第四篇

评价学方法

第十二章

评价学通用方法

本章首先提出了评价的建模问题，然后阐述评价学通用方法，包括基于现实世界系统的观察方法、基于自包含系统的实验方法和基于对象因果模型的建模方法。

12.1 评价的建模问题

评价学概念定义 14 (评价系统，评价模型，完美评价模型)**.** 用于评价的系统称为评价系统。现实世界系统是一种评价系统。

评价系统的简化表示称为评价模型 (*Evaluation Model*)。自包含系统、对象因果模型均是一种评价模型，前者是物理模型，后者是数字模型。

完美评价模型 (*perfect evaluation model*) 是指在最理想的情况下能够准确推断出评价对象 (EO) 的真实影响 e_t 的评价模型。不完美的评价模型，即使在最理想的情况下，也无法准确推断出评价对象 (EO) 的真实影响 e_t。

评价学概念定义 15 (评价结果差异，评价结果差异阈值)**.** 评价结果差异定义为真实影响 e_t 与推断影响 e_i 之间的差值。评价结果差异越小，评价结果准确性越高。

评价结果差异阈值 ϵ，即具体评价场景中可容忍的推断影响与真实影响之间的偏差上限。

本节基于我们之前的工作 [339]。我做了一些简化。

在不同场景中，提升评价有效程度与效率的关键在于建立一系列能确保主要特征可传递的评价模型。为了表达的清晰性，本节讨论的影响限于量的变化，暂时不考虑结构、机制或行为的变化、对象的消失或者新对象的出现。

在现实中，评价的关键是在控制评价结果差异，也就是在确保评价结果不超出评价结果差异阈值的前提下，如何保证评价的有效程度和效率。

在构建评价模型时，评价结果差异阈值可能对评价结果产生深远影响 。在某些情况下，尤其在安全关键型任务 (safety-critical tasks) 中，这种影响会引发严重关切，因为一旦失败，可能导致伤害、生命损失或重大环境破坏等严重后果。

因此，在充分理解利益相关方的评价需求后，可预先定义一个风险函数 $\gamma(\cdot)$，用于量化不同评估结果所对应的风险水平。

当评价对象引发的风险较高、评价结果可能带来重大影响时，必须设置较小的评价结果差异阈值。因为错误或者建模不合理的评价模型引发的后果（影响）会更加严重。

在建立评价模型时，我使用 $m(\cdot)$ 符号来表示这个建模过程。评价模型的准确性由 $m(\cdot)$ 决定。提高评价模型的准确性或者全面探索评价模型的空间确实会提高评价的有效程度。然而，这种方法也带来一个显著缺陷——其高昂的成本或开销。

基于上述设定，评价成本与评价准确性之间的权衡问题可被建模为一个优化问题：目标是在确保评估结果差异 $\mathrm{disc}(\mathrm{M_g}, \mathrm{M_p})$ 不超过预设评价结果差异阈值 ϵ 的前提下，最小化评价成本。

问题定义 5（评价建模的问题定义）. 评价结果差异函数 $\mathrm{disc}(\cdot)$ 用于衡量在特定评价模型 $\mathrm{M_g}$ 和完美评价模型 M_p 下评价结果之间的偏差，定义如下：

$$\begin{cases} \text{discrepancy threshold } \epsilon = |\mathrm{e_t} - \mathrm{e_i}|, \\ \text{discrepancy threshold } \epsilon = \gamma^{-1}(\cdot), \\ \mathrm{M_p} = \mathrm{M_p}(\mathrm{k_1}, \cdots, \mathrm{k_n}), \\ \mathrm{disc}(\mathrm{M_g}, \mathrm{M_p}) = \mathrm{disc}\Big(\rho\big(\mathrm{m}(\mathrm{M_p})\big), \rho(\mathrm{M_p})\Big), \\ \mathrm{accuracy}(\mathrm{M_g}) = \mathrm{accuracy}\big(\mathrm{m}(\mathrm{M_p})\big), \\ \mathrm{cost} = \psi(\|\mathrm{M_g}\|). \end{cases} \tag{12.1}$$

评价建模的优化问题可以具体表示如下：

$$\arg\min \mathrm{cost}(\mathrm{M_g}) \quad \textit{subject to} \quad \mathrm{disc}(\mathrm{M_g}, \mathrm{M_p}) < \epsilon, \tag{12.2}$$

or

$$\arg\max \mathrm{accuracy}(\mathrm{M_g}) \quad \textit{subject to} \quad \mathrm{cost}(\mathrm{M_g}) < \varepsilon. \tag{12.3}$$

在公式 12.1中，$\rho(\cdot)$ 是一个测量函数 。此外，我们将评价成本的定义简化为常数 ψ 与 M_g 所需空间容量的乘积。显式地引入成本因素使我们能够将评价过程中相关的资源限制和实际约束纳入考量。

另一个相关问题是：如何确保评价的可溯源性？ 这要求将评价结果差异归因于评价条件（ECs）之间的差异，从而建立清晰、透明的溯源能力。

从概念上讲，可溯源性要求建立一个量化映射关系，用以描述测量函数 $\rho(\cdot)$ 的输入差异与其在评价结果差异之间的关系，而这一过程正是由我们前述数学模型所刻画的。我们的前期工作 [339] 表明，这一概念与函数梯度的数学表达高度契合：函数梯度给出了每个输入变量变化时，输出随之变化的速率。

在评价背景下，评价结果的梯度可以表示如下，它可以是一个张量或者向量：

$$\nabla\rho(m(M_p)) = \nabla\rho\Big(m\big(M_p(k_1, \cdots, k_n)\big)\Big) = \Big(\frac{\partial\rho}{\partial m}\frac{\partial m}{\partial M_p}\frac{\partial M_p}{\partial k_1}, \cdots, \frac{\partial\rho}{\partial m}\frac{\partial m}{\partial M_p}\frac{\partial M_p}{\partial k_n}\Big). \tag{12.4}$$

“∇” 是梯度算子，在此表示一个标量函数对其各个变量的偏导数组成的向量。“∂” 是偏导数算子，表示在保持其他变量不变的情况下，函数对其中一个变量的变化率。

在评价中，各类评价条件（EC）组件的 解析数学表达式（闭式数学表达式）并不总是可获得的。尽管如此，我们仍可借鉴数值方法中的梯度求解思路：通过对不同输入变量对应的 EC 施加微小扰动，并观察评价结果差异的变化，从而近似估算出梯度。

12.2 评价学通用方法

在本节中，我将介绍基于评价学（Evaluatology）通用方法。在本书的这个版本里，我暂时不介绍逆评价和基于评价学的设计，这部分内容将在后续版本或者其他专著里介绍。

在评价学中，观察、实验与建模是三种基本的手段。

观察或者实验是基本的求索方法，它能够从不同的角度揭示对象之间的相互影响。观察和实验存在三点差异：实验相对观察提供了更多的求索条件；实验存在更高的限制条件，能改变求索对象或者求索条件；实验成本高于观察成本。

建模是一种心智方法，它可以基于观察或者实验，或者其他心智方法，例如猜想和推理。

现实世界系统、自包含系统和对象因果模型分别是对应观察、实验和建模的评价系统或者模型。

在评价学通用方法里，现实世界系统是主体正在观察的系统，用以揭示对象的影响或者事件的原因。

自包含系统（SCS）是精巧设计的实验系统。自包含系统是对对象之间相互影响的物理建模，能排除不相关对象的影响，尽量减少相互作用的对象数目。设计精巧的自包含系统对于理解对象之间的相互影响至关重要。历史上，卡文迪许的万有引力实验、拉瓦锡的燃烧实验和孟德尔的豌豆实验都设计了精巧的自包含系统。自包含系统与蛮力法完全不同，会极大地降低实验成本，能在实验过程中迭代地检验因果的假设。

对象因果模型（OCM）是基于影响机制建立的对象之间的因果模型。它能够解释被动对象的任何变化。建立基于对象之间相互影响的因果模型，这依赖于专业知识。完全忽略专业知识，建立所谓的因果模型，通常具有误导性。在 [224] 中，吸烟和致癌之间因果的推断，朱迪亚 · 珀尔在和专业医生有关癌症发病的机理的讨论中，就暴露了缺少专业领域知识建立因果模型的弱点。

正如求索铁林寨模型所示，现实世界系统、自包含系统和对象因果模型的关系不是一个完全替代的关系，相反，它们是平行或者互补的关系。可以直接基于现实世界系统建立对象因果模型，或者在自包含系统基础上建立对象因果模型。如果无法建立数学模型，也可以建立物理模型，也就是自包含系统。

根据第 六章，检验（Testing）是确认模型有效性的最终方法。在评价学里，无论是物理模型还是数学模型，都需要验证。这一步的关键是确定检验基准（Test Oracle）。

评价学通用方法遵循综合和迭代的过程，通常起始于观察或实验，也可以借助于猜测或者推理建立模型，通过测试对模型进行检验。在这个循环基础上不断进行修改和迭代。

12.3 可复用的基本评价流程

在详细阐述基于三个核心概念的不同评价方法之前，本节先给出不同方法可以复用的四个评价流程：包括定义评价对象、影响对象、混杂对象以及推断和检验 EO 的影响，如图 12.1 所示。可复用的基本评价流程包括四个步骤，每个步骤都需要检验。本章主要贡献者有高婉铃、王晨曦和王磊。

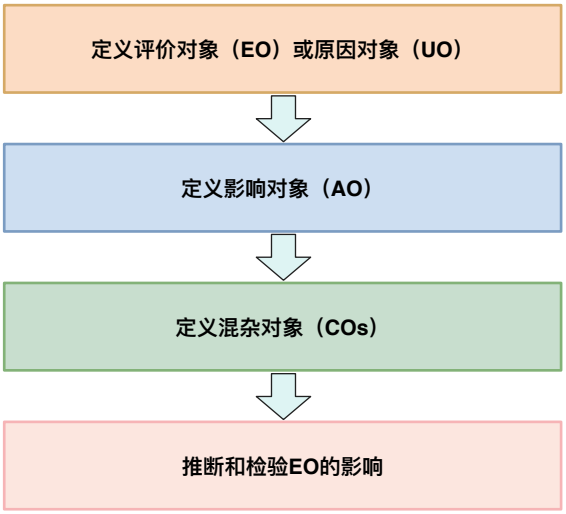


图 12.1: 评价学通用方法四个基本流程，基于观察、实验或者建模的评价方法均可以复用。

12.3.1　定义评价对象

首要且最根本的步骤，是对 EO 进行清晰、严格的刻画与定义。若缺乏精确的 EO 定义，则不同 EO 之间，乃至”同一”对象不同实例之间的比较，均会产生歧义或者无效的结果。

例如，两个对象均被称为“处理器”，但若一个是 CPU，另一个是 GPU，则二者的结构、机制、属性与行为差异巨大，直接比较会产生误导或者无意义的结论。即便限定在功能较窄的 CPU 范围内，EO 也不是铁板一块 (monolithic)：同样的处理器可以在不同配置下工作，例如是否启用加速，或者使用不同的热或功耗管理设置。若不明确指出 EO 的配置是如何规定的，评价结果可能会大相径庭，且难以预测。

采用第 八 章定义的符号，明确定义的 EO 配置空间的一个元素可以定义为：$o_i = \{o_{i1}, o_{i2}, ..., o_{ix}, ...\}$。这里，$i$ 表示 EO 的特定实例（如处理器型号 X），而 x 枚举了可能的配置（例如，“默认模式”、“启用涡轮加速”、“低功耗模式”等）等。每个元素 o_{ix} 对应 i_{th} 组件的特定配置。选择待评价的配置不仅是一个技术选择，而且是一个科学选择，因为它决定了评价结果的可解释性和可比较性。

从实际的角度来看，配置空间 O_i 往往极大，不可能，也没有必要对所有配置进行穷尽性评价。因此，评价人员通常通过选择代表性配置 o_{ij} 来减少这个状态空间，例如在广泛的 EC 集合下，表现最优或性能稳定的配置，或者体现最常见的真实用户场景的配置被选择。例如，在 CPU 评价中，评价者可能将处理器固定为默认设置，或者选择在大多数标准工作负载上产生最佳性能的配置。

然而，这种策略存在一个关键的注意事项： 没有任何一种配置能保证在所有评价条件 (ECs) 下都表现最优 。在计算密集型工作负载中表现出色的 CPU 设置可能在能耗受限的场景中表现不佳。因此，虽然采用代表性配置可以大大降低评价成本和复杂性，但也存在过度简化的风险。

在“完备性（评价所有可能的配置）–效率（简化配置空间）”这一矛盾之间取得平衡，

构成了 EO 定义的核心工作。如果目标是广泛的科学表征，则可能需要包括更多的配置。如果目标是实际的部署指导，那么少量精心选择的配置可能就足够了。

综上，定义 EO 并非一个简单的前期步骤，而是一个决定整个评价有效性的基础性过程。它确保了评价结论能准确归因于 EO 本身，而不是其配置中隐藏的或不受控制的变化。

12.3.2　定义影响对象

在明确定义 EO 后，第二步是定义 AO——在它上面 EO 影响能被直接测量或测试（见第 8.1 节定义）。AO 是*接受影响的系统*。AO 的定义非常关键，它决定了哪些量能被直接测量或者测试，这是建模或者推理的基础。

AO 是一个由多个组件构成的系统：

$$a_j = \{c_{j1}, c_{j2}, ..., c_{jk}\}, \tag{12.5}$$

其中 c_{jk} 表示第 j_{th} 个*不可或缺组件*的第 k_{th} 种配置。不可或缺组件是指它是一个最小功能集合的组成部分，这个最小功能集合整体上使得 EO 的影响可以直接被测量或者测试。AO 必须具有“足够完备性”以使得 EO 的影响可以直接被测量或者测试，又“不过度冗余”，从而降低评价系统或者评价模型的复杂性。

在一类评价问题中，EO 无法独立运行，必须嵌入到一个为其提供必要运行基础设施的更大系统之中。例如，当 EO 是 CPU 处理器时，它无法脱离其他组件单独进行评价，因为执行任何有意义的负载都依赖于内存、存储以及输入/输出子系统。在这种情况下，AO 即为一个能够支持计算的最小运行环境，包含所有不可或缺的组件。EO 是属于 AO 的一部分。

在另一类评价问题中，AO 并不包含 EO 本身。在许多领域，尤其是工程领域之外，AO 可能作为一个独立系统存在，用以响应 EO 的干预或作用。例如，在药理学评价中，EO 是药物化合物，而 AO 是人体这一完整的生物系统，EO 的生化影响正是在其中得以体现。

因此，根据具体领域与因果结构的不同，EO 与 AO 之间的关系可以概括为两类：

- *包含型关系*：AO 将 EO 作为其不可或缺的内部组件之一，EO 作为子系统运行于一个更大的可测量的整体之中（例如计算机系统中的 CPU，或车辆中的发动机）；
- *外部关系*：AO 独立于 EO 存在，接受 EO 的影响，例如人体接受药物。

无论属于哪一种关系类型，定义 AO 都必须在两项基本准则（criteria）之间取得平衡：*保证独立运行的完备性与避免引入额外的复杂性*。范围过窄的 AO 将无法保证独立运行的完备性，导致影响无法直接被测量或者测试。而范围过宽的 AO 则可能引入额外的负载性。

定义一个合适的 AO 在实践中往往涉及反复的抽象与简化，这一过程高度依赖专业经验。评价人员通常先识别出 AO 正常运作所需的所有组件，再系统性地剔除非必要元素，直至定义一个在完备性和简单性之间平衡的 AO。

12.3.3 定义混杂对象

第三个基础步骤是定义混杂对象（COs）。形式上，混杂对象集合可表示为：

$$COs = \{co_1, co_2, \cdots, co_i, \cdots, co_l\}, \tag{12.6}$$

其中，co_i 表示一个可能影响 AO 测量或测试结果的对象。在评价系统中，EO 与 COs 通常共同作用于 AO，产生可测量或可测试的总体影响。需要强调的是，只有那些能够直接或者间接影响 AO 的对象，才构成评价意义上的混杂对象。

从影响源的角度可以区分不同的混杂对象。一种混杂对象不通过 EO 直接影响或者间接影响 AO。另外一种混杂对象通过 EO 间接影响 AO。

以 CPU 评价为例，若 EO 是待评价处理器，AO 是能够运行工作负载并产生性能、功耗或能效结果的计算机系统，则 COs 至少可能包括测试负载、数据集、不同设置的线程系统和进程系统。上述对象会显著影响 AO 上观察到的结果。例如，同一 CPU 在不同测试负载、不同线程数或不同数据集规模下可能呈现完全不同的性能表现；若这些 COs 未被明确规定，评价结果将缺少内部有效性。

12.3.4 推断和检验 EO 的影响

测量或测试总体影响

明确定义 EO、AO 与 COs 后，下一步是获取 EO 与 COs 共同作用于 AO 所产生的总体影响。

评价学的实证阶段正是通过这一过程体现：在 AO 上直接测量或者测试 EO 与 COs 共同在 AO 上产生的总体影响。

总体影响并非仅由评价对象（EO）产生的真实影响，而是 EO 与 COs 共同在 AO 上产生的影响。在实际情况中，这是可以直接获得的观察数据。

获得 EO 的推断影响

在获得总体影响后，下一步，也是最关键的一步是*推断 EO 对 AO 的影响*，即所谓的*推断影响*。需要注意的是，推断影响与真实影响并不等同：真实影响是推断影响所逼近的目标。

这一阶段可以借助第 七章阐述的推理学或者第 九章阐述的实验学方法。

对推断影响进行检验

在获得 EO 的推断影响之后，评价学的最后一步是对该推断影响进行验证。这可以采用第 六章阐述的测试学方法。

12.4 基于现实世界系统的评价方法

基于现实世界系统的评价方法是一类以观察为主要手段的评价方法。在这类方法中，评价者通常不能操作 EO、AO 或 COs，也不能任意构造评价条件，而是在现实世界中

已经存在的系统运行过程中，观察 EO 与 AO 之间的影响关系，并通过检验过的模型对 COs 的影响进行调整。

由于评价发生在真实环境中，评价结果能够反映对象在实际部署条件下的影响。然而，其主要挑战在于 COs 往往复杂且无法控制。因此，基于现实世界系统的评价方法并不依赖物理隔离，而是依赖于对评价条件的系统记录、对混杂对象的识别以及对观察结果的因果调整。

