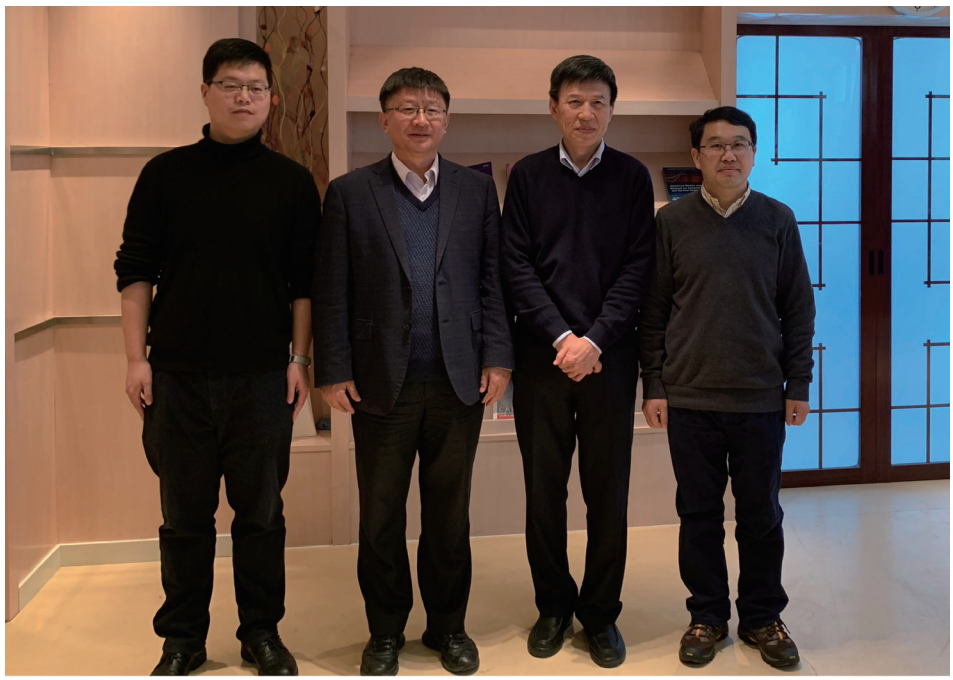
[11] 此书英文原名是 *A System of Logic*，直接翻译过来是《一个逻辑体系》，严复先生认为“逻辑学”就是中国的“名学”，这一学派兴起于先秦，代表人物有公孙龙等。这本书在英语世界产生过很深远的影响，其中五条“穆勒方法”总结了归纳推理中，获得因果知识的一些准则。严复先生是北京大学从“京师大学堂”更名后的第一任校长，也曾任复旦大学校长。

[12] 原书这章的题目是“On observation and experiment”。按照现在的习惯，“experiment”统一翻译成“实验”。前面用到的“临床试验”对应着“clinical trial”。“实验”和“试验”的意思似乎差别不大；中文英文皆如此。

屈原的《天问》反映了中国古人对自然和历史的好奇心（图片来网络）

是“论观察试验”[12]。又如，洪谦先生师承奥地利逻辑实证主义学派（logical positivism）的莫里兹 • 石里克（Moritz Schlick），于 1934 年在维也纳大学完成博士论文，题为“现代物理学中的因果律问题”。再如，金岳霖先生也对休谟和穆勒的哲学有独到的见解。到了现代，越来越多的中国哲学家也参与了有关因果关系的话题的讨论。

欧美的统计因果推断研究有很早的萌芽，比如内曼在 1923 年的论文，又如 Jerome Cornfield 等人于 1959 年关于吸烟和肺癌因果关系的研究，再如 William Cochran 对观察性研究的探索。但是，很多其他的统计学家则对因果推断充满了怀疑甚至敌意；仅有的这些早期研究也很零散、不成体系。鲁宾在 Cochran 的影响下，系统地研究因果推断，用数学的语言来描述一些应用统计学家已知的直觉和很多大家未知的奥妙。他在对因果推断充满敌意的氛围中，艰难地发表了一系列文章，坚持进行这方面的研究，培养了几代因果推断的学者。哈佛大学一直是因果推断研究的中心，这种状态持续到鲁宾退休、受聘到清华大学丘成桐数学中心。现在，美国各大统计系都有因果推断的研究者。在中国，北京大学数学科学学院的耿直教授，是国内统计因果推断研究的先驱，早在上世纪九十年代因果推断还是冷门话题的时候，就开始相关研究，坚持了三十多年，亲历了因果推断从冷门发展成热点的过程。在美国，鲁宾和珀尔学派相互批评对方的研究范式；但是在中国，耿直的研究整合了鲁宾和珀尔的研究范式，两者并行而不悖，在此基础上，产生了风格独特、思想深刻的研究成果。他曾应邀在国际工业与应用数学大会（International Congress on Industrial and Applied Mathematics，2011）作一小时大会报告。另外，耿直还培养了很多年