形式上，现实世界评价系统可表示为：

$$RWS = \{EO, AO, COs\}, \tag{12.7}$$

其中，RWS 表示现实世界系统。在现实世界中，EO 与 COs 共同作用于 AO，产生可观察结果：

$$Y = \rho(AO \mid EO, COs). \tag{12.8}$$

评价的目标是从观察到的整体结果 Y 中，推断 EO 对 AO 的影响。

基于现实世界系统的评价通常包括以下步骤。首先，评价者需要定义 EO、AO 与 COs，明确哪些对象是评价关注的原因对象，哪些对象承载评价结果，哪些对象可能对结果产生混杂作用？其次，需要在现实系统中记录尽可能完整的评价条件，包括对象配置、运行环境、输入数据、负载状态以及其他可能影响结果的上下文因素。再次，评价者需要借助虚拟化实验方法（见第 9.4节）对 COs 的影响进行调整，从而估计 EO 的推断影响。

在潜在结果框架（见十七章）下，若 T 表示不同的 EO 干预，$Y(1)$ 与 $Y(0)$ 分别表示 AO 在不同 EO 干预下的潜在结果 [253]，则基于观察的评价学方法的关键问题是如何在不能保证真随机化情况下逼近 [133, 241]：

$$\mathbb{E}[Y(1) - Y(0)]. \tag{12.9}$$

由于现实系统中 T 往往并非随机分配，COs 可能同时影响 T 与 Y，因此需要在可观察混杂对象 $X \subseteq COs$ 条件下进行调整 [133, 241]：

$$\mathbb{E}[Y(t)] = \sum_{x} \mathbb{E}[Y \mid T = t, X = x] P(X = x). \tag{12.10}$$

这一过程本质上是通过协变量调整，在无混杂成立的条件下获得 EO 对 AO 的影响。

在结构因果模型 [221, 224] 中（见第 十九章），当满足一定条件下，评价者可以进一步将 EO、AO 与 COs 的关系表示为 DAG 结构，并通过后门准则、前门准则或 do 演算等方法识别 EO 的影响 [221, 224]。例如，当 COs，EO, AO 可以简化为变量时，当存在一组可观察 COs 阻断 EO 与 AO 之间的非因果路径 [224] 时，可以通过控制该组 COs 来估计 EO 对 AO 的影响。这类方法的优势在于能够显式表达变量之间的影响路径，但其有效性依赖于因果结构假设的合理性和 COs 观察记录的充分性。

请注意，在第 21.8节，我将讨论潜在结果理论和结构因果模型在特定条件下可以成为评价学通用方法的简化、特例或者补充方法。

基于现实世界系统的评价方法适用于难以进行物理实验但能够收集大规模真实运行数据的场景。其优势是贴近真实应用，能够反映复杂环境中的对象影响；其局限性是评价结果依赖于对 COs 的识别和控制。若关键 COs 未被观察或因果结构假设不成立，则推断影响可能偏离 EO 的真实影响。

12.5 基于自包含系统的评价方法

基于自包含系统（SCS）的评价方法是基于实验的评价方法。与现实世界系统不同，自包含评价系统并不试图完整保留真实世界的全部复杂性，而是通过精心设计一个可运行、可隔离、可重复的实验系统，使 EO 对 AO 的影响能够在尽可能清晰的边界内被实验和推断。

SCS 作为一个自洽且可隔离的环境，使得 EO 的因果影响能够被有意义地观察、测量、测试或分析。“自洽”意味着 SCS 纳入了必要的 COs；“可隔离”则意味着能够将 SCS 中包含的对象与其他无关对象区分开来。

SCS 可被形式化表示为：

$$SCS = \{EO, AO, COs\}, \tag{12.11}$$

(1) SCS 的结构性作用

SCS 定义了评价范围的边界。在实际中，EO 与 COs 会同时作用于 AO，所测量或测试得到的结果即为*总体影响*——EO 与 CO 共同作用于 AO 的结果。因此，评价的核心目标并非仅仅测量或测试这一总体影响，而是要在其中*分离并量化 EO 的真实影响*。只有将 EO 的真实影响与 COs 的影响区分开来，才能确保观察到的差异真实反映了 EO 带来的影响。

(2) 组成与相互作用 SCS 描述了三个组件之间的结构性与因果关系：

从 EO 指向 AO 的箭头表示评价关注的核心因果路径，即我们试图揭示的直接影响；而 COs 对 AO 的影响则代表外部干扰的作用。在许多情况下，EO 可能通过 COs 传递间接影响到 AO。因此，SCS 的设计必须对 COs 进行审慎选择与控制，以确保 EO 的影响可识别。

在生物医学评价中，若 EO 为药物化合物，AO 为人体，则 SCS 还应包含诸如环境条件、生活方式或心理压力等 COs。这些 COs 同样会影响测量或者测试结果（如血药浓度、生理反应），必须将药物的真实影响与外部干扰区分开来。

(3) 最小完备性

SCS 必须包含必要的对象，使得 EO 的影响可以在 AO 上直接测量或者测试——不多也不少；

(4) 物理实验方法. 基于 SCS 的评价方法核心在于通过物理实验方法来获得 EO 的真实影响。根据处理机制的不同，可以概括为三类基本方法：随机化、隔离和存在算子。

- 随机化方法 [295, 224, 133]：通过物理的随机分配机制处理 COs 的混杂。当评价者能够将 AO 实例随机分配到不同 EO 干预时，随机化可以使干预组与对照组在统计意义上具有等价性。在样本规模足够大的条件下，即使某些 COs 未被直接观测，它们在不同组中的分布也可以趋于平衡。

  形式上，若 T 表示 EO 的干预状态，$Y(1)$ 与 $Y(0)$ 表示潜在结果，随机化要求：

  $$(Y(1), Y(0)) \perp\!\!\!\perp T. \tag{12.12}$$

  在该条件下，观察到的组间差异可以用于估计 EO 的平均影响：

  $$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0]. \tag{12.13}$$

需要指出的是，随机化获得的是 EO 对不同 AO 实例的平均影响，但当 AO 只能是单个对象实例，或实验成本极高时，随机化并不总是可行。

- *隔离方法*：通过屏蔽或固定 COs 来处理混杂的实验方法。与随机化不同，隔离不要求形成大规模对照组，也不依赖 AO 实例的随机分配。它的基本思想是：通过物理、工程或制度化手段，使某些 COs 的影响不进入评价系统，或对 COs 进行控制，从而提高 EO 影响的可识别性。

  假设评价系统中存在 EO、AO 与 COs。隔离方法试图构造一个受控的 SCS，使得：

  $$Y = \rho(AO \mid EO, COs = c_0), \tag{12.14}$$

  其中 c_0 表示被控制或者屏蔽的 COs 配置。在这种情况下，AO 中观察到的结果差异更容易被解释为 EO 的影响。隔离方法在物理、化学和工程评价中被广泛使用。其优势在于可以在较小规模甚至单个 AO 实例上获得因果影响；其局限在于过度隔离可能降低评价结果的外部有效性。因此，隔离并不是简单地降低系统的复杂性，而是在内部有效性与外部有效性之间取得平衡。

- *存在算子方法*：从 EO 本身出发识别其影响的实验机制。它通过模拟实现 EO 的存在与否来观察 AO 结果随之发生的差异，从而推断 EO 的影响。

  与隔离方法不同，存在算子并不是从 COs 出发屏蔽混杂对象，而是从原因对象 EO 出发，直接操作 EO 的存在状态。

  需要强调的是，通过存在算子获得的影响并不必然等同于隔离 COs 后获得的 EO 影响。假设系统中只有原因对象 A、影响对象 B 与混杂对象 C。若 A 不仅直接影响 B，还通过影响 C 间接影响 B，则对 C 进行隔离后获得的 A 对 B 的影响，并不等同于通过存在算子演算 [1]获得的 A 对 B 的影响。前者更接近受控条件下的直接影响，后者则可能包含由 A 经由 C 传递的间接影响。

(5) 实践层面的意义。合理定义 SCS 并非纯粹的形式化步骤，而是对实验设计、结果可复现性与有效性具有实际意义。定义不当的 SCS 可能导致结果被混杂，使得在 AO 上直接测量或测试到的差异并非源于 EO，而是来自未受控的 COs 变化。相反，过度简化的 SCS 则可能消除具有现实意义的混杂对象，从而削弱结果的外部有效性。评价的艺术在于，找到系统控制（内部有效性）与现实代表性（外部有效性）之间的恰当平衡。

12.6　基于对象因果模型的评价方法

基于对象因果模型的评价方法是基于建模的评价方法。与基于现实世界系统的观察方法不同，它并不只是对通过观察获得的评价结果进行调整；与基于自包含评价系统的实验方法不同，它也不一定要求评价者能够物理操作 EO 或 COs。其核心思想是：基于领域知识、机制理解和已有证据，显式建立 EO、AO 与 COs 之间的对象影响结构，并在该结构上推断 EO 对 AO 的影响。

[1]本书的当前版本还没有阐述存在算子演算。

对象因果模型（Object Causal Model, OCM）的目标是通过数学方程的方式刻画评价系统中对象之间的影响路径。一个通过检验的 OCM 能够揭示对象及其 COs 如何影响 AO。

(1) 建模的角色：结构先于观察。 OCM 的建立基于：

- 对 SCS 中对象组成的理解；
- 对领域机制（mechanism）的认知；
- 对可能影响路径的结构性假设。

因此，OCM 并不是对观察数据的简单拟合，而是对“什么会影响什么，以及如何影响”的形式化表达。观察数据仅在模型建立之后，作为验证或求解该模型的输入。

(2) 对象与影响关系的形式化 。在 OCM 中，SCS 被表示为对象及其影响关系的集合：

$$OCM = (\{A_i\}, \{A_i \Rightarrow A_j\}), \tag{12.15}$$

其中 $A_i, A_j \in EO, AO, COs$。

这里的核心不是变量之间的相关性，而是对象之间的影响机制：

- $A \Rightarrow B$：表示 A 对 B 存在直接影响；
- $A \not\Rightarrow B$：表示无直接影响；
- $A \Rightarrow B \rightsquigarrow C$：表示通过中介对象产生的间接影响。

这些关系在建模阶段被显式定义。

*(3) 直接与间接影响的结构化描述。*OCM 的核心内容在于区分不同类型的影响路径：

- 直接影响（*direct effect*）：$E_{A \Rightarrow B}$；
- 间接影响（*indirect effect*）：$E_{A \Rightarrow B \rightsquigarrow C}$；
- 复合路径：多级传递 $A \Rightarrow B \rightsquigarrow C \rightsquigarrow D$。

在这一阶段，这些影响尚未被量化，而仅作为结构组成的对象存在。在 OCM 模型中，直接影响或者间接影响本身可以被视为衍生对象，从而参与更高层次的影响传递。

(4) 对象因果方程（Object causal equations）的建立 。在明确影响结构之后，需要为关键对象建立影响方程，用以刻画其状态如何由其他对象决定：

$$B = f_B(\text{Parents}(B)), \tag{12.16}$$

其中 Parents(B) 表示所有对 B 有影响的对象集合。

需要强调的是：

- 方程形式来源于领域知识，而非数据拟合；
- 方程是确定的，而不一定是概率形式；

- 对象之间的影响机制是多样化的。

这一组方程构成了一个可推测或者可预测的系统，使得在给定输入条件下，可以推导 AO 的状态变化。

(5) 与 SCS 的关系。

SCS 与 OCM 分别从两个互补的层面刻画评价系统：SCS 提供一个可运行、可实验的物理系统，而 OCM 则提供该系统的数学表示，用于刻画对象之间的影响机制及其在更广泛配置空间中的行为。

- *SCS 是物理实验系统，它包含 EO、AO 与 COs，可以精准地控制实验，从而产生实际的观察数据。然而，受限于实验条件与资源，SCS 通常只能覆盖有限的配置空间；*
- *OCM 是数学模型（abstract model）：它以对象及其影响关系为基本单元，刻画系统的作用机制，并在形式上允许对更广泛（甚至潜在穷尽）的配置空间进行推测、预测或者分析。*

12.7 本章小结

本章首先给出了建立评价模型的问题定义，然后介绍了评价学的通用方法，包括基于现实世界系统的观察方法、基于自包含系统的实验方法以及基于对象因果模型的建模方法。

第十三章

评价学通用方法在不同学科应用的讨论

本章中，我将初步讨论评价学通用方法在不同领域的应用。为方便起见，本章后续将为每个案例研究分配一个唯一的案例编号，以示区分。

我和其他贡献者区分了两类情况：一种是回顾性案例，一种是前瞻性案例。本章首先介绍回顾性案例，然后介绍前瞻性案例。

13.1 回顾性案例

这一节，我将介绍物理、化学等不同学科的回顾性案例。

13.1.1 一个物理学例子：牛顿的苹果树

问题背景

有一个广为人知的传说：一棵位于剑桥大学的苹果树，后来以“牛顿树”名垂千古，启发了千年一遇的天才艾萨克·牛顿去思考为什么苹果会从树上掉落（案例一）[301]。案例一体现了评价的逆问题（de-evaluation），我在第 8.4.3 节中将其正式称为“逆评价”，即追溯现象的原因。

牛顿观察到了苹果从树上掉落这一现象，即苹果这个对象发生了变化，从树上掉落，并开始思考其背后的原因。在案例一中，直接 AO 明确无误就是苹果本身。牛顿可以测量这个直接 AO 的多个物理量，这些量综合反映了原因对象及其大量混杂对象的总体影响。牛顿在解释这一现象时面临了三大严峻挑战。

- 哪些主要的原因对象导致了苹果从树上掉落?
- 原因对象通过什么样的影响机制影响直接 AO?
- 如何推断原因对象产生的影响?

图 13.1: 要找出导致苹果从树上掉落的原因，很难将一个自包含系统（SCS）完全隔离出来。

自包含系统

案例一具有三个独特特征。首先，很难将一个自包含系统（SCS）完全隔离出来。即便牛顿胆识过人，他仍不得不在系统中纳入地球、太阳、月亮、风和空气等对象。此外，可能还存在其他未知对象，如图 13.1所示。

第二，在这个自包含系统中几乎无法人为改变不同条件 。然而，牛顿是一位天才，他能够设计出多种极端的自包含系统 (Extreme SCS)，以隔离其他混杂对象的影响——这一思路后来由亨利 · 卡文迪许出色地实现。

比如，可以用橘子、梨子或桃子来替代苹果，这表明影响机制本身影响对象 （AO）无关。一个充满真空的 评价条件（EC）可以消除风和空气的影响，而在地球不同地理位置设置真空实验环境，则能揭示该影响机制随地理位置变化的规律。

亨利 · 卡文迪许在 1797 年测量引力常数（G）的实验 [47, 167]，通过一系列精密的实验技巧提供了非常简单的自包含系统，如图 13.2所示。

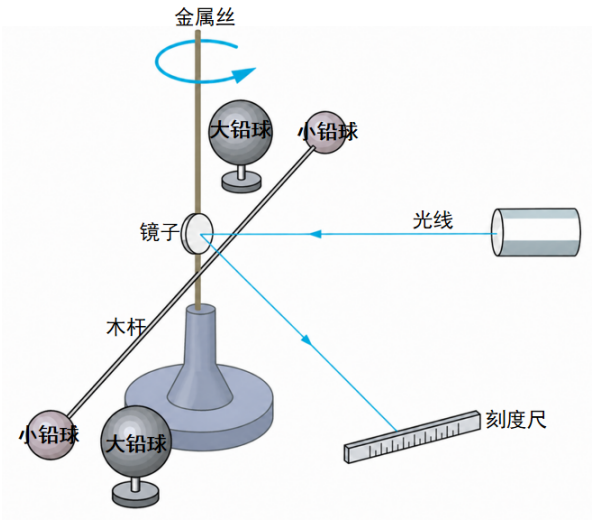


图 13.2: 卡文迪许实验图 [47]。何谦重新绘制。

亨利 · 卡文迪许如何设计自包含系统来测量万有引力常数 (G)

实验装置

- 一个扭秤，由一根六英尺长的木杆组成，木杆通过一根金属丝悬挂，两端各固定一个较小的铅球（每个重 1.6 磅）。
- 两个较大的铅球（每个重 350 磅）被放置在较小铅球的附近，两者相距约九英寸。
- 金属丝上附有一面镜子，可将一束光线反射到刻度尺上，从而放大并精确测量因引力作用引起的微小扭转运动。

方法

- 大铅球与小铅球之间的引力导致木杆发生轻微扭转，从而使光束发生偏转。
- 通过测量这一偏转量，卡文迪许计算出两个质量之间的引力大小，并结合牛顿的平方反比定律推导出了引力常数 G。

消除其他混杂对象的影响以隔离自包含系统

屏蔽地球引力的影响

- 对称布置 ： 扭秤两侧受到地球引力相互平衡，从而抵消其对扭转运动的净影响。
- 局部质量对称 ： 大铅球在小铅球周围对称放置，以减小引力梯度的不对称性。

附近对象的隔离

- 密闭外壳：实验装置被安置在密闭的空间内，以消除空气流动并减少温度变化的影响。
- 非磁性材料：选用铅是因为其具有非磁性特性，可避免来自外部磁场等磁性干扰对实验造成影响。

关键发现

- 卡文迪许确定了万有引力常数 G　6.754×10^{-11} $\mathrm{N\cdot m^2/kg^2}$，非常接近当前该常数的计算值 (6.6732×10^{-11})。
- 根据 G 的值, 他计算出了地球的平均密度（水密度的 5.481 倍）和质量。

卡文迪许实验的误差源

环境扰动

- 气流：即使是很微弱的空气流动，也可能导致扭秤产生虚假的偏转。
- 温度波动：悬挂金属丝或铅球的热胀冷缩可能改变所测得的扭矩，影响实验精度。

仪器的局限性

- 轻微的非理想质量：铅球并非完美的质点，这种实际质量分布与理想点质量的偏差会导致对牛顿平方反比定律的偏离。
- 扭丝缺陷：金属丝弹性不均匀或制造过程中的瑕疵可能影响扭力常数的准确性，进而干扰测量结果。

测量不确定性

- 轻微的光学对准误差：镜面未对准或在追踪光斑位移时出现刻度视差，会影响角度测量的准确性。
- 周期性振荡问题：扭力振荡的阻尼不完全，导致在测量振荡周期时产生计时误差，进而影响扭转常数的确定。

即使能定义自包含系统，确定苹果落地的原因仍然充满挑战。通过牛顿的工作，我们知道原因对象是地球，影响机制是万有引力。

第三，地球对苹果的影响本身是一个纯粹的物理问题，讨论它的“价值”在物理层面没有意义。但如果我们考虑地球的影响延伸到了苹果的甜度——这是果农或消费者真正在意的特性——那么在这个语境下，评估地球的价值就变得有意义了。这一点在我们设想将苹果树移植到月球或其他潜在位置时尤为关键，那时地球作为原因对象的角色可能会被替代或改变。我们在第 十一章讨论了价值问题。

最后，即便只是研究地球对苹果的影响，也存在多种视角。例如，地球不仅提供引力，还为苹果树的种植提供了不可或缺的生存环境，包括适宜的温度、大气、水分和土壤等条件。

对象因果模型

可以基于牛顿的万有引力定律和其他物理领域知识构建卡文迪许实验的对象因果模型。

13.1.2　一个化学例子：拉瓦锡的燃烧实验

案例二是一个著名的化学实例：拉瓦锡的燃烧实验 [168]，由王磊撰写。

问题背景

拉瓦锡的燃烧实验被认为是化学史上的一个里程碑，因为它揭示了氧气在燃烧过程中的关键作用。在 18 世纪，化学界普遍信仰“燃素说”（phlogiston theory），该理论主张所有可燃物中都蕴含一种称为“燃素”的隐秘物质，燃烧的过程即是物质释放燃素至空气中的过程。在拉瓦锡的燃烧实验之前，人们认为原因对象是“燃素”，且燃素是这些直接影响对象内在的一部分。

拉瓦锡的燃烧实验揭示导致 AO 燃烧的原因对象实际上是参与反应的氧气，而非 AO 释放的燃素。他通过严谨的实验和测量推翻了燃素说，确立了质量守恒和氧化反应的基本原理，为现代化学奠定了基石。

评价学的基本概念有助于解释拉瓦锡的实验。

现实世界系统

现实世界中存在大量的燃烧现象，可燃物在一定条件下会燃烧。在燃烧现象中，直接影响对象（AO）是指那些能够被燃烧的物质，如汞、磷和硫等。我们可以直接测量这些直

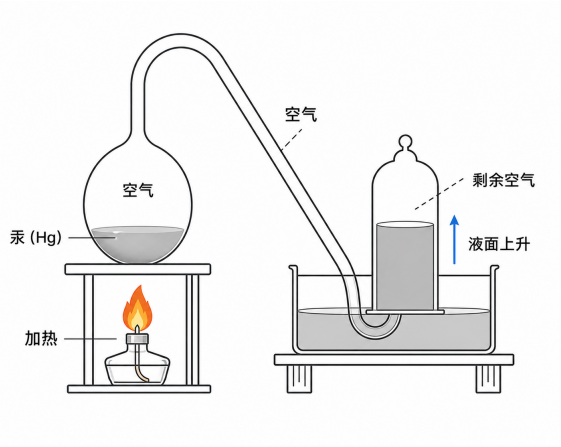


图 13.3: 拉瓦锡经典燃烧实验示意图。

接影响对象 AO 的量，如质量、气体体积和温度的变化。而原因对象是导致 AO 燃烧的对象。

自包含系统

拉瓦锡的经典燃烧实验是其对物质转化和空气成分研究的核心实验，该实验设计巧妙且执行严谨，如图 13.3 所示。

在这个实验中，他将汞置于蒸馏烧瓶中加热，下方设有汞收集器以捕获蒸发的汞蒸气。蒸气随后通过管道导入单独的容器，在那里通过冷凝器冷却并重新液化。实验分三个阶段进行：

在第一阶段，加热持续了十二天，在此期间，汞与空气中的成分发生反应，生成了红色的氧化汞。系统内的气体体积缩减了约五分之一，而汞和氧化汞的总质量保持不变，这初步支持了质量守恒的原理。在第二阶段，使用第一阶段反应后剩余的气体与汞一同加热，汞不再发生氧化，也观察不到燃烧现象。在第三阶段，收集得到的氧化汞在蒸馏瓶中经受高温分解，生成了一种能够支持燃烧和呼吸的气体。值得注意的是，这种气体的体积恰好填补了之前缩减的五分之一空间，而剩余的五分之四气体则不具备这样的性质。

通过这一实验，拉瓦锡不仅证明了汞的气化和氧化过程是可逆的——证实了质量并非消失而是发生了转化——而且还最终证明了空气由“可燃的氧”（约占五分之一）和“惰性的氮”（约占五分之四）组成。

这一开创性的发现有效地检验并推翻了燃素说（phlogiston theory），确立了质量守恒和氧化反应的基本原理，为现代化学奠定了基石。

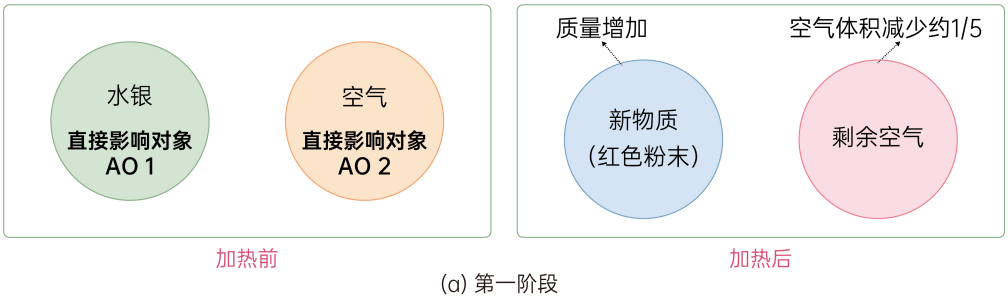
该实验的目标在于找出引发燃烧的主要原因对象。因而可以视为逆评价：原因对象是某种有待确定的神秘物质。

拉瓦锡设计了以封闭空间为中心的自包含系统（SCS），这些装置将其他对象与原因对象、影响对象和混杂对象隔离开来。

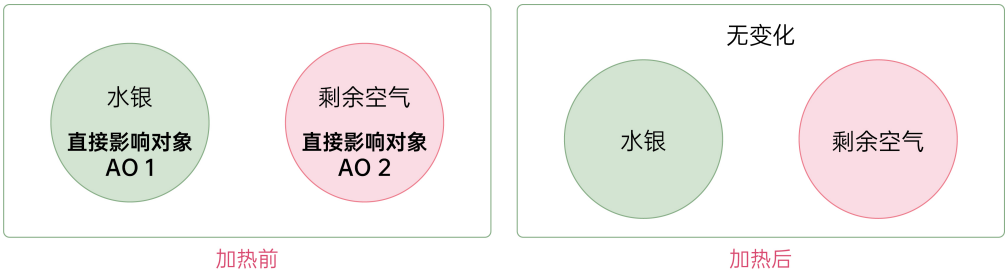
他将汞（直接影响对象 AO）密封在蒸馏瓶中加热，观察密闭空间内气体体积的变化，并收集燃烧产生的气体进行分析。

关键步骤包括：

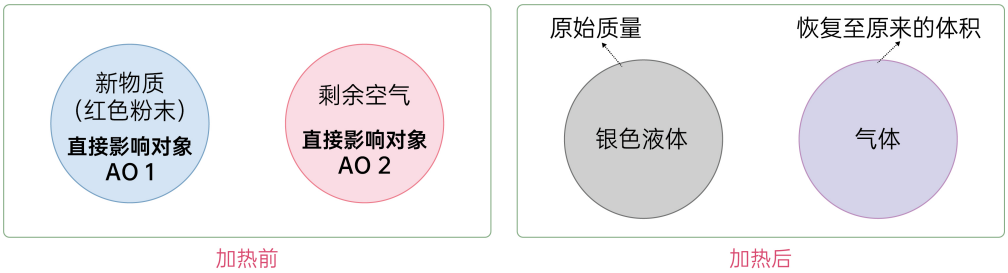
- 加热汞：将汞密封在玻璃容器中，汞形成红色氧化汞（HgO），空气体积大约减少了五分之一。
- 气体分析：剩余的气体（后来被称为氮气）无法支持生命或燃烧。
- 氧化汞的分解：加热时，HgO 释放出一种高度可吸入的气体（氧气），气体体积相应恢复。



(a) 第一阶段



(b) 第二阶段



(c) 第三阶段

图 13.4: 拉瓦锡实验：加热后三个阶段的变化。

与其他案例显著不同，案例二具有独特的特点：封闭空间能够隔离其他对象的影响；然而，尽管处于同一封闭空间，直接影响对象 AO 和混杂对象 COs 却在不同的阶段发生变化，如图 13.4所示。

第一阶段，存在两个直接影响对象 AO：汞（银色液体）和空气，当时拉瓦锡对后者的了解有限。加热后，汞发生变化，生成了一种新物质（红色粉末），其质量增加；空气体积减少约五分之一。

第二阶段，存在两个直接影响对象 AO：汞（银色液体）和第一阶段剩余的空气。加热后，汞未发生变化，未观察到燃烧现象。

第三阶段，存在两个直接影响对象 AO：新物质（红色粉末）和第一阶段产生的剩余空气。加热后，红色粉末变为银色液体，质量恢复原状，并释放出一种高度可吸入的气体。最后，气体恢复至原来的体积。

通过比较不同阶段的自包含系统 (SCS)，同时使用封闭空间隔离其他对象的影响，拉瓦锡推导出燃烧依赖于空气中的一个反应成分（后来被命名为氧气），且按体积计算，氧气约占空气的五分之一。他最终推翻了盛行的燃素理论，该理论认为燃烧是由于“燃素”的释放，并确立了“燃烧是物质与氧气的氧化反应”的科学理论。

此外，他还使用了其他可燃材料（直接影响对象 AO），如磷和硫，来验证燃烧机制的普遍性。

对象因果模型

拉瓦锡揭示的对象因果模型可以用化学方程式进行表示：

- 燃烧（*Combustion*）：$2\mathrm{Hg} + \mathrm{O_2} \xrightarrow{\Delta} 2\mathrm{HgO}$ （银色液体 → 红色粉末）
- 分解（*Decomposition*）：$2\mathrm{HgO} \xrightarrow{\Delta} 2\mathrm{Hg} + \mathrm{O_2}\uparrow$ （红色粉末 → 银色液体 + 气体）

此外，拉瓦锡得出了以下结论：1）燃烧时质量的增加源于氧的吸收，而非“燃素”的释放；2）空气大约由 20% 的氧气（对燃烧至关重要）和 80% 的氮气组成。

尽管关于发现的优先权和可解释性存在争议，拉瓦锡使用精确的天平和严谨的质量计量方法巩固了氧气燃烧理论。他的实验表明，燃烧是物质与氧气之间的化学反应，根本改变了我们对化学过程的理解。他的工作发现的两个关键定律为现代化学奠定了基石：1）质量守恒定律，即反应中的总质量保持不变；2）氧化理论，即燃烧是物质与氧气的氧化反应。

13.1.3　一个生物学例子：格雷戈尔 · 约翰 · 孟德尔的豌豆

生物性状，比如说人的身高和体型、猫的直毛或卷毛、玉米的甜糯，代表了生物个体之间可观察到的差异。这些生物性状由什么决定？为什么高茎豌豆和矮茎豌豆杂交产生的后代全都是高茎？而这些高茎豌豆自花授粉产生的某些后代会出现矮茎？

格雷戈尔 · 约翰 · 孟德尔给出了这些问题的答案。通过精心设计的豌豆实验（案例三），格雷戈尔 · 约翰 · 孟德尔提出了经典的基因分离定律和自由组合定律 [193]。案例三由王晨曦撰写。

问题背景

植物的性细胞统称为配子，分为雄性配子和雌性配子。生物体的性状一般都是由等位基因决定的，但是特殊情况下，不能完全匹配。例如，XY 染色体（部分植物拥有）是决定生物性别的染色体，不一定完全匹配，因而没有等位基因。

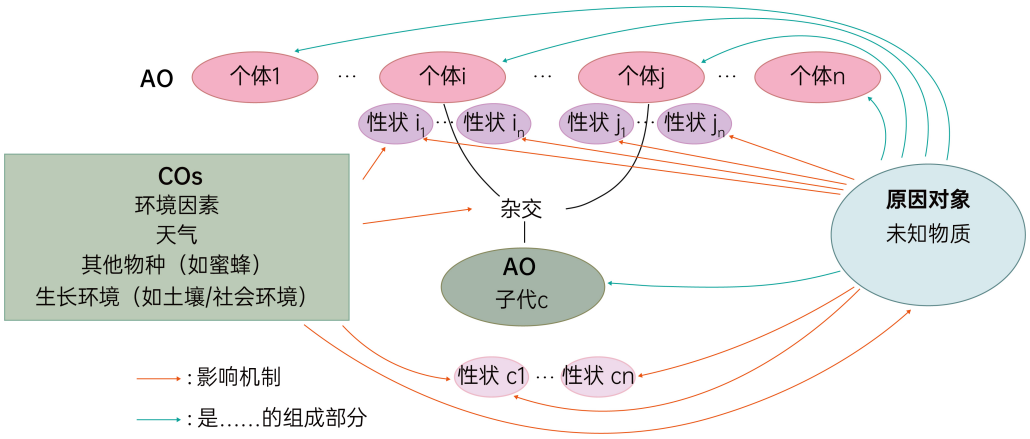


图 13.5: 自然环境下生物性状遗传的现实世界系统。这个系统里的原因对象、AO 以及 COs 之间存在复杂的关系。

纯合子指决定遗传性状的的两个等位基因相同的生物体。

现实世界系统

自然环境下生物性状遗传的现实世界系统如图 13.5 所示。在自然条件下，直接 AO 是生物体亲本；混杂对象集合 COs 包括自然环境条件，如天气，其他生物体，如蜜蜂，生长环境，如土壤环境等；原因对象是某种未知的遗传物质；原因对象对 AO 的影响较为复杂：首先是 AO 开花，授粉（包括自花授粉和异花授粉），产生种子（影响对象的新结构），最后种子发芽成长为子代（新对象）及其性状的出现。生物性状可以直接观察。在自然条件下观察生物体可以发现子代既有和亲本相似的性状，也有表现出和亲本完全不一样的性状。

自包含系统

与本章的其他案例不同，案例三有一个独特的特点。当两株豌豆杂交时，它们的基因都是原因对象，而直接 AO 则是进行杂交的豌豆及其子代基因，子代豌豆的出现是影响（新对象）也是间接 AO（其基因受原因对象的影响）。

亲本基因对亲本豌豆的影响是决定了其表现的生物性状，不同基因下豌豆的性状不同，如豌豆的高茎与矮茎。而亲本基因杂交的直接影响结果是子代基因这个新对象的产生，同时子代基因作为原因对象又决定了子代豌豆（子代基因影响的直接 AO，亲代基因影响的间接 AO）的性状。图 13.6 展示了原因对象、AO 和 COs 之间复杂的关系。

孟德尔选择花园豌豆作为 AO 极具眼光，也是他能成功发现孟德尔遗传定律的关键：

- 纯种品系：豌豆可以自花授粉，孟德尔鉴定出一些品种，当自花授粉时，它们能持续且稳定地孕育出与亲本在特定性状上（例如，种子形状）相同的后代。这些是他的 P 代（亲本）。
- 易于区分、可遗传的性状：他选择了七种表现清晰、对比鲜明的性状（参见表 13.1），例如种子的圆粒与皱粒。这些性状中没有模棱两可的中间性状。

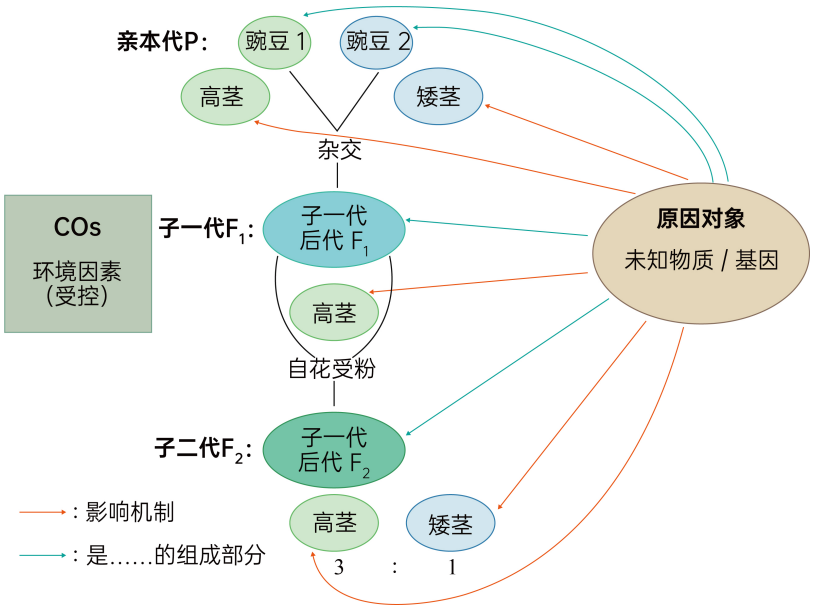


图 13.6: 格雷戈尔 · 约翰 · 孟德尔豌豆实验 [193] 的 SCS。这个 SCS 里的原因对象、AO 以及 COs 之间存在复杂的关系。

- 受控杂交：豌豆花的物理结构使孟德尔能够轻松地防止其自花授粉，并对选定的亲本进行异花授粉（杂交）。
- 产生子代的时间短：可以在合理的时间范围内研究多代性状。

性状 Trait	显性性状 Dominant Trait	隐性性状 Recessive Trait
种子形状 Seed Shape	圆粒 Round	皱粒 Wrinkled
种子颜色 Seed Color	黄色 Yellow	绿色 Green
花的颜色 Flower Color	紫色 Purple	白色 White
豆荚形状 Pod Shape	饱满 Inflated	缢缩 Constricted
豆荚颜色 Pod Color	绿色 Green	黄色 Yellow
花的位置 Flower Position	腋生 Axial	顶生 Terminal
茎长 Stem Length	高茎 Tall	矮茎 Dwarf

表 13.1: 孟德尔研究的七种豌豆性状 [193]。

孟德尔最初花费数年时间让豌豆自花授粉，从而筛选出能够稳定遗传特定性状（后代表现出相同的性状）的品种，称为“纯合子”。例如，高茎豌豆的后代将始终是高茎。他最初的实验一次只关注一种性状，即单基因杂交。

- 他对两株在同一性状上（例如，高茎与矮茎）表现不同的纯种（纯合子）豌豆进行杂交。

- 他收集杂交产生的豌豆种子，并进行种植—这就是子一代（F_1）。

观察一： F_1 子一代总是只与双亲之一表现出相同性状。例如，纯种高茎与纯种矮茎豌豆杂交产生的所有子代都是高茎。孟德尔称在 F_1 子一代中出现的性状为显性性状（高茎）。似乎消失的性状称为隐性性状（矮茎）。

孟德尔随后让 F_1 子一代豌豆自花授粉，产生子二代（F_2）。

观察二：在 F_1 子一代中消失的隐性性状，在 F_2 子二代中重新出现了！此外，孟德尔统计了每种性状的数量，发现了一个近似的稳定比例：*3 显性性状: 1 隐性性状*。