学术界的“四世同堂”：耿直（右二）、学生郭建华（左二，东北师范大学副校长），学生的学生朱文圣（右一，东北师范大学数学与统计学院副院长），学生的学生的学生王鹏飞（左一，东北财经大学讲师）

轻的、从事因果推断研究的学者，他们在国内外统计系担任教职，并且活跃于国内和北美的统计界，成为若干主流杂志非常重要的贡献者和这个领域的引领者。下面我简单评述一下耿直教授的一部分研究成果。

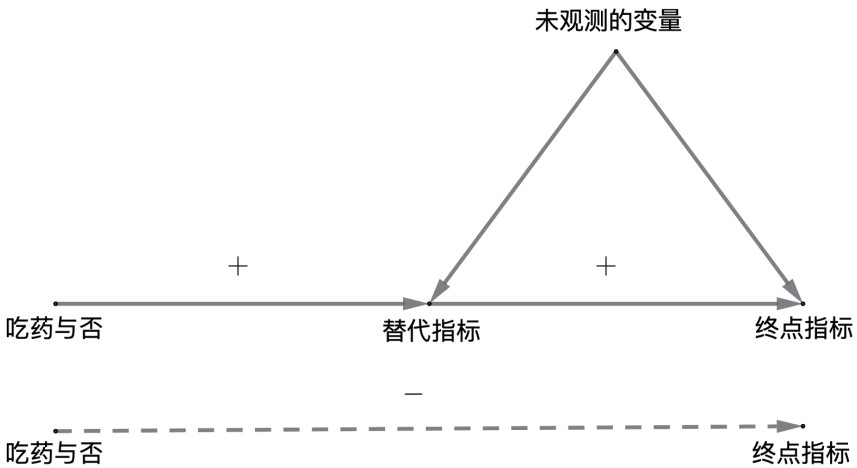
6.1 混杂因素

统计学里有个很有名的 Yule-Simpson 悖论：由于忽略某个变量，使得两个变量间的相关关系出现逆反现象。例如，某药对男性有效，对女性也有效，但是合并男和女后，发现该药对总体无效。这个悖论与前面休谟的质疑有些联系，即，从经验归纳不出因果关系。在这个悖论中被忽略的那个变量，被称为混杂因素（confounder）。它是因果推断的关键。前面鲁宾的可忽略性也被称为无混杂性，即排除了未观测的混杂因素，他的理论才成立。

因果推断需要关于混杂因素的假定，而判断某个变量是否是混杂因素，又需要关于因果关系的假定，这似乎有点循环论证。因此，确定什么是混杂因素是非常困难的。耿直探讨混杂因素的定义，提出了各种判断混杂因素的条件。其中一个结果是：如果不需要关于因果关系的假定，可以判断一个变量不是混杂因素，但不能确定一个变量是混杂因素。珀尔在《为什么》（*The Book of Why*）中写到，混杂因素问题的完整解决方案是因果革命的主要亮点之一。他声称利用因果图可以完美解决判断混杂因素的问题。但是，因果图常常是未知的，应该是因果推断的目标，而不是前提条件。耿直的研究，在一定程度上

弥补了珀尔研究的缺陷。这一系列文章发表在统计学顶级期刊 *Journal of the Royal Statistical Society, Series B* 上[13]。