为了解释这些结果，孟德尔提出了一个革命性的遗传模型 [193]：

- 遗传因子：遗传由离散的“因子”（现在称为基因）控制，这些因子不经改变地从亲代传递给子代。
- 等位基因：对于每种性状，子代生物体从每个亲本那里各继承一个遗传因子，总共获得两个遗传因子。这些决定同一组生物性状基因的不同形式称为等位基因（例如，高茎基因和矮茎基因）。
- 显性基因/隐性基因：在杂合子（具有两个不同等位基因）中，显性基因决定生物体的外观，而隐性基因没有明显效果。
- 基因的分离定律：遗传性状的两个等位基因在配子（性细胞）形成过程中彼此分离，因此每个配子只携带每种性状的一个等位基因。

孟德尔接着问道：当两种性状同时遗传时会发生什么？为了回答这个问题，他进行了双基因杂交，杂交的亲本在两种性状上存在差异—例如，一个纯种的圆粒黄色种子豌豆（$RRYY$）和一个纯种的皱粒绿色种子豌豆（$rryy$）。

- F_1 子一代：所有子代都是 $RrYy$，并具有圆粒黄色种子。
- F_2 子二代：当 F_1 子一代植物自花授粉时，F_2 子二代表示出四种性状，比例稳定：

9 圆粒黄色 (1 RRYY + 2 RRYy + 4 RrYy + 2 RrYY) : 3 圆粒绿色 (1 RRyy + 2Rryy) : 3 皱粒黄色 (1 rrYY + 2 rrYy): 1 皱粒绿色 (1rryy)

9:3:3:1 的比例是一个双基因杂交比例，可以从两个独立的单基因杂交的 3:1 比例推导出来 ((3:1) × (3:1) = 9:3:3:1)。这引出了基因的自由组合定律。

- 基因的自由组合定律：决定不同性状的基因在配子（性细胞）形成过程中是独立分配的。一个配子获得一种性状的等位基因不影响它获得另一种性状的等位基因。

在孟德尔的豌豆实验中，原因对象是参与杂交的等位基因，AO 指的是亲本豌豆（具体参与的是它的性细胞），它可以被任何生物体所替代，AO 的影响体现为开花、结果及其它们配子的等位基因变化。

对于亲本基因来说，它作为原因对象影响亲本开花、结果并且决定着子代的基因。影响机制是基因的分离定律和自由组合定律。子代基因直接影响着子代豌豆的性状，因此对于亲本基因来说，子代豌豆是其影响的间接 AO。

对象因果模型

通过对亲本已知基因组按照基因分离定理和自由组合定律进行计算，可以预测子代遗传的基因组以及表现出的性状，这就是本案例的对象因果模型。具体来说，这个对象因果模型中对象有父本基因 AA，父本生物体 F，母本基因 aa，母本生物体 M，亲本杂交产生的子代基因 Aa 以及子代生物体 C。对象因果模型建模了以下对象之间的因果关系：

- *基因决定生物性状：* 原因对象是生物体基因，直接 AO 是携带此基因的生物体。$AA \Rightarrow F$、$aa \Rightarrow M$、$Aa \Rightarrow C$。
- *孟德尔遗传定律（如基因的分离定律）：* 原因对象是亲本基因，直接 AO 是子代基因。$AA \Rightarrow Aa$、$aa \Rightarrow Aa$。
- *生物体遗传现象：* 原因对象是亲本基因，通过影响子代基因来间接影响子代生物体（间接 AO）。$AA \Rightarrow Aa \rightsquigarrow C$、$aa \Rightarrow Aa \rightsquigarrow C$。

13.1.4　一个天文学例子：开普勒三大定律

案例四是一个天文学例子：开普勒三大定律 [151, 152]，由王磊撰写。

问题背景

天文学研究的对象多为宇宙中真实存在的天体，其运行过程在现阶段无法进行人工干预，只能进行自然观察。研究者难以像卡文迪许扭秤实验那样精心设计受控实验，人工屏蔽混杂对象 COs 的影响；或是和药物评价一样开展随机对照实验，随机分配 AO 到干预组和对照组；天体运行轨道的研究高度依赖于长期自然观察的数据。开普勒正是基于第谷二十多年积累的高精度行星位置观察数据，通过严密的逻辑推理与数学计算，最终归纳出行星围绕太阳运动的三大定律：开普勒第一定律（轨道定律）指出，太阳系行星的运行轨道都是一个以太阳为焦点的椭圆；第二定律（面积定律）指出，连接行星和太阳的线段在相等时间间隔内扫过的面积相等；第三定律（周期定律）指出，对于绕太阳运行的所有行星，其轨道周期的平方与其轨道半长轴的立方之比为常数。

现实世界系统

在恒星系统（如太阳系）中，中心恒星（原因对象）的引力会对围绕其运行的行星（直接影响对象 AO）的运动轨迹产生影响。在开普勒之前的经典天文学中，由于深受“完美匀速圆周运动”这一先验假设的束缚，例如在托勒密地心体系的表示中行星沿本轮（epicycle，即较小的圆形轨道）运动，而本轮的圆心又沿均轮（deferent，即较大的圆形轨道）运动，因此研究者往往试图通过复杂的本轮与均轮叠加来拟合观察数据。

开普勒基于第谷长达二十多年高精度的行星观察数据，通过跨星体的几何推算，在数学上成功剔除了由于地球公转与自转带来的观察位移，揭示了太阳对行星产生的真实影响，即行星并非做完美的匀速圆周运动，而是存在轨道偏心与速度的周期性变化。

在开普勒研究的现实世界系统里，原因对象 UO 是太阳（太阳系中唯一的恒星），影响对象 AO 是太阳系中的行星（以火星为典型代表），混杂对象 COs 是宇宙中其他天体。推导的核心在于预测太阳系行星的运行轨道，也即预测太阳对行星运行轨道产生的影响。

第谷长达二十多年在地球上观测到的行星运行轨道数据，实际上是原因对象和混杂对象对影响对象的总体影响（Overall Effect）。这些数据不仅包含了太阳（EO）对行星（AO）的影响，还混杂了宇宙中其他天体对行星的影响以及由于地球自身运动（公转、自转）与观察角度变动所带来的偏差。开普勒基于第谷观测记录的太阳系行星运行轨道数据，通过跨星体的交叉分析、敏锐大胆的数学直觉以及严密的几何计算，成功拟合出第谷的观察数据。

对象因果模型

开普勒揭示的行星运动的三大定律（对象因果模型）描述如下为 $B \Rightarrow C$，B（太阳）和 C（火星）均为显式对象，但存在其他对象 B_i（如地球、其他天体）对观察结果有影响且无法屏蔽。直接观察到的是总体影响，而开普勒通过求解几何与数学方程，成功推断出了 B 对 C 的真实影响（True Effect）。公式如下：

- 轨道定律 *Ellipses:*

$$r(\theta) = \frac{p}{1 + e\cos\theta} \quad \left(\begin{array}{l} r: \text{极径，B 到 C 的距离;} \\ \theta: \text{极角，B 到 C 连线相对于某个参考方向的角度;} \\ p: \text{半正焦弦，描述 C 轨道的形状和尺度;} \\ e: \text{离心率，描述 C 轨道的形状} \end{array} \right)$$

- 面积定律 *Equal Areas:*

$$\frac{dA}{dt} = \text{constant} \quad \left(\begin{array}{l} A: \text{B 到 C 的向径扫过的面积;} \\ t: \text{时间;} \\ \text{constant}: \text{面积速率常数} \end{array} \right)$$

- 周期定律 *Harmonic Law:*

$$\frac{T^2}{a^3} = k \quad \left(\begin{array}{l} T: \text{C 绕 B 公转周期;} \\ a: \text{C 轨道半长轴;} \\ k: \text{开普勒常数} \end{array} \right)$$

上述定律揭示了评价对象–太阳（太阳系中唯一的恒星）对影响对象–太阳系中的行星（以火星为典型代表）运动的影响。

另一方面，开普勒三大定律虽然刻画了第谷的观察数据，实现了数学上因果规律的总结，但开普勒本人并没有给出太阳系行星运行轨道为什么是椭圆的物理因果解释。站在开普勒肩膀上回答这个问题的正是艾萨克 · 牛顿，他提出的万有引力定律解释了太阳对行星的真实影响是万有引力，太阳系行星在太阳万有引力作用下绕太阳运动。因此，万有引力定律是一个对象因果模型，可以预测任何两个物体之间的万有引力大小。值得特别说明的是，开普勒第三定律（周期定律）正是牛顿万有引力公式中距离平方反比的基石，而卡文迪许则通过扭秤实验测量得到了牛顿万有引力公式中的万有引力常数。

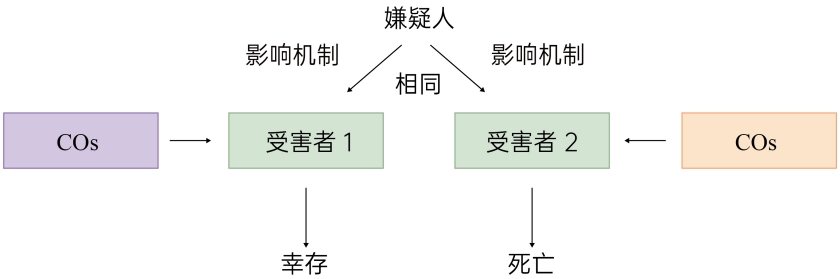


图 13.7: 嫌疑人用同样强烈的暴力言语对两名受害者进行持续恐吓，但这两名受害者受到的总体影响（可观察的结果）却明显不同。

13.1.5　一个法律责任示例：确定嫌疑人责任。

案例五是一个法律案例，这里简单地分析它的现实世界系统，因为伦理原因，实验显然并不适合。

假设有一名嫌疑人使用强度相同的暴力言语，持续恐吓两名受害者，并因此可以认为它产生了相同的影响。其中一名受害者家庭支持有力，生活没有压力；而另一名受害者则独自生活，经济拮据，本身已承受较大压力。前者最终幸存，而后者不幸选择自杀。

该案例呈现出两个独特特征：作用机制明显不同于其他案例 。这里的原因对象是暴力言语，用精神的方法施压影响对象。原因对象及其混杂对象的影响机制及其影响都很难被检验。其次，混杂对象（COs）成为决定总体影响的关键 。尽管两名受害者面对的原因对象及其影响（可能难以检验）相同，但由于各自所处环境不同，最终的总体影响却截然不同，呈现出显著差异的结果，如图 13.7所示。

13.1.6　一个社会科学例子：政策评价

案例六是一个社会科学的例子，由祝弘华和康国新撰写。

问题背景

20 世纪 90 年代，冰岛青少年精神活性物质使用问题较为突出。1999 年调查显示，10 年级学生中曾吸烟者占 56%，过去 12 个月内曾醉酒者占 56%，曾使用大麻者占 15%。这说明，青少年吸烟、饮酒和大麻使用已成为需要公共治理回应的现实问题 [121]。

面对这一形势，冰岛政府与地方社区自 20 世纪 90 年代中期开始推行“冰岛预防模式”（Icelandic Prevention Model，IPM）。该政策以降低青少年精神活性物质使用率为目标，以系统改善家庭、学校和社区环境为主要路径。

二十年后，同一批指标令人震惊地逆转：至 2015 年，曾吸烟者降至 16%（欧洲仍为 46%），过去 30 天内饮酒者降至 9%（欧洲 48%），大麻终身使用率降至 5%（欧洲 16%）——在参与欧洲学校酒精与毒品调查（ESPAD）的 35 个国家中位居最低或次低 [162]。这一现象构成了一个典型的评价问题：在既有观察数据的情况下，研究者需要从影响对象 A 上可测量指标及其随时间的变化，追溯原因对象及混杂对象对 AO 的影响，最终推断出准确性高的影响模型 [163, 164]

现实世界系统

在冰岛预防模式（IPM）的案例中，原因对象 UO 是冰岛预防模式这一政策，影响对象 AO 是冰岛青少年群体，构建自包含系统的挑战是准确地识别可能的混杂对象，将其引入自包含系统。

混杂对象 COs 可能包括家庭、学校、同伴群体、社区组织、课外活动体系以及地方政府与社区协作网络等。评价的核心在于，通过观察青少年行为与状态的变化，推测该政策对青少年产生的真实影响。

在这一现实世界系统 RWS 中，青少年吸烟、饮酒、大麻使用、课后活动参与、与父母相处时间等观察数据，反映的并不是政策单独施加的影响，而是政策与 COs 共同作用于青少年所形成的总体影响（Overall Effect）。也就是说，观察结果既包含了政策干预的作用，也混杂了家庭监督、学校参与、同伴关系、社区活动供给以及地方资源配置等多种社会因素的影响。因此，评价的关键并不只是记录青少年行为是否发生变化，而是要在这些总体变化中识别并分离出政策所带来的真实影响（True Effect）。

IPM 的特点在于它并非一次性干预，而是一个持续迭代的政策系统。通过组建地方联盟、建立长期资金支持、开展年度普查、反馈社区数据、设定本地目标并调整政策实施配置，政策在每一轮循环中不断修正其配置，并逐步改进对政策影响的估计。这种复杂的动态迭代过程增加了建立对象因果模型的挑战。

进一步对该案例（案例六）进行分析，可以得出三点观察结论：

首先，政策的具体实施难以被精确控制，也难以在其他地区复制。

第二，混杂对象 COs 识别难度更高 ，参与总体影响的混杂对象更难识别和控制。宏观经济条件、媒体文化环境乃至青少年对政策的主观态度都是难以操纵的社会力量。

IPM 的应对策略是以年度全校普查作为一种调查的手段：每年对各学区所有在校学生（而非样本）进行匿名问卷调查，系统追踪各种因素（混杂对象）以及可测量指标随时间的变化趋势（数据回收率通常超过 80%）。这一持续监测机制为每个社区建立了动态比较的基线。

第三，不同参与者（即直接 AO）之间的相互影响也显著影响评价结果，如图 13.8 所示。每个直接 AO 的状态和行为都可能受到其他人的状态和行为的影响，呈现出明显的社会网络效应或群体互动特征。例如，当某所学校主流的同伴行为规范转向不使用成瘾性物质时，个体的” 默认选择” 便可能发生根本改变；同样，关键意见领袖的行为转变可能触发连锁效应。这意味着，EO 对单个 AO 的影响并不等于 EO 对整个 AO 群体的净影响——AO 之间还传递了间接影响。

这一讨论揭示了社会政策评价的复杂性：不仅实施难以标准化，且受主观态度和人际互动影响深远，使得因果推断更具挑战性。

13.2 前瞻性案例

13.2.1 一个药理学示例：药物评价

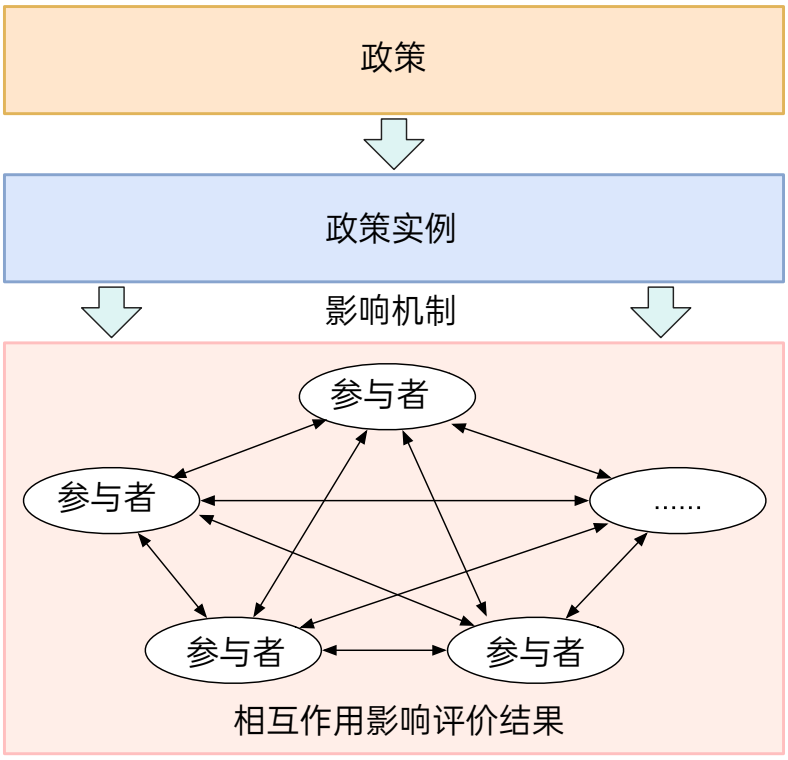


图 13.8: 在政策评价中，直接 AO 的相互影响会显著影响结果。

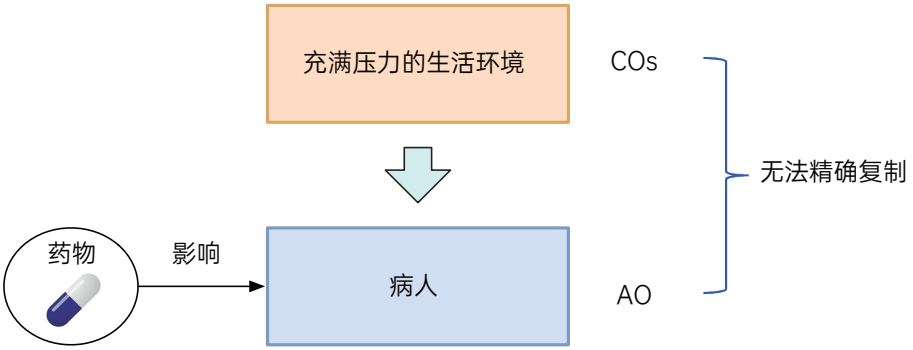


图 13.9: 在药物评价的案例里，无法精确地复制直接影响 AO 和混杂对象 COs。

案例七涉及对一种药物的评价，我们在此简单地讨论它的自包含系统。EO 指的是特定药物，而直接 AO 则包括这些干预措施所针对的受试者（人类）。图 13.9 展示了 SCS 的构成。

简单地分析案例七，可以得出三点观察结论：直接的 AO 与 EO 相互独立；总体影响可以在直接 AO 上进行测量或测试，其中包含混杂对象（COs）的影响，例如患者承受的压力以及患者所生活的环境。

最后但同样重要的一点是（COs）与直接（AO）的特性。在案例七中，我们无法非常准确地复制 COs 或直接 AO。例如，我们无法精确复制一个患者，更不用说包含在 COs 中的其他关键因素。这一区别凸显了药物评价相较于其他案例在外部有效性上的挑战，尤其是在涉及人类受试者时，个体差异和环境复杂性决定着研究成果的有效性。

13.2.2　一个教育案例：学习干预的效果评价

案例八探讨的是对教育干预措施（如数字化学习平台或新型教学方法）的评价，由高婉铃撰写。

现实世界系统

这个案例中，教育干预本身——如特定的 AI 辅助教学系统或教学策略——是评价对象，个体学习者是直接影响对象。所观察到的评价结果并非由 EO 与 AO 单独决定，现实世界评价系统中必然包含一系列外部情境因素（COs），如教师专业水平、课堂环境、社会经济背景乃至家庭支持。这些 COs 与学习过程深度交互，在研究者试图识别 EO 的真实影响时，COs 会造成严重的混杂，使得研究者难以在有限的自然观察中剥离出干预措施的真实影响。

值得强调的是，该模型中的 EO、AO 与 COs 均不是静态实体，而是随时间演化的系统，共同构成一个反映学习动态的时间序列。

出于伦理原因，教育干预措施的评价可能无法采用实验方法。

对象因果模型

对象因果模型可能能为上述场景提供数学模型，体现 EO 和 COs 对 AO 的影响。

例如，模型可以模拟 “若同一学习者在不同的家庭支持水平（COs）下接受同一教学干预（EO），其学习成效将如何演变”。这种数学化的重构，使评价从局部的经验描述转向了对因果关系的推测，不仅实现了对 COs 混杂的校正，并且能提高教育干预评价的泛化能力。

13.2.3 对抗演习评价：系统性对抗的一个案例

案例九是一个系统性对抗的场景，不同于被动对象，防御方与进攻方具有自由意志。案例九由范帆达撰写。

在系统性对抗中，例如防御方与进攻方之间的防御演习，评价的核心问题并不仅仅是判断谁胜谁负。更重要的是，要从演习的总体结果中揭示出评价对象（EO）的实际影响，例如一个防御型指挥与控制系统在实战环境中的真实影响。

评价的挑战在于观察到的结果（例如成功拦截攻击的次数、响应延迟或系统生存能力）是多种相互交织因素共同作用的结果，并不仅限于评价对象（EO）本身。因此，要进行有效的评价，必须构建一个自包含系统（SCS），以涵盖所有相关的交互对象，并能够准确归因不同对象的真实影响。

在这种对抗场景中，自包含系统（SCS）与其他案例不同，呈现出动态性和相互依赖的特点：每个对象不仅会对其他对象做出反应，还可能主动改变自身的后续状态；进一步，EO 做出的每一个决策都会引发 AO 的反制反应（行为），而这反过来又会改变 EO 的运行状态。因而，EO, AO 以及 COs 的角色本身也会发生变化。这种循环因果关系挑战了传统的线性因果分析模型，凸显出必须将评价视为一种 系统性因果推断过程 。

对抗性场景为检验评价学（Evaluatology）的基本原理提供了丰富而真实的背景。在对抗性或竞争性系统中，评价学的作用因此从验证结果扩展到揭示生成这些结果的 影响结构 。通过明确建模 EO–AO–COs 之间的相互作用，评价者即便在持续适应与对抗的环境中，也能分离出系统的真实影响，从而实现评价学的最终目标：在一个自包含但动态交互的世界中，有效地揭示因果影响。

13.2.4 暗物质

案例十有关暗物质，由王磊和王晨曦撰写。

问题背景

在开普勒、牛顿以及卡文迪许向人们揭示天体运行轨道规律以及万有引力定律后，天文学家的天体观察仍在继续，数百年来积累了大量观察数据。大量的宏观测量实验表明宇宙中除了能够被人们观察到的天体（物质）外，还可能存在大量未被人们所观察到的物质和能量 [25]。

20 世纪 70 年代女天文学家薇拉 · 鲁宾（Vera Rubin）观察螺旋星系时发现，星系边缘的恒星旋转速度极快，该星系的万有引力远不足以提供圆周运动的向心力 [256]；同时天文学家发现，某些地方的光线弯曲程度极其夸张，有可能存在大量隐藏的物质。基于这些现象，科学家推测系统中存在大量弥散的暗物质。暗物质还未被人们所直接观察到，物理学界对于暗物质到底是什么仍处于猜想阶段 [25]。

基于自包含系统的前瞻实验

物理学有关暗物质的主流猜想是弱相互作用大质量粒子（WIMPs），物理学界已经设计类似卡文迪许扭秤实验的类似装置来捕获暗物质。例如，科学家在中国锦屏山地下 2400 米（PandaX 实验）和其它深地实验室放置巨大的液氙探测器，希望捕捉暗物质粒子穿过地球时与原子核的碰撞 [44]。在此类前瞻性实验中：

- **原因对象：**暗物质粒子（预期产生碰撞能量的源头）。
- **直接影响对象（AO）：**探测器介质（如液氙原子核），通过记录其受撞击后产生的闪光（电子信号）来表征影响。
- **混杂对象（COs）：**宇宙射线、环境放射性噪声等。
- **自包含系统（SCS）：**可以剔除其他对象的干扰。例如，将探测器部署在中国锦屏山地下 2400 米深处，利用巨厚地壳岩层作为天然屏蔽体。

从前瞻实验角度，若要成功捕获暗物质产生的真实影响（True Effect），AO 的选取与部署位置至关重要。目前液氙探测器尚未获取确凿证据，这引发了方法论上的深层讨论：是否需要更换对暗物质更敏感的 AO 探测介质？以及如何进一步排除其他对象的干扰？这些都是科学家还在探索的问题。

13.2.5 另一个物理学例子：理论和模型的检验

案例十一涉及到验证一个表面物理模型的理论，理论验证的问题可以见于任何学科。迈克尔 · 斯克里文博士曾以该案例 [274] 说明评价科学在几乎所有学科中的基础性作用。然而，我认为这不应被归类为“评价”，而应属于“测试”范畴，因为该实践的目的在于确认理论或模型是否符合测试所依据的“检验基准”（test oracle）。我们在第 6.10节讨论了如何应用测试学来验证宇称不守恒理论。

13.3 本章小结

本章讨论了评价学通用方法在不同学科领域的应用。根据不同案例的性质，有些采用了观察的方法，有些则采用了实验或者建模的方法。这些讨论展现了评价学通用方法的潜力。

本书的第 六部分将提供评价学通用方法的一些具体应用。

第十四章

实验设计

本章将介绍实验设计（Design of Experiments，DoE）的基本概念、问题定义、假说（assumption）、原理、方法学以及案例研究。王晨曦是本章的主要贡献者，我从评价学的角度进行了浅显易懂地解读，并对全文进行了修改。

14.1 实验设计的基本概念

本节介绍的实验设计的基本概念基于前人编写的经典著作 [138, 307, 155, 87, 86]。

DoE 概念定义 1 (实验单元 (experimental units))**.** 实验单元是进行实验研究的对象，既可以是一个个体，比如说一个人，也可以是集体（比如说一个班级）。

实验单元的含义和评价学中的影响对象的含义相近，但 DoE 没有区分对象、系统和组件的概念。更没有区分不同的对象性质。

DoE 概念定义 2 (响应 (response))**.** 与分类（*categorical*）变量不同，响应是在实验单元上通过测量得到的数值指标*(quantitative metrics)*。

响应的概念对应于评价学中对象的量，一种可计数的属性。

DoE 概念定义 3 (因子集（factors）)**.** 因子集是一组参数（*parameter*）[1]，改变参数取值将导致测量得到的响应变化。一个因子有多个水平（*level*)。因子水平既可以是分类变量又可以是数值变量。

因子的概念对应于评价学对象的属性。

DoE 概念定义 4 (交互（interaction）)**.** 交互是因子之间系统性的依赖关系。

DoE 概念定义 5 (实验条件组（Experimental Condition Groups）)**.** 实验条件组是遍历所有因子水平取值的组合。举例来说，如果总共有 n 个输入因子，每个因子 i 有 m_i 个因子水平取值，那么总共有 $\prod_{i=1}^{n} m_i$ 个实验条件组。

实验条件组不同于评价学的评价条件的概念。评价学里的评价条件是从自包含系统概念里严格推导出来的，是自包含系统排除掉评价对象后得到的衍生系统。而实验条件组缺少严谨的定义。

[1]这里的参数和第 三章定义的统计学里有关全体的参数不同。

DoE 概念定义 6 (实验)**.** 实验是将不同的实验条件组施加给实验单元，每个实验条件组由确定以及受控的因子水平组成。这里的实验概念类似于评价学的实验，但缺少严谨的定义。

DoE 概念定义 7 (因子效应（Factor Effects）)**.** 因子效应是平均实验条件组下测量得到的响应差值（*response difference*），每个实验条件组选取不同因子水平。

因子效应的定义不同于评价学的影响定义，不是从对象之间相互影响的角度来定义的。因此，尽管英文相同，我选择了不同的中文词汇。

DoE 概念定义 8 (模型方程)**.** 模型方程是刻画响应与因子效应之间关系模型的方程。

DoE 概念定义 9 (全因子设计（Full Factorial Designs）)**.** 全因子设计通过模型方程单独估计选取的因子及其交互的所有效应。全因子设计包括 $2^k r$ 因子设计、$l^k r$ 因子设计以及通用因子设计。其中 2 或者 l 代表因子水平，k 代表实验设计选取的因子数，r 代表每个实验条件组重复测量（*replication*）的次数。通用因子设计也是选取 k 个因子，并重复测量 r 次响应，但每个因子选取的因子水平数没有限制为相同的固定值。

例如 $2^k r$ 因子设计中每个因子必须选取 2 个因子水平，$l^k r$ 因子设计中每个因子必须选取 l 个因子水平，这些因子设计中所有因子选取的因子水平数限制为同一个常数。

DoE 概念定义 10 ($l^k r$ 因子设计（Factorial Designs）)**.** $l^k r$ 因子设计（*Factorial Designs*）是一种全因子实验设计，它选取了 k 个因子，每个因子选取了 l 个因子水平，每个实验条件组重复 r 次。其模型方程能够计算出所有选取因子及其交互的所有效应。

DoE 概念定义 11 (部分因子设计（Fractional Factorial Designs）)**.** 部分因子设计通过模型方程计算选取因子及其交互的聚集效应（*Aggregate Effect*）。

典型的部分因子设计包括 $2^{k-p} r$ 因子设计。

DoE 概念定义 12 ($2^{k-p} r$ 因子设计)**.** $2^{k-p} r$ 因子设计是一种部分因子实验设计，它选取了 k 个因子，每个因子选取了 2 个因子水平，每个实验条件组重复测量 r 次。为了减少实验开销，此实验设计中 2^p 个因子或交互的效应混杂在一起。其模型方程计算得到 2^p 个因子或交互的聚集效应（*aggregated effects*）。

请注意，这里的聚集效应（aggregated effects）概念不同于第 十二章定义的总体影响（overall effect）概念。

DoE 概念定义 13 (随机误差（Random error）)**.** 随机误差表示响应变量的变化中，无法由描述响应与因子效应之间关系的模型所解释的部分。

DoE 概念定义 14 (同方差（Homoscedasticity）)**.** 同方差表示所有因子不同因子水平取值下随机误差的方差是相同的，异方差表示随机误差的方差随着因子水平取值变化而变化。

14.2　实验设计可以用来干什么？

这一节，我先从评价学的角度浅显易懂地解释实验设计 (DoE) 可以用来干什么。

有一个影响对象，也就是实验单元，它有多个可测量或测试的量，也就是 DoE 里的响应概念，同时也有多个因子，它本质上是不同对象的属性，每个属性可以取不同的值（因子水平）。DoE 里没有解释这些对象之间的关系。

DoE 方法可以认为是解决逆评价的问题，用不同的因子或者因子交互之间的效应来解释响应的变化。DoE 假定这些效应都是线性的，可以直接相加。

14.3　问题定义

实验设计是一个结构化统计方法，它用于系统地推断一个过程或者系统输出的响应和多个输入因子之间的线性模型 [307, 138, 87]。实验设计已经成为工程、制造、生物学以及行为科学等领域的基本方法，为研究人员提供了一套优化、质量改进以及假设检验的强大工具集 [155, 86, 87]。

问题定义 6 (DoE 问题定义)**.** 给定一个实验单元或实验单元总体（*population*），具有一个或多个可测量响应以及一组可控输入因子，每个因子可设定为不同水平。因子效应未知，每一项构成响应期望值的组成部分，可相加。

DoE 问题可以正式表述为："如何有策略地选择有限但具有代表性的因子水平组合进行试验，以便高效地同时估计各因子效应，从而建立一个能够描述响应变量与因子效应之间关系的有效模型，同时尽量减少未控制或未知因子的混杂 *[307, 155, 87]*。"

实验设计的输入包括因子，每个因子的水平取值、实验单元以及测量得到的响应。*DoE* 的输出是每个响应与因子效应之间的关系。实验设计的约束在于实验资源是有限的；由于成本过高或时间限制，运行因子水平的所有可能组合是不可行的。此外，观测到的响应通常会受到未控制或未知因素的干扰，这些因素可能混杂响应结果，从而掩盖因子效应。

14.4　基本假说

DoE 假说 1 (线性)**.** 响应变量与因子效应（包括因子间交互的效应）和随机误差（*random error*）之间的关系可用线性模型充分表示 *[138]*。形式上，响应 Y（因变量）是因子效应（自变量）的线性函数*:*

$$Y = \mu + \beta_1 + \beta_2 + \cdots + \beta_{i,j} + \cdots + \epsilon, \tag{14.1}$$

其中 Y 是响应，μ 是所有响应的均值，带下标的 β 表示因子效应（包括因子间交互的效应），ϵ 是随机误差（*random error*）项。单一下标的 β 表示每个因子的主效应，下标与因子相对应；多个下标的 β 表示不同因子之间的交互效应，因子由下标表示。

DoE 假说 2 (效应的可加性（Additivity of Effects）)**.** 不同因子的效应本质上可分离，且可相加 *[138, 307, 87]*。也就是说，改变某一因子所引起的响应变化，在一阶近似（*a first approximation*）下，不依赖于其他因子的水平。任何因子之间的交互都可以显式地

识别（*identified*）[2]，并建模为模型方程中的一个独立项，这就是因子效应的可分离原理（*principle of effect separation*）。

DoE 假说 3（随机化（Randomization））. 实验进行的顺序应该是随机的 *[155, 87]*。没有受到控制或未知因子可能混杂响应，随机化用于将这些潜在效应均匀地分布在所有实验条件组中。该过程骤降低了干扰因子（*nuisance factors*）混杂所研究因子集的系统效应的风险，从而保证了效应的无偏估计（*unbiased estimates*）。

DoE 假说 4（独立同分布（i.i.d.））. 模型方程中的随机误差（*random error*）项是独立同分布的（*independently and identically distributed*，*i.i.d.*）*[138, 307, 155, 87]*。具体地，这些项服从一个均值为零，方差 σ^2 为常数的正态分布：

$$\epsilon \overset{i.i.d.}{\sim} N(0, \sigma^2). \tag{14.2}$$

这意味着随机误差的独立性、同方差性以及正态性。

DoE 假说 5（同方差（homoscedasticity））. 所有因子不同因子水平下的随机误差的方差是相同的，也就是同方差（*homoscedasticity*）的属性，记作 $Var(\epsilon) = \sigma^2$ *[138, 307, 155, 87]*。同方差的属性对于标准假设检验（如 F 检验）的有效性（*validity*）和从模型中推导的置信区间的正确性至关重要。异方差（*Heteroscedasticity*）会使得这些推理无效。

14.5 基本原理

DoE 使用一个线性模型（称为模型方程），该模型对来自不同因子的效应求和，这些效应可以直接相加 [138, 307, 155, 87, 86]。对于每个从实验单元上测量得到的响应，它的效应包括响应平均值、单个因子的主效应、因子之间交互效应和随机误差，更多细节信息请参考第 14.6.2节。

如 14.6.2节详细阐述，方差分析（Analysis of Variance，ANOVA）将总体响应方差分解为组间方差、组内方差和随机误差（random error）平方和的分量。通过方差分析 [267]，可以定量计算每个因子对响应方差的贡献，并且检验因子及其交互的统计显著性。具体来说是在给定显著性水平α 下进行 F 检验，若计算的 F 值大于等于对应自由度与显著性水平下 F 分布查找的 F 值（p-value $\leq \alpha$），则说明计算 F 值分子效应相比于分母效应是统计显著的，反之则不是统计显著的。

通过系统地将因子水平组合成实验条件组，由此施加给实验单元进行实验，实验设计能够在结构化和受控条件下测量响应。总体效应分为三个分量：1）单个因子的主效应，2）因子之间的交互效应，3）偶然波动造成的随机误差 [138, 307, 155, 87, 86]。

实验设计并非一次只改变一个因子，而是引入了一个可以同时改变多个因子的框架，从而能够识别因子间的交互效应，并且构建预测模型。通常基于全因子或部分因子设计，实验设计（DoE）通过引入随机化（randomization）、重复（replication）和区组化（blocking）来减少随机误差，并增强推理的稳健性。

[2] 在多种评价方法里，可识别（identified）具有独特的含义。正如第 2.5节定义的，它是指：在给定因果结构（causal structure）假设与观察数据 分布（observed data distribution）的条件下，目标因果量（target causal quantity，即研究者希望从数据中推断的因果影响，例如干预分布 P(Y |do(X)) 或平均处理影响 ATE 等因果估计量是否可以被唯一确定。

14.6　方法学

14.6.1　典型的因子设计

因子设计包括全因子设计和部分因子设计。全因子设计通过模型方程分别估计所研究因子的全部主效应和因子之间的交互效应，而部分因子设计则计算多个因子或其交互的整体效应。尽管全因子设计需要更高的实验与计算开销，但它提供了对每个因子效应和因子间交互效应的准确估计。相比之下，部分因子设计通过估计聚集效应来降低实验开销，但这些结果可能受到混杂因素的影响，原因在于多个因子的贡献具有累积性（cumulative nature）[138, 307, 155, 87]。

- 全因子设计：全因子设计包括 $2^k r$、$l^k r$ 以及通用因子设计。在 $2^k r$ 因子设计中，每个因子选取 2 个因子水平；$l^k r$ 因子设计每个因子选取 l 个因子水平；通用因子设计允许不同因子选取可变的因子水平。全因子设计的模型方程总共计算 $(2^k - 1)$ 个效应，包括 $\binom{k}{1}$ 个因子主效应，以及 $\sum_{i=2}^{k} \binom{k}{i}$ 个因子间交互效应。每次执行重复 r 次，会引入一个额外的随机误差项。具体讨论详见 14.6.2节。$\binom{k}{i}$ 表示“从 k 个里面选取 i 个”的意思。
- 部分因子设计：部分因子设计，诸如$2^{k-p} r$ 因子设计，估计 $(2^{k-p} - 1)$ 个效应，每个效应是 2^p 个因子或交互的聚集效应。与全因子设计类似，部分因子设计也包括 r 次重复，在模型方程中也包含一个随机误差项。