6.2 替代指标悖论和准则、统计和因果关系的传递性



替代指标悖论的图模型。此图表示一个随机化实验中，“吃药与否”是随机化的，所以和“未观测的变量”都独立，但是这些“未观测的变量”可能同时影响“替代指标”和“终点指标”。即使“吃药与否”对“终点指标”没有直接的影响，替代指标悖论也会发生：“吃药与否”对“替代指标”有正作用，“替代指标”对“终点指标”有正作用，但是“吃药与否”对“终点指标”的作用却是负的。这个悖论类似于前面提到的 Yule-Simpson 悖论，它的关键是存在“未观测的变量”同时影响“替代指标”和“终点指标”。如果“吃药与否”对“终点指标”有直接的影响，那情况则更复杂，悖论更加不可以避免。注意，这个图和前面提到的“前门准则”有本质的不同。

在科学研究中，由于终点指标很难观测，所以常常选择替代指标。例如，在艾滋病的临床试验中，关心的终点指标是患者的生存寿命，但是需要等待很长时间才能被观测到，因此，有一些研究采用免疫力细胞 CD4 数目作为替代指标，药物能提高 CD4 数目就被认为是有效的。在深入研究了 Yule-Simpson 悖论的基础上，耿直教授发现了新的悖论，并称其为“替代指标悖论”：虽然新药对替代指标有正的因果作用，替代指标对终点指标也有正的因果作用，但是新药对终点指标可能有负的因果作用。

这项成果不仅有理论价值，而且对医学研究也有指导意义。有一本书《致命的药物》（*Deadly Medicine*）报告了一个真实的案例。医生的常识是，心

[13] Geng (1992) pp. 585-593; Geng and Asano (1993), pp. 741–747; Guo and Geng (1995), pp. 263-267; Geng, Guo and Fung (2002), pp. 3-15; Ma, Xie and Geng (2006), pp. 127-133。

律失常是猝死的危险因素，因此他们将纠正心律失常作为替代指标。一种新研制的药物能有效纠正心律失常，于是获得了美国食品药品监督管理局的批准。令人惊讶的是，该药物增加了数万人猝死，超过越南战争中美国士兵的死亡人数。这就是替代指标悖论的现实后果。几位杰出的统计学家，Ross Prentice, 唐纳德 • 鲁宾，Steffen Lauritzen[14]，分别都提出了关于替代指标的准则，不过他们的准则都无法避免替代指标悖论。耿直的文章，澄清了这些准则的缺陷，并且提出了新的准则，可以避免悖论出现。这一系列文章发表在统计学顶级期刊 *Journal of the Royal Statistical Society, Series B* 上[15]。Tyler VanderWeele 在他的综述文章中[16]，回顾并高度评价了耿直教授的这一系列工作。

耿直在这方面的精深研究，不仅在统计和医学上有意义，还对科学哲学有所增进。上面介绍的替代指标悖论，在数学上是不可思议的：如果 $Y = f(X)$, $Z = g(Y)$ 且 f, g 都是单调增函数，那么 Z 一定是关于 X 的单调增函数。在统计和因果推断中，由于随机性和隐变量的存在，这种传递性（transitivity）一般情况是不成立的。但是，科学研究和人类认知常常依赖这种传递性。它的理论根基是不完整的。耿直做出了奠基性的工作。著名数学家陶哲轩，也对类似的问题表现出了兴趣，他曾在博客中讨论“相关性何时可传递？”（*When is correlation transitive*?）[17]。他回顾了一些基本的不等式，有助于研究传递性。但是，这方面的数学结果还不算丰富。

6.3 因果图的结构探索

如上面所述，珀尔关于因果作用可识别性的理论依赖一个完整已知的图模型。一个更有挑战性的问题是：如何从数据中学习未知的图模型？耿直提出了分解和局部学习的方法，化繁为简，有针对性地构建图模型。在数据不能完全确定变量间因果图结构的情况下，他提出了一种实验设计的方法，干预最少的变量，将相关关系的图转变为因果关系的图。这对科学研究中的实验，有指导意义。这一系列文章发表在机器学习领域的顶级期刊 *Journal of Machine Learning Research* 上[18]。

[14] Prentice 曾获得年轻统计学家的最高奖 COPSS 奖章，终身成就奖“费希尔讲座”，他是美国医学院院士。鲁宾是因果推断的奠基人之一，曾获得终身成就奖“费希尔讲座”，美国科学院院士。Lauritzen 是英国皇家学会院士。