14.6.2　模型方程与方差分析形式

我们以通用因子设计作为案例研究，解释其模型方程和方差分析（ANOVA）的形式。请注意，所有其他因子设计都是通用因子设计的特殊情形，拥有相同的结构化形式。

考虑一个通用因子设计，该设计用于研究影响响应的 k 个因子。其中，第 i 个因子有 n_i 个因子水平，并且每个实验重复执行 r 次。模型方程将每个测量响应分解为一个线性可加模型 [138, 307, 155, 87]，该模型包含以下部分：1）整体均值：所有实验条件组测量得到的响应平均值；

2）因子主效应：总共 $\binom{k}{1} = k$ 项，代表每个因子产生的单独效应；

3）因子间交互效应：总共 $\sum_{i=2}^{k} \binom{k}{i} = 2^k - 1 - k$ 项，包括 $\binom{k}{2}$ 个双因子交互效应、$\binom{k}{3}$ 个三因子交互效应，直到 k 因子交互效应；

4）随机误差：用于解释重复实验测量时引入的偶然波动。

方差分析 [267] 通过对所有效应进行平方求和，将响应总方差分解为不同的组成部分：因子主效应、因子间交互效应和随机误差。

为了更具体地说明，我们以 $k = 3$ 个因子为例来演示模型方程和方差分析的形式，重点介绍效应是如何被构建和分析的。其他因子设计遵循类似的表达形式。

在 $k = 3$ 个因子的通用因子设计中，因子 A 有 n_1 个因子水平，因子 B 有 n_2 个因子水平，因子 C 有 n_3 个因子水平。每个实验组合重复 r 次。此因子设计的模型方程表示为：

$$Y_{i,j,k,l} = \mu + a_i + b_j + c_k + d_{i,j} + e_{i,k} + f_{j,k} + g_{i,j,k} + \epsilon_{i,j,k,l}. \tag{14.3}$$

在这个模型方程中，$Y_{i,j,k,l}$ 表示第 l 次实验测量得到的响应结果，因子 A 选取第 i 个因子水平、因子 B 选取第 j 个因子水平、因子 C 选取第 k 个因子水平。μ 表示所有响

应的平均值。a_i 表示因子 A 的主效应，b_j 表示因子 B 的主效应，c_k 表示因子 C 的主效应。两个因子间的交互效应由 $d_{i,j}$ (因子 A × 因子 B)、$e_{i,k}$ (因子 A× 因子 C) 以及 $f_{j,k}$ (因子 B× 因子 C) 表示，而 $g_{i,j,k}$ 表示所有三个因子的交互效应。最后，$\epsilon_{i,j,k,l}$ 用于解释实验重复过程中由于偶然波动而引入的随机误差。

为了计算模型方程中的这些组件，我们首先要计算以下分组平均值。

$$\overline{Y}_{.,.,.,.} = \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r} Y_{i,j,k,l}}{n_1 \times n_2 \times n_3 \times r}, \tag{14.4}$$

$$\overline{Y}_{i,.,.,.} = \frac{\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r} Y_{i,j,k,l}}{n_2 \times n_3 \times r}, \tag{14.5}$$

$$\overline{Y}_{.,j,.,.} = \frac{\sum_{i=1}^{n_1}\sum_{k=1}^{n_3}\sum_{l=1}^{r} Y_{i,j,k,l}}{n_1 \times n_3 \times r}, \tag{14.6}$$

$$\overline{Y}_{.,.,k,.} = \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{l=1}^{r} Y_{i,j,k,l}}{n_1 \times n_2 \times r}, \tag{14.7}$$

$$\overline{Y}_{i,j,.,.} = \frac{\sum_{k=1}^{n_3}\sum_{l=1}^{r} Y_{i,j,k,l}}{n_3 \times r}, \tag{14.8}$$

$$\overline{Y}_{i,.,k,.} = \frac{\sum_{j=1}^{n_2}\sum_{l=1}^{r} Y_{i,j,k,l}}{n_2 \times r}, \tag{14.9}$$

$$\overline{Y}_{.,j,k,.} = \frac{\sum_{i=1}^{n_1}\sum_{l=1}^{r} Y_{i,j,k,l}}{n_1 \times r}, \tag{14.10}$$

$$\overline{Y}_{i,j,k,.} = \frac{\sum_{l=1}^{r} Y_{i,j,k,l}}{r}. \tag{14.11}$$

使用这些按组计算的平均值，公式 14.3中模型方程的组件可以计算为：

$$\mu = \overline{Y}_{.,.,.,.}, \tag{14.12}$$

$$a_i = \overline{Y}_{i,.,.,.} - \overline{Y}_{.,.,.,.}, \tag{14.13}$$

$$b_j = \overline{Y}_{.,j,.,.} - \overline{Y}_{.,.,.,.}, \tag{14.14}$$

$$c_k = \overline{Y}_{.,.,k,.} - \overline{Y}_{.,.,.,.}, \tag{14.15}$$

$$d_{i,j} = \overline{Y}_{i,j,.,.} - \overline{Y}_{i,.,.,.} - \overline{Y}_{.,j,.,.} + \overline{Y}_{.,.,.,.}, \tag{14.16}$$

$$e_{i,k} = \overline{Y}_{i,.,k,.} - \overline{Y}_{i,.,.,.} - \overline{Y}_{.,.,k,.} + \overline{Y}_{.,.,.,.}, \tag{14.17}$$

$$f_{j,k} = \overline{Y}_{.,j,k,.} - \overline{Y}_{.,j,.,.} - \overline{Y}_{.,.,k,.} + \overline{Y}_{.,.,.,.}, \tag{14.18}$$

$$g_{i,j,k} = \overline{Y}_{i,j,k,.} - \overline{Y}_{i,j,.,.} - \overline{Y}_{i,.,k,.} - \overline{Y}_{.,j,k,.} + \overline{Y}_{i,.,.,.} + \overline{Y}_{.,j,.,.} + \overline{Y}_{.,.,k,.} - \overline{Y}_{.,.,.,.}, \tag{14.19}$$

$$\epsilon_{i,j,k,l} = Y_{i,j,k,l} - \overline{Y}_{i,j,k,.}. \tag{14.20}$$

通过将整体均值 μ 从公式 14.3的等式右边移项到等式左边，我们得到以下公式：

$$Y_{i,j,k,l} - \mu = a_i + b_j + c_k + d_{i,j} + e_{i,k} + f_{j,k} + g_{i,j,k} + \epsilon_{i,j,k,l}. \tag{14.21}$$

针对每一项响应，将方程两边平方，然后求和，等式左边代表响应的总方差，而等式右边则分解为各效应的平方和（Sum of Squares，简称 SS）。请注意，由于正交项的性质，不同效应之间的交叉相乘项求和结果为零。因此，效应的总方差被分解为因子主效应、因子间交互效应和随机误差的贡献之和。用 SS 表示平方和，我们有以下公式：

$$SS_Y = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r} Y_{i,j,k,l}^2, \tag{14.22}$$

$$SS_0 = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r} \mu^2, \tag{14.23}$$

$$SS_T = SS_Y - SS_0, \tag{14.24}$$

$$SS_A = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r} a_i^2, \tag{14.25}$$

$$SS_B = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r} b_j^2, \tag{14.26}$$

$$SS_C = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r} c_k^2, \tag{14.27}$$

$$SS_{AB} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r} d_{i,j}^2, \tag{14.28}$$

$$SS_{AC} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r} e_{i,k}^2, \tag{14.29}$$

$$SS_{BC} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r} f_{j,k}^2, \tag{14.30}$$

SS	df
SS_Y	$n_1 \times n_2 \times n_3 \times r$
SS_0	1
SS_T	$n_1 \times n_2 \times n_3 \times r - 1$
SS_A	$n_1 - 1$
SS_B	$n_2 - 1$
SS_C	$n_3 - 1$
SS_{AB}	$(n_1 - 1) \times (n_2 - 1)$
SS_{AC}	$(n_1 - 1) \times (n_3 - 1)$
SS_{BC}	$(n_2 - 1) \times (n_3 - 1)$
SS_{ABC}	$(n_1 - 1) \times (n_2 - 1) \times (n_3 - 1)$
SS_E	$n_1 \times n_2 \times n_3 \times (r - 1)$

表 14.1: 每个平方和（SS）的自由度（df）。

$$SS_{ABC} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r} g_{i,j,k}^2, \tag{14.31}$$

$$SS_E = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r} \epsilon_{i,j,k,l}^2. \tag{14.32}$$

因此，我们有以下公式：

$$SS_Y - SS_0 = SS_T = SS_A + SS_B + SS_C + SS_{AB} + SS_{AC} + SS_{BC} + SS_{ABC} + SS_E. \tag{14.33}$$

为了量化每个效应对响应结果方差的贡献度，我们计算特定效应的平方和（SS）与响应总方差（SS_T）的比值。通过将每个效应的平方和（SS）除以其对应的自由度（degrees of freedom，简称 df），我们可以得到均方（Mean Squares，简称 MS）。各平方和对应的自由度（df）详见表 14.1。

我们按照以下公式计算效应的均方（MS）：

$$MS_A = \frac{SS_A}{n_1 - 1}, \tag{14.34}$$

$$MS_B = \frac{SS_B}{n_2 - 1}, \tag{14.35}$$

$$MS_C = \frac{SS_C}{n_3 - 1}, \tag{14.36}$$

$$MS_{AB} = \frac{SS_{AB}}{(n_1 - 1) \times (n_2 - 1)}, \tag{14.37}$$

$$MS_{AC} = \frac{SS_{AC}}{(n_1 - 1) \times (n_3 - 1)}, \tag{14.38}$$

$$MS_{BC} = \frac{SS_{BC}}{(n_2 - 1) \times (n_3 - 1)}, \tag{14.39}$$

$$MS_{ABC} = \frac{SS_{ABC}}{(n_1 - 1) \times (n_2 - 1) \times (n_3 - 1)}, \tag{14.40}$$

$$MS_E = \frac{SS_E}{n_1 \times n_2 \times n_3 \times (r - 1)}. \tag{14.41}$$

将方程 14.34 至方程 14.41 中任意两个效应对应的均方（MS）相除计算得到的比值是一个 F 统计量。通过将该比值与具有相应自由度（df）以及给定显著性水平 α 的 F 分布中的临界 F 值进行比较，F 检验可以检验这两个效应的统计显著性。F 检验的原假设 H_0 是分子效应相比于分母效应不是统计显著的，备择假设 H_1 是分子效应相比于分母效应是统计显著的。如果计算的 F 值大于等于临界 F 值，对应于 p 值小于等于显著性水平 α，此时需要拒绝原假设，选择备择假设，即在给定显著性水平 α 下分子效应相比于分母效应是统计显著的。例如，我们计算 $F_c = \frac{MS_A}{MS_E}$。在 0.05 的显著性水平下，我们参考分子自由度 $df_1 = n_1 - 1$，分母自由度 $df_2 = n_1 \times n_2 \times n_3 \times (r - 1)$ 的 F 分布，从而获得对应于 95% 累积概率的临界值 F_v。如果 $F_c \geq F_v$，则可得出结论：相对于随机误差，因子 A 的主效应在 0.05 的显著性水平下是统计显著的。

14.7　案例：推断苹果产地对苹果采购价的效应

我们接下来以苹果采购价格作为案例研究，展示如何执行实验设计框架里的方差分析以及 F 检验。

实验设计 Design of Experiment

案例研究 Case study

一箱苹果的采购价格（5kg）与其产地以及规格的组合有关。产地包括两个地区：河南省杞县（产地 1）和山东省烟台市（产地 2），规格分为三种：小果（规格 1）、中果（规格 2）和大果（规格 3）。不同产地与规格组合对应的苹果价格如下表所示。

我们将分析苹果的产地和规格对采购价格的效应，并确定产地还是规格对采购价格的效应更具有统计显著性。本案例中的因子及其水平设定将贯穿第 四部分的所有章节。

	规格 1	规格 2	规格 3
产地 1	¥25	¥45	¥50
产地 2	¥55	¥65	¥90

表 14.2: 一箱 (5kg) 苹果的采购价。

计算

本案例的模型方程如下：

$$Y_{i,j,k} = \mu + a_i + b_j + c_{i,j} + \epsilon_{i,j,k}.$$

其中 μ 是整体均值效应，a_i 是产地效应，b_j 是规格效应，$c_{i,j}$ 是产地和规格间的交互效应，$\epsilon_{i,j,k}$ 是随机误差效应。在本案例中，采购价格并不会波动，因此 $\epsilon_{i,j,k} = 0$。

我们按如下方式计算每个均值：

$$\begin{aligned}
\overline{Y}_{.,.,.} &= \frac{\sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=0}^{0} Y_{i,j,k}}{2 \times 3 \times 1}, \\
\overline{Y}_{i,.,.} &= \frac{\sum_{j=1}^{3}\sum_{k=0}^{0} Y_{i,j,k}}{3 \times 1}, \\
\overline{Y}_{.,j,.} &= \frac{\sum_{i=1}^{2}\sum_{k=0}^{0} Y_{i,j}}{2 \times 1}, \\
\overline{Y}_{i,j,.} &= Y_{i,j,k}.
\end{aligned}$$

模型方程中的每个效应为：

$$\begin{aligned}\mu &= \overline{Y}_{.,.,.,},\\ a_i &= \overline{Y}_{i,.,.,} - \overline{Y}_{.,.,.,},\\ b_j &= \overline{Y}_{.,j,.,} - \overline{Y}_{.,.,.,},\\ c_{i,j} &= \overline{Y}_{i,j,.,} - \overline{Y}_{i,.,.,} - \overline{Y}_{.,j,.,} + \overline{Y}_{.,.,.,}.\end{aligned}$$

我们接着按以下公式计算每个效应的平方和（SS）：

$$\begin{aligned}SS_Y &= \sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=0}^{0} Y_{i,j,k}^2 = 20,500,\\ SS_0 &= \sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=0}^{0} \mu^2 = 18,150,\\ SS_T &= SS_Y - SS_0 = 2,350,\\ SS_A &= \sum_{i=1}^{2}\sum_{j=1}^{3} a_i^2 = 1,350,\\ SS_B &= \sum_{i=1}^{2}\sum_{j=1}^{3} b_j^2 = 900,\\ SS_{AB} &= \sum_{i=1}^{2}\sum_{j=1}^{3} c_{i,j}^2 = 100,\\ SS_E &= \sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=0}^{0} \epsilon_{i,j,k}^2 = 0.\end{aligned}$$

因子	贡献度
产地	57.45%
规格	38.30%
产地-规格	4.25%

表 14.3: 一箱 (5kg) 苹果采购价的方差分析结果。

结论

对于这个案例，每个因子的自由度如下表所示。

因子	自由度
产地	1
规格	2
产地-规格	2

表 14.4: 每个因子的自由度。

我们接着按以下公式计算每个效应的均方（MS）：

$$MS_A = \frac{SS_A}{df_a} = \frac{1,350}{1} = 1,350,$$
$$MS_B = \frac{SS_B}{df_b} = \frac{900}{2} = 450,$$
$$MS_{AB} = \frac{SS_{AB}}{df_{ab}} = \frac{100}{2} = 50.$$

我们在 0.05 的显著性水平下比较采购一箱（5kg）苹果时，产地相对于规格的统计显著程度。计算得到的 F 值 f-compute 是：

$$f\text{-compute} = \frac{MS_A}{MS_B} = \frac{1,350}{450} = 3.$$

分子自由度为 1，分母自由度为 2 下的 F 分布在 0.05 显著性水平下的临界值为 18.51。

$$f\text{-table} = 18.51,$$
$$f\text{-compute} < f\text{-table}.$$

因此，在 0.05 的显著性水平下，我们无法得出采购一箱（5kg）苹果时，产地相比规格更具有统计显著性。

14.8　局限性

实验设计是分析响应与各因子效应之间关系的有效工具，它能够计算不同因子对响应的方差贡献程度以及其统计显著性。然而，研究人员在应用实验设计时也需注意以下局限性。

首先，实验设计要求对所有实验条件组进行详尽的实验，每个实验条件组由每个研究因子的具体水平取值组合构成。若缺少任一实验条件组的响应数据，就无法进行方差分析和 F 检验。选定因子及其水平值后，可绘制多维网格图：每个因子代表不同维度，每个实验条件组对应网格上的点。实验时需将各条件组测量得到的响应结果填入对应网格位置。若网格中任一点的响应数据缺失，方差分析和 F 检验均无法进行。

其次，随着实验设计（DoE）中选定因子数量的增加，进行实验和分析数据的相关成本会呈指数级增长 [87, 86]。此外，方差分析（ANOVA）和 F 检验需要计算不同实验条件组下响应值的平方和（SS），相较于仅计算均值的方法，其计算成本更高。因此，实验设计的成本仍然是一个需要权衡的关键因素。

最后，实验设计中进行方差分析和 F 检验时，需要检验是否满足相应的统计假设 [138]，这就需要验证这些假设是否适用于响应变量。

14.9　本章总结

实验设计采用模型方程来描述因子与响应之间的影响。其模型方程量化了：1）所有响应的总体均值；2）单个因子的主效应；3）因子的交互效应；4）由偶然波动引起的随机误差。根据因子水平对实验响应进行分组并计算相应的组均值，可以估计每个模型方程项的分量值。总响应方差可以分解为所有建模的效应（包括随机误差）的平方和，在此基础上可以进行 F 检验，确认被检验的分子效应相比分母效应是否具有统计显著性。F 检验的原假设 H_0 是分子效应相比于分母效应不是统计显著的，备择假设 H_1 是分子效应相比于分母效应是统计显著的。如果计算的 F 值大于等于临界 F 值，对应于 p 值小于等于显著性水平 α，此时需要拒绝原假设，选择备择假设，即在给定显著性水平 α 下分子效应相比于分母效应是统计显著的。

第十五章

结构方程模型

本章介绍了结构方程模型（Structural Equation Model，简称 SEM）的基本概念、问题定义、假设条件、原理、假说以及案例研究。本章主要贡献者为王磊，我从评价学的角度进行了浅显易懂地解读，并对全文进行了修改。

15.1 结构方程模型的基本概念

SEM 概念定义 1 (理论模型(Theoretical Model))**.** 理论模型是对假设关系（*hypothesized relationships*）的形式化表述，它们由测量模型和结构模型定义 *[145]*。

SEM 概念定义 2 (隐藏变量（Latent Variables），可观察变量（Observed Variables）)**.** 隐藏变量 [1] 是无法被直接测量或者测试，只能通过理论模型间接推断的变量 *[145]*。它们通常用来表示不可观察的结构体，如心理特质（例如动机、智力）或社会学概念（例如社会影响、群体规范）。而可观察变量是指其取值可以在研究中直接获得的变量 *[145]*，例如问卷调查结果、测试分数或其他任何可量化的数据。在结构方程模型（*SEM*）中，它们作为隐藏变量的观察指标，即统计层面的观察变量，同时其样本协方差矩阵（*sample covariance matrix*）为模型估计提供了实证基础。

SEM 里的隐藏变量与可观察变量等同于评价学里的隐藏变量与显式变量。

SEM 概念定义 3 (因子（Factor)，因子载荷（factor loadings)，验证性因子分析 (Confirmatory Factor Analysis，CFA))**.** 因子在统计层面上代表隐藏变量，并通过其与观察指标（*observed indicator*）之间的系统性关系（*systematic relationships*）来识别 *[145]*。因子载荷用于衡量隐藏变量对其观察指标影响强度和方向的系数。因子分析（*Factor Analysis*）旨在通过少数隐藏的共同因子来解释多个观察变量之间的相关性或协方差结构 *[144]*。验证性因子分析 “专门处理观察度量或指标（如测试题目、测试分数、行为观察评分等）与隐藏变量或因子之间的关系 *[36]*。”

SEM 概念定义 4 (内生变量（Endogenous Variables）)**.** 内生变量是 “由模型解释的变量” *[145]*。它们可以是可观察变量或隐藏变量。在结构方程模型中，内生变量通常由模型内其他变量及其干扰项共同解释。

[1] 其他中文文献里翻译为潜变量，我统一翻译为隐藏变量，更容易直观地理解

SEM 概念定义 5 (外生变量（Exogenous Variables）)**.** 外生变量是“由外部给定的变量” *[145]*。在结构方程模型中，外生变量通常作为理论模型中的输入变量，其取值不由当前模型中的其他变量来解释，而是用于解释或预测内生变量。它们可以是可观察变量或隐藏变量，在 *SEM* 中，它们可以通过路径关系用于解释或预测内生变量。

SEM 里的内生变量与外生变量和评价学里的内生变量与外生变量含义接近，区别在于 SEM 的内生和外生相对于模型，而评价学里的内生和外生相对于自包含系统。

SEM 概念定义 6 (测量模型（A Measurement Model）)**.** 测量模型是将隐藏变量与其可观察指标（*observed indicators*）相联系的模型 *[32]*。它具体规定了每个隐藏变量是如何在可操作的层面上被定义和测量。测量模型中通过引入误差项来表示测量误差。结构方程模型（*SEM*）中的测量模型本质上是一种验证性因子分析（*confirmatory factor analysis*, *CFA*），明确了观察变量如何反映其底层的隐藏变量。

测量模型只是体现了隐藏变量与观察变量（显式变量）之间的关系，这一点和 DoE 相似，和评价学定义的因果关系不同。

SEM 概念定义 7 (路径分析，路径系数（path coefficients）)**.** 路径分析是一种统计方法，它根据假设的因果结构来估计一组观察变量之间的影响。这种结构在路径图中以可视化方式呈现，其中单向箭头代表了理论设定的方向性影响 *[330]*。路径系数（*path coefficients*）表示一个变量对另一个变量直接影响的强度和方向。其中，直接路径（*direct path*）是指源变量与目标变量之间由单个单向箭头直接连接的路径，与该路径对应的系数表示模型中的直接影响；间接路径（*indirect path*）是指由两个或多个方向一致的单向箭头依次连接，并经过一个或多个其他变量的有向路径。在线性路径模型中，一条特定间接路径所对应的间接影响通常等于该路径上各路径系数的乘积 *[157]*。

SEM 概念定义 8 (结构模型)**.** 结构模型规定了变量之间（特别是隐藏变量之间）假设的因果关系 *[145]*。在此上下文下，因果关系是指方向性的影响，即理论上认为自变量的变化会引起因变量的变化。结构模型是一个更广泛的概念框架，它泛化（*generalize*）并扩展了路径分析的原理。

和评价学不同，SEM 不是定义在对象之间相互影响上的，结构模型并不能表达对象之间的因果影响，而是量之间的方向性的影响。

SEM 概念定义 9 (残差)**.** 在结构模型中无法解释的干扰项（*disturbance terms*）称为残差 *[157]*。

SEM 概念定义 10 (参数 (parameters))**.** 参数是线性方程求解的未知系数，主要包括因子载荷、路径系数、方差（*variances*）以及协方差（*covariances*）。方差主要包括外生变量、残差以及测量误差的方差，而协方差主要指外生变量之间的关联，通常会标准化为相关系数以便解释 *[157]*。

15.2　结构方程模型可以用来干什么？

这一节，我先从评价学的角度浅显易懂地解释结构方程模型（SEM）可以用来干什么。

存在一组对象可观察、可测量或者可测试的变量（可观察变量），基于这些数据出发，SEM 用两种不同的方法同时建模。

首先，它建立线性的测量模型，对外生的或者内生的隐藏变量与可观察变量之间的关系进行线性建模。SEM 将隐藏变量称为因子，将衡量隐藏变量对其观察指标影响强度和方向的系数称为因子载荷。所谓的内生是指受模型内其他变量的影响，而外生则不受模型内其他变量的影响，是模型的外部输入。

其次，它建立线性结构模型，对内生的或者外生的隐藏变量之间的影响进行方向性的建模。SEM 将一个变量对另一个变量直接影响的强度和方向称为路径系数。SEM 没有严谨地定义影响。

最后基于一系列的假设，SEM 去估计这些方程组的参数。最后通过和实际观察的数据进行统计学的验证来检验模型的正确性。

在实际操作中，SEM 通常把可观察变量视为完全识别的单指标隐藏变量。这是一种技巧性的操作：假设存在一个方程组，可以唯一地建立单个可观察变量和单个隐藏变量之间的关系。这样能够用统一的结构模型来建模。

15.3　问题定义

结构方程模型（SEM） [145, 315, 129, 10, 157] 是一种强有力的统计方法，能够帮助研究者分析变量之间的复杂关系。SEM 广泛应用于需要多变量分析的研究领域，尤其是在行为科学中，用于刻画心理、社会以及经济等构造体（constructs）。SEM 是在一组线性结构方程中估计未知参数的通用方法 [145]。

问题定义 7 (SEM 问题定义)**.** *SEM 的核心问题可以表述为：如何基于可观察数据，构建并验证一个包含多个抽象概念（隐藏变量）及其复杂因果关系的理论模型 [157]。SEM 通过估计参数来量化影响的方向和强度，从而使理论模型具有可操作性，并揭示外生变量如何影响内生变量。*

具体而言，它同时解决了两个问题：如何准确测量抽象概念？以及“这些概念之间的因果网络关系是什么？”

该问题还可以进一步表述为一个详细的分析问题。分析以两个主要输入为起点：一是实证数据，即对可观察变量（如问卷回答或测试分数）的测量；二是理论模型的设定。理论模型设定利用 SEM 的基本概念阐明了假设关系：它定义了隐藏变量（不可观察的构造体）并将其与可观察变量（指标）联系起来，同时也假定了外生变量（作为外部输入或原因，可以是可观察变量或隐藏变量）与内生变量（其变化由模型解释的结果变量）之间的因果网络。该模型受五个关键约束条件的限制（详见第 15.4 节）。SEM 的核心任务是估计这组线性结构方程中的未知系数，并检验所设定的模型对观察数据的解释程度 [157]。

对抽象概念进行测量是 SEM 的一种表达，根据本书的评价学理论和计量学理论，抽象概念不能测量，只可以推理。

15.4　基本假说

SEM 假说 1 (测量误差的独立性)**.** 与观察指标相关的误差项其期望值为零，彼此之间互不相关，且与它们所要测量（根据评价学理论，应为推测或者推理的）的隐藏变量也不相关（*uncorrelated*）*[157]*。

SEM 假说 2 (多元正态分布)**.** 观察变量服从多元正态分布 *[157]*。

SEM 假说 3 (线性)**.** *SEM* 通常假定其各组成部分之间存在线性关系，包括隐藏变量与其观察变量之间的线性关系，以及隐藏变量彼此之间的线性关系 *[157]*。

SEM 假说 4 (结构残差 (Structural Residuals) 与外生变量的独立性)**.** 结构模型的干扰项，即结构方程中的残差项，其期望值为零，并且与模型中的外生隐藏变量不相关 *[157]*。

SEM 假说 5 (模型可识别性)**.** 模型必须是可识别的（*identified*），即存在一组唯一的参数取值，使其与观察变量总体协方差矩阵相一致 *[157]*。

15.5　基本原理

结构方程模型（Structural Equation Modeling，SEM）是一种统计分析框架，它将验证性因子分析（*Confirmatory Factor Analysis*，*CFA*）与路径分析（*Path Analysis*）相结合，通过比较模型所预测的协方差结构（通过协方差矩阵表示）与实证数据中实际观察到的协方差结构之间的拟合程度，来检验理论模型。SEM 允许同时估计测量模型和结构模型部分，从而对隐藏变量与观察变量之间的关系提供一个整体而系统的刻画 [157]。

测量模型（*Measurement Model*）规定了隐藏变量如何通过观察指标加以测量。CFA 用于验证这样一个假设：一组观察变量能够反映某一隐藏的概念 [157]。

结构模型（*Structural Model*）则用于说明隐藏变量之间的因果关系。结构模型以类似于联立回归方程系统（system of simultaneous regression equations）的方式刻画由直接路径表征的直接影响（direct effect）和由间接路径表征的间接影响（indirect effect），从而允许变量之间存在复杂的相互依赖关系 [157]。

SEM 的一个关键优势在于其能够显式地处理测量误差，这不同于传统回归模型通常假定自变量不存在测量误差的做法。SEM 同时为隐藏变量和观察变量引入误差项，从而获得更加准确、无偏的参数估计 [157]。

15.6　方法学

SEM 方法学包含一个流程，始于模型的正式设定（formal specification ），继而进行参数估计，最终以模型检验结束。

结构方程模型（SEM）通常由一组线性方程表示，这些方程一般分为两个部分：

测量模型。 测量模型 (Measurement Model) 定义了隐藏变量如何通过可观察的指标来体现 [157]。形式上，外生和内生测量的关系由以下公式给出：

$$\mathbf{x} = \Lambda_x \boldsymbol{\xi} + \delta, \tag{15.1}$$

$$\mathbf{y} = \Lambda_y \boldsymbol{\eta} + \epsilon, \tag{15.2}$$

其中，$\mathbf{x}$ 表示与外生隐藏变量 $\boldsymbol{\xi}$ 相关的观察指标，$\mathbf{y}$ 表示与内生隐藏变量 $\boldsymbol{\eta}$ 相关的观察指标。矩阵 Λ_x 和 Λ_y 表示相应的因子载荷，而 δ 和 ϵ 分别表示 $\mathbf{x}$ 和 $\mathbf{y}$ 的测量误差。

结构模型　结构模型（Structural Model）部分描述了隐藏变量之间的因果关系 [157]。具体而言，内生隐藏变量由其他内生因子和外生输入共同决定：

$$\boldsymbol{\eta} = B\boldsymbol{\eta} + \Gamma\boldsymbol{\xi} + \zeta, \tag{15.3}$$

其中，$\boldsymbol{\eta}$ 表示内生隐藏变量的向量，$\boldsymbol{\xi}$ 表示外生隐藏变量向量。矩阵 B 捕捉了内生变量之间的回归关系，而 Γ 则表示外生变量对内生变量的影响。项 ζ 代表模型无法解释的残差（干扰项）。在实际建模中，通过将测量模型中的测量误差固定为 0，并将因子载荷固定为 1，可以将观察变量视为完全识别的单指标隐藏变量，从而将其纳入上述结构模型的框架中。

这些方程构成了结构方程模型（SEM）的基础，能够同时对隐藏变量的测量结构以及它们之间的结构（因果或关联）关系进行建模 [157]。

结构方程模型（SEM）中参数估计的主要方法是 *最大似然（ML）估计*，该方法假设观察变量服从 *多元正态分布*。最大似然估计通过最小化观察样本协方差矩阵与假设模型所隐含的协方差矩阵之间的差异来估计模型参数——如因子载荷、路径系数、方差和协方差。在多元正态性的假设下，最大似然估计提供了有效且无偏的参数估计 [157]。

最大似然（ML）估计方法还生成了关键的 *拟合优度指标*，帮助检验模型与观察数据结构的拟合程度。估计模型后，研究人员通过一组 *拟合优度指标* 来检验模型的整体拟合度。虽然 *卡方（χ^2）统计量* 用于检验观察协方差矩阵与模型隐含的协方差矩阵之间的精确拟合程度，但它对样本量极为敏感，因此需要谨慎解读。在实际研究中，研究人员依赖于 *绝对拟合指标*（例如 *RMSEA*、*SRMR*）和 *增量拟合指标*（例如 *CFI*）来检验模型的适配性 [157]。

常用的拟合标准包括：

- *CFI*（*Comparative Fit Index,* 比较拟合指数） *[157]*: 大于 0.90 的值通常被视为模型拟合的可接受指标，而超过 0.95 的值则表示良好的拟合。
- *RMSEA*（*Root Mean Square Error of Approximation,* 均方根误差近似度） *[157]*: 小于 0.08 的值通常表示可接受的拟合，而小于 0.05 的值则表示较好或精确的拟合。
- *SRMR*（*Standardized Root Mean Square Residual,* 标准化均方根残差） *[157]*: 小于 0.08 的值通常被解读为可接受的拟合。

这些指标应结合解读，同时考虑模型的复杂性、理论预期和样本特征。

15.7　示例：推断苹果产地对购买价格的影响

本例使用 SEM 方法探讨苹果产地如何通过关键质量特征影响购买价格。我们将产地建模为隐藏构造体，而苹果特性和价格则为直接观察变量。为符合经典的结构模型定义，苹果特性和价格在模型中被处理为完全识别的单指标隐藏变量。这种方法在保持方法论严谨性的同时，也兼顾了实际的可解释性。

结构方程模型

变量说明

该模型结合了基于农业研究中的隐藏变量和观测变量：

- 产地（OR）：隐藏变量，表示区域种植优势，通过三个观察指标测量：
  - $OM1$：气候适宜性（1-5 级，5 = 最适合苹果生长）
  - $OM2$：土壤有机质含量（g/kg，15-25 g/kg = 最适宜范围）
  - $OM3$：年平均温度（°C，8-12 °C = 有利于糖分积累的理想温度）
- 甜度（SW）：观察变量（在结构模型中视为单指标隐藏变量），以布里克度（Brix）为单位（连续变量）：
  - 苹果甜度的典型范围：10-16 Brix
- 大小（SI）：观察变量（在结构模型中视为单指标隐藏变量）（连续变量）：
  - 苹果大小的典型范围：1.2-2.2 百克
- 保质期（SL）：观察变量（在结构模型中视为单指标隐藏变量），以天为单位（连续变量）：
  - 苹果常温下存储的典型范围：3-14 天
- 购买价格（PP）：观察变量（在结构模型中视为单指标隐藏变量），以 ¥/kg 为单位：
  - 直接观察到的苹果市场交易价格

请注意：在路径模型中，SW、SI、SL 和 PP 作为内生变量（受其他变量影响），而 OR 作为外生隐藏变量。

理论框架与假设

基于农业经济学研究，我们假设优越的产地条件直接提高苹果的质量特性，进而提高市场价值。提出的因果路径如下：

$$OR \to SW, \quad OR \to SI, \quad OR \to SL, \quad OR \to PP$$
$$SW \to PP, \quad SI \to PP, \quad SL \to PP$$

理论依据：

- 更好的产地条件（最佳气候、土壤、温度）促进糖分积累（↑ 甜度）、果实发育（↑ 大小）和结构完整性（↑ 保质期）
- 因消费者偏好较甜、较大、且保质期较长的苹果，更高的质量特性而在市场上获得价格溢价
- 产地还可能通过区域品牌和声誉影响对价格产生直接影响

模型说明

1. 测量模型（仅针对外生隐藏变量 *产地*）：

$$
\begin{aligned}
OM1 &= \lambda_{OM1} OR + \delta_{OM1}, \quad (\lambda_{OM1} = 1，\text{标记变量用于识别}) \\
OM2 &= \lambda_{OM2} OR + \delta_{OM2} \\
OM3 &= \lambda_{OM3} OR + \delta_{OM3}
\end{aligned}
$$

其中，λ = 因子载荷，δ = 测量误差。

2. 结构模型（所有变量之间的路径分析）：

$$SW = \beta_1 OR + \zeta_1 \qquad (\text{产地} \rightarrow \text{甜度}) \tag{15.4}$$

$$SI = \beta_2 OR + \zeta_2 \qquad (\text{产地} \rightarrow \text{大小}) \tag{15.5}$$

$$SL = \beta_3 OR + \zeta_3 \qquad (\text{产地} \rightarrow \text{保质期}) \tag{15.6}$$

$$PP = \gamma_1 OR + \gamma_2 SI + \gamma_3 SW + \gamma_4 SL + \zeta_4 \qquad (\text{各变量对价格的影响}) \tag{15.7}$$

其中，β、γ = 路径系数，ζ = 结构干扰项。对于显式变量 SW、SI、SL 和 PP，通过将其测量误差固定为 0，因子载荷固定为 1，使其在逻辑上等同于被完美测量的隐藏变量。

参数估计与模型识别

估计结果（基于农业研究的假设但合理的参数）：

- **测量模型参数：**

$$\lambda_{OM1} = 1.00 \quad \text{（用于识别，固定）}$$
$$\lambda_{OM2} = 0.75, \quad \lambda_{OM3} = 0.68 \quad \text{（自由估计）}$$
$$\text{Var}(\delta_{OM1}) = 0.20, \quad \text{Var}(\delta_{OM2}) = 0.25, \quad \text{Var}(\delta_{OM3}) = 0.30$$

- **结构模型参数：**

$$\beta_1 = 0.55, \quad \beta_2 = 0.48, \quad \beta_3 = 0.62$$
$$\gamma_1 = 0.25, \quad \gamma_2 = 0.18, \quad \gamma_3 = 0.35, \quad \gamma_4 = 0.20$$

所有参数在 $p < 0.05$ 时均具有统计显著性，支持假设关系。

数值预测示例

本例展示了如何使用模型进行预测。根据估计的参数，价格的结构方程为：

$$\hat{PP} = 0.25 \times OR + 0.18 \times SI + 0.35 \times SW + 0.20 \times SL + \zeta_4$$

对于具有特定质量特征的苹果：

$$SW = 14 \text{ Brix}, \quad SI = 1.67 \text{ 100g}, \quad SL = 7 \text{ 天}$$

对于一个产地条件高于平均水平的地区（该地区的产地条件评分 OR = 3.8，计算方法是将该地区的 OM1、OM2 和 OM3 值加权组合得到，权重根据测量模型参数确定），从观察因子得出的预测价格分量为：

$$\begin{aligned}\hat{PP} &= 0.25 \times 3.8 + 0.18 \times 1.67 + 0.35 \times 14 + 0.20 \times 7 \\ &= 0.95 + 0.30 + 4.90 + 1.40 = 7.55 \text{ ¥/kg}\end{aligned}$$