[15] Chen, Geng and Jia (2007), pp. 911-932; Ju and Geng (2010), pp. 129-142; Jiang, Ding and Geng (2016) pp. 829-848。

[16] 文章是 VanderWeele (2013) Surrogate measures and consistent surrogates. *Biometrics*, 69, 561-565。 VanderWeele 曾获 COPSS 奖章。

[17] https://terrytao.wordpress.com/2014/06/05/when-is-correlation-transitive/

[18] Xie and Geng (2008), pp. 459-483; Ma, Xie and Geng (2008), pp. 2847-2880; He and Geng (2008), pp. 2523-2547; Liu et al. (2020)。

7 统计因果推断的未来

虽然因果推断已经有了一些基础性工作，但是这些工作还不足以回应现实世界向我们发出的挑战。理论上，目前的研究范式还不能完美地应对复杂的实际工作需要。一些学者考虑了因果推断和微分方程的关系，但是这方面的研究还在草创阶段。不管是鲁宾还是珀尔的范式，对于有反馈的因果系统，都有致命的缺陷，这也是值得思考的问题。另外，现有的工作大多数都是在评估某个给定的原因对某个给定的结果的作用，而科学研究的本质是探索未知的原因。虽然因果图的结构学习对探索原因有帮助，但是这方面的理论还不够丰富。因果推断对整个思想界都有更深刻的意义，它是一种独特的思辨方式，很多层面上是传统的数学和概率论所不具备的。更广地来说，研究因果推断，对于丰富我们的精神世界，大有裨益。

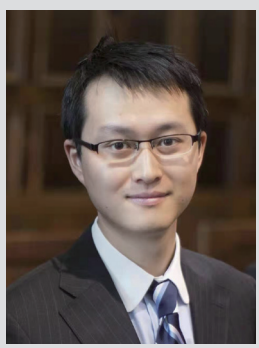
身处大数据时代，如何从海量数据中挖掘因果关系，也是一个非常有挑战性但是引人入胜的话题。由于研究深度学习（deep learning）而获得 2018 年图灵奖的计算机科学家约书亚 • 本希奥（Yoshua Bengio）最近转向因果推断的研究。他认为，机器学习和因果推断两种思想过去虽然独立发展，但是在未来会相互交织而产生新的成果 [19]。

从应用的角度，因果推断一直和很多学科发生深刻的联系。比如，经济学家深入研究的工具变量（instrumental variable），是探求因果关系的有力工具。又如，心理学家发明的因子分析（factor analysis），是研究隐变量的有力工具，这对研究不完全观测的图模型，大有帮助。我个人的研究，很大程度受到应用工作者的启发，他们研究的问题常常超越了现有的因果推断理论，成了新的理论研究的源头活水。

因果推断的研究，对规范我国药物批准和政策评估，也大有帮助。比如，前面提到的 Prentice 和鲁宾，都常常为美国食品药品监督管理局做咨询，解决他们在评估药效方面遇到的困难。我国的生物医药行业在未来有很大的腾飞空间，因果推断的学者们将发挥他们的巨大作用。再如，美国顶级高校的公共政策学院或者政府学院，都有研究因果推断的专家，他们研究公共政策对社会福利的影响，对于优化社会资源，起着重要作用。研究因果推断的学者，以后也应该走出象牙塔，承担社会责任。

[19] 本希奥的文章 Towards Causal Representation Learning 出现在 https://arxiv.org/abs/2102.11107。

致谢：郭建华（东北师范大学）、蒋智超（美国马萨诸塞大学）、苗旺（北京大学）、张俊妮（北京大学）、潘昆峰（中国人民大学）、黎波（清华大学）、刘中华（香港大学）、鞠念桥（美国哈佛大学）和宁少阳（美国威廉姆斯学院）给作者提出了宝贵的建议。美国密歇根大学生物统计系的宋学坤教授仔细阅读并修改了本文的初稿。



作者简介：

丁鹏，2004 年至 2011 年在北京大学数学科学学院获得本科和硕士学位，2015 年获哈佛大学统计学博士学位，2016 年起任教于加州大学伯克利分校统计系，2021 年晋升为副教授。其主要研究方向是因果推断。