影响分解分析

我们将苹果产地对购买价格的总影响进行分解。

1. 间接影响：

$$\begin{aligned}\text{通过甜度：}&\quad \beta_1 \times \gamma_3 = 0.55 \times 0.35 = 0.1925\\ \text{通过大小：}&\quad \beta_2 \times \gamma_2 = 0.48 \times 0.18 = 0.0864\\ \text{通过保质期：}&\quad \beta_3 \times \gamma_4 = 0.62 \times 0.20 = 0.1240\\ \text{总间接影响：}&\quad 0.1925 + 0.0864 + 0.1240 = 0.4029\end{aligned}$$

2. 直接影响：

$$\text{直接影响} = \gamma_1 = 0.25$$

3. 总影响：

$$\text{总影响} = \text{直接影响} + \text{间接影响} = 0.25 + 0.4029 = 0.6529$$

基于上述分析可知产地优势（OR）对价格（PP）的总影响为 0.6529。这意味着，当 OR 增加一个单位时，PP 平均增加 0.65 ¥/kg。直接影响为 0.25，而间接影响为 0.4029，间接影响通过甜度 (0.1925)、大小 (0.0864) 和保质期 (0.1240) 产生中介作用。直接影响和间接影响分别占比为：直接影响占比 38.3% 和间接影响占比 61.7%。图 15.1 为其图形化表示。

模型拟合与验证

拟合优度指标（假设结果）：

- $\chi^2/df = 2.15$ （可接受：< 3.0）
- CFI $= 0.94$ （良好拟合：> 0.90）
- RMSEA $= 0.06$ （良好拟合：< 0.08）
- SRMR $= 0.05$ （良好拟合：< 0.08）

本示例采用结构方程模型方法，探讨苹果产地如何通过关键品质特征影响采购价格。在该模型中，产地被设定为一个隐藏构造体，而苹果属性与价格则被定义为观察变量。为与经典结构模型的定义保持一致，苹果属性和价格均被视为完全识别的单指标隐藏变量。这种方法在方法论严谨性与实际可解释性之间取得了平衡。

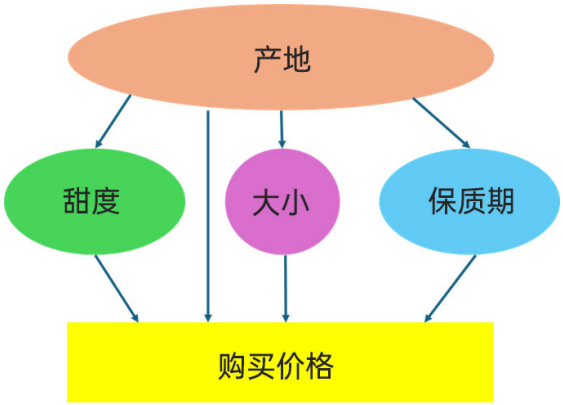


图 15.1: 苹果定价模型的结构方程路径图

总而言之，这个例子展示了结构方程模型（SEM）在农业价格分析中的应用——通过标准化指标（10-16 Brix 甜度、1.20-2.20 百克大小、3-14 天保质期），模型的结果对生产者和营销者更具可操作性。影响分解结果明确表明，苹果的产地主要通过提高甜度来影响购买价格。这一见解可以帮助各地区集中精力研发提高糖分聚集的技术，从而提升苹果的市场价值。

15.8　局限性

SEM 是一个强大的工具，也存在一些局限性，研究人员必须加以考虑 [145, 315, 129, 10, 157]。它通常需要满足许多假设，包括线性、正态性和误差的独立性。它还通常要求较大的样本量，尤其是对于复杂模型，小样本可能导致不可信的分析结果。模型设定错误（model misspecification）也可能会产生误导性结论。必须强调模型设计要扎根于理论基础和实证证据的重要性。此外，SEM 可能在计算上较为密集，尤其是当引入非线性关系或类别变量时，需要功能强大的软件和计算资源。最后，尽管 SEM 可以识别变量之间的关联，但它本质上并不等同于确立因果关系，它并没有可操作性地定义什么是因果。例如在结构因果模型里，影响是用 do 算子的形式定义 [224]；在对象因果模型里，影响是以因果对象存在与否，对影响对象产生的影响的差异来定义的。因此，在做出因果推断时必须极其谨慎。

15.9　本章小结

结构方程模型是一种用于分析多元数据中复杂关系的强大工具。它使研究人员能够对隐藏变量进行建模、处理测量误差，并研究不同领域（如心理学、社会学、教育学和市场营销）中的关系结构。结构方程模型对于理解多个变量如何同时交互具有不可替代的作用，因此在行为科学领域具有极高的应用价值。然而，在使用时必须密切关注模型设定（model specification）、数据质量和样本量，以避免诸如过度拟合或模型设定错误等常见陷阱。

第十六章

随机对照实验

本章介绍随机对照实验（Randomized Controlled Trials，简称 RCTs）的基本概念、问题定义、假设、原理、方法学以及案例研究。王晨曦是本章的主要贡献者。我从评价学的角度浅显易懂地解释了 RCTs 可以用来干什么，同时修改了全文。

16.1 随机对照实验的基本概念

本节定义的基本概念来自于前人编写的经典著作 [295, 224, 86, 87]。

RCTs、PO 和准实验概念定义 1 (干预 (intervention or treatment))**.** 干预是指明确定义的过程，这些过程的影响（*effect*）或安全性（*safety*）正在被研究。这里的明确定义仍然适用评价学里定义的明确定义的概念。干预可以是一种药理学制剂（药物）、医疗器械、外科技术、行为疗法、教育项目，或任何其他施用于受试者（*participants*）以引发可测量响应的操作。

干预通常用二元指示符 T 表示。对每个单元 i，$T_i \in \mathcal{T}$ 为单元 i 所接受的干预；在二元情形下，$\mathcal{T} = \{0, 1\}$，其中 $T_i = 1$ 表示干预处理，$T_i = 0$ 表示对照处理 *[253]*。

干预是评价学理论中对象的一种特例。需要指出的是 RCTs 并没有区分对象、系统和组件，也没有区分对象不同的性质。

RCTs 和 PO 概念定义 1 (受试者（participant）或单元（unit）)**.** 受试者或单元是指符合预设入选准则（*criteria*），并正式加入研究的个体，它可以指“特定时间点上的物理对象，例如一块土地、一个个体，或在时间维度上没有变化的一个人” *[253]*。将按照试验方案接受调查程序、数据收集和随访。

RCTs 和潜在结果理论 (PO) 里的受试者或者单元是评价学中的影响对象的一个特例，它属于一个显式对象。

RCTs 概念定义 1 (结果（outcome）)**.** 结果是从受试者那里通过测量获得的数值指标，它不是分类变量。

RCTs 的结果是评价学中影响对象的量。

RCTs 概念定义 2 (干预组)**.** 干预组是随机分配接受干预的受试者队列（*cohort*），该干预正是研究的主要关注点。

RCTs 的干预组是评价学里一组影响对象。

RCTs 概念定义 3 (对照组（control group）)**.** 对照组是指被随机分配接受对照措施（*comparator*）的受试者队列，该对照措施用于评估治疗效果。这种对照可以是安慰剂（一种惰性物质）、标准治疗方案、不同治疗方法，或不接受任何治疗，具体取决于研究问题和伦理考量。

RCTs 的对照组也是评价学里一组影响对象。

RCTs 概念定义 4 (随机化（randomization）)**.** 随机化意味着每位受试者都有同等的机会被分配到干预组或对照组。

RCTs 概念定义 5 (研究者（investigator）)**.** 研究者是指在试验现场负责实施随机对照实验（*RCTs*）的合格人员。主要研究者（*principal investigator, PI*）对研究的伦理合规性和按照试验方案（*protocol*）执行负有全面责任，包括管理研究团队，以及保护受试者的权利和数据的完整性。

RCTs 概念定义 6 (对照比较（controlled comparison）)**.** 对照比较 是随机对照实验（*RCTs*）中一项基本的方法学原则，指系统性比较干预组和对照组的结果。其核心目标是通过研究设计使所有其他无关变量（*extraneous variables*）保持不变，从而隔离出干预的因果影响。

RCTs 概念定义 7 (盲法（blinding）)**.** 盲法是随机对照实验（*RCTs*）中的一种方法学程序，指研究中一个或多个相关方对治疗分配情况保持未知（即不知道某位受试者属于治疗组还是对照组）。其主要目的是防止因意识或无意识的偏倚（*bias*）影响感知到的结果、行为或对结果的解释。

RCTs 概念定义 8 (平均干预影响（average treatment effect，简称 ATE）)**.** 平均干预影响是干预组和对照组之间期望的影响差异。对于任何受试者 i 都存在两种潜在结果：如果分配到干预组，则观察到的结果为 $Y_i(1)$；如果分配到对照组，则观察到的结果为 $Y_i(0)$。然而，同一受试者 i 的两种潜在结果不能同时被观察到。随机对照实验利用随机化来计算平均干预影响的无偏估计量（*an unbiased estimator*）。

16.2　随机对照实验可以用来干什么？

这一节，我先从评价学的角度浅显易懂地解释 RCTs 可以用来干什么。

随机对照实验是从评价对象和影响对象的角度出发，通过实验设计构造控制组来消除混杂对象的影响。

RCTs 是通过构造干预组和对照组来推导一个具体的干预（评价对象）对一组受试者（影响对象）的影响，不要求保证这些受试者完全相同。在干预组里，待评价的干预施加给一组受试者；而在对照组里，不同的评价对象施加给另外一组受试者。同一名受试者要么属于干预组，要么属于对照组。RCTs 通过比较干预组和对照组上观察到的差异来推断干预的影响。

RCTs 假设受试者的数目足够多的时候，通过随机分配受试者到干预组和对照组，就可以近似保证两组中评价条件在不同维度上分布的一致性，类似评价学中等价评价条件。

RCTs 的想法是通过随机化地分配受试者到干预组或者对照组，近似地保证两组的评价条件的配置空间接近相同。因此，就可以推断两个对象在等价的评价条件配置空间下的影响差异，这些影响实际上是个分布。

RCTs 无法推断一个评价对象对另外一个具体影响对象的影响。当影响对象的数目足够多的时候，它能获得两个不同的评价对象对这些影响对象集合的平均影响的差异。

16.3 问题定义

随机对照实验是推测干预和结果之间关系的一个系统性方法 [295, 224]。

问题定义 8 (RCTs 问题定义)**.** 给定受试者总体和提出的干预作为输入，每个受试者 i 都有两种潜在结果 *[295, 86]*：接受干预时的 $Y_i(1)$ 和接受对照干预时的 $Y_i(0)$。其挑战在于，对于同一受试者 i，只能观察到其中一种结果。请注意，在评价学（*Evaluatology*）方法里，在某些情况下，我们确实可以观察到不同结果，正如我们在第 十二 章中所讨论的。

随机对照实验问题可以表述为“如何使用随机机制（随机化）将受试者分配到干预组或对照组，从而提供平均干预影响 *(ATE) [295]* 的无偏估计。”

随机对照实验的输入是受试者、干预组、对照组以及每位受试者观察到的结果。在随机将受试者者分配到干预组和对照组的约束下，随机对照实验输出平均干预影响 *ATE*。

16.4 基本假说

RCTs 和 PO 假说 1 (稳定单元干预值假设 （stable unit treatment value assumption，SUTVA） [66, 253])**.** 任何受试者或者单元（例如，个体患者）的潜在结果不受其他受试者或者单元的特定干预分配的影响。此假设包含两个关键组成部分：

1. 无干扰性：一个单元的干预分配不影响任何其他单元的潜在结果（*poential outcome*）。这确保了单元的独立性，它是标准统计推断的先决条件。

2. 无隐藏的干预变化：对于每个被分配到干预的单元，其干预方案都是相同的。不存在可能对结果产生不同影响的多种、未明确说明的干预或对照干预。

SUTVA 是支撑随机对照实验（RCTs）和第 十七章介绍的潜在结果理论（PO）的最基本假设，因为它允许为每个单元在每种干预状态下定义唯一的潜在结果。

首先，“无干扰”条件要求分配给某一单元的干预不会改变其他单元的潜在结果。例如，在评价在线学习平台中的个性化推荐系统时，如果我们假设推荐给学生 A 的课程不会影响学生 B 的学习结果，则该假设的这一部分成立。然而，在实际应用中，该假设可能被违反。假设平台中某些课程名额有限，如果学生 A 被推荐并选修了一门热门课程，学生 B 可能因此无法选修该课程，从而其学习结果受到间接影响。在此类情境下，单元之间的干扰影响需要被显式建模，以避免因果估计产生偏差。

其次，“不存在隐藏处理版本”的假设要求每一个干预水平都必须具有唯一且一致的定义。例如，若将干预定义为“参加某项培训项目”，则必须保证该培训在所有参与者之间相同。如果部分学员仅参加为期三天的短期培训，而另一些学员则接受为期三个月的强化课程，那么“培训”这一标签实际上对应了多种不同的干预。在这种情况下，潜在

结果无法唯一对应到单一干预水平上，违反了 SUTVA。解决该问题的一种常见策略是细化干预定义，例如区分“短期培训”和“长期培训”，使每一个潜在结果都清晰对应于一个明确定义的干预。

RCTs 假说 1 (可忽略性，ignorability)**.** 此假设也称为 无混杂性（*unconfoundedness*）或 可互换性，*exchangeability*。它指出，在随机化分配机制下，干预分配与潜在结果是条件独立的 *[295]*。形式上，

$$(Y(1), Y(0)) \perp\!\!\!\perp T, \tag{16.1}$$

其中 $Y(1)$ 和 $Y(0)$ 分别是干预组和对照组下的潜在结果，T 是干预分配指示符。

随机化是一种物理设计特征，它 确保 可忽略性假设成立。即使存在未观察到的因素，它使得干预组和对照组在统计上具有等价性，或称为在期望（expectation）意义上可 互换（exchangeable）。因此，观察到的结果中的任何系统性差异都可以因果地归因于干预。

RCTs 假说 2 (一致性 consistency)**.** 受试者在接受特定干预水平时观察到的结果，正是该受试者在相同干预水平下的潜在结果 *[295]*。形式上，如果 $T_i = t$，那么 $Y_i^{obs} = Y_i(t)$。

这个假设意味着干预在所有受试者上的应用或测量中没有歧义。它将概念上的潜在结果与实际观察到的数据联系起来。

16.5　基本原理

随机对照实验（RCTs）的主要动机在于通过确保干预组和对照组在基线（baseline）时具有统计等价性，从而消除混杂偏倚（confounding bias）。这种等价性通过将受试者随机分配到不同实验条件来实现，从而保证观察到的和未观察到的协变量在各组之间平均而言是平衡的 [295, 224, 86]。

随机对照实验通常涉及将一种或多种干预与对照干预进行比较，并经常结合盲法和安慰剂对照，以最大限度地减少预期和测量偏倚。该方法的特点是具有明确定义的协议（protocol）、干预分配与潜在结果独立，以及严格遵循统计学原则（如意向性治疗分析，intention-to-treat [1]） [295, 224, 86]。

随机对照实验（RCTs）广泛应用于临床医学、社会科学和公共政策评估领域，使研究者能够在受控且可重复的条件下，对干预措施的效果做出高置信度的推断 [295]。

随机化是随机对照实验的核心。它确保每位受试者都有同等的机会被分配到干预组或对照组。这一点至关重要，因为它确保研究在开始时，各组在各个方面平均来说是相似的，这种平均来说的相似性不仅仅体现在年龄或性别，还包括无数我们甚至无法测量的其他因子（如遗传、饮食或自然恢复能力）。随机对照实验（RCTs）通过随机化控制了处理组/对照组之外的其他因素（包括已知和未知因素），将组间差异的影响减少至随机波动。

由于干预组和对照组具有可比性，研究结束时任何结果的显著差异都可以更可信地归因于干预本身，而不是各组之间预先存在的差异。

[1] 意向性治疗分析（Intention-to-Treat Analysis, ITT）是统计学和临床研究中一种关键的分析方法，其核心原则是： 所有被随机分配到各组的受试者，无论是否实际接受分配的治疗、是否完成试验或出现方案偏离，都应按照最初随机分组的组别进行数据分析 。

16.6　方法学

16.6.1　平均干预影响的形式化定义

我们将正式定义随机对照实验中的对照比较（controlled comparison）。目标是估计平均干预影响 (average treatment effect，ATE)，即干预在整个研究总体中的平均因果影响 [295]。

我们首先定义随机对照实验中的关键步骤

- 令 $i = 1, 2, ..., N$ 表示试验中的全部 N 名受试者。
- 受试者 i 的干预分配表示为 T_i:

$$T_i = \begin{cases} 1, & \text{如果受试者 } i \text{ 被分配到干预组,} \\ 0, & \text{如果受试者 } i \text{ 被分配到对照组.} \end{cases} \tag{16.2}$$

- 如果 $T_i = t$，对受试者 i 来说观察结果（比如说苹果采购价格的下降）表示为 $Y_i(t)$。

根据上述讨论，受试者 i 观察结果取决于干预分配:

$$Y_i^{obs} = Y_i(T_i) = T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0). \tag{16.3}$$

随机化的根本优势在于它允许进行简单而有力的比较。平均干预影响（ATE）通过计算两组平均结果的差异来估计:

$$ATE = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)], \tag{16.4}$$

其中:

- $E[Y_i(1)]$ 是干预组中所有受试者结果的期望值（算术平均值）。
- $E[Y_i(0)]$ 是对照组中所有受试者结果的期望值。

在实际使用中，我们计算样本算术平均值来估计 ATE:

$$\widehat{ATE} = \frac{1}{N_1} \sum_{i:T_i=1} Y_i(1) - \frac{1}{N_0} \sum_{i:T_i=0} Y_i(0). \tag{16.5}$$

其中:

- N_1 是分配到干预组中的受试者数目。
- N_0 是分配到对照组中的受试者数目。
- $N = N_0 + N_1$

16.7　案例：推测苹果产地对苹果采购价的影响

我们将进行一项案例研究，执行随机对照实验以调查苹果产地对采购价格的影响。

随机对照实验

案例研究

- 假设您的研究中有 10,000 颗（$N = 10,000$）苹果种子。
- 您使用计算机随机分配 5,000 颗（$N_1 = 5,000$）苹果种子种植在河南省杞县（干预组）。
- 另外 5,000 颗（$N_0 = 5,000$）种子种植在山东省烟台市（对照组）。

十年后，您测量了两组种植苹果的平均每公斤采购价格。
假设获得的平均每公斤采购价格如下表所示。

试验分组	OR_i	$E[PP_i(OR_i)]$
干预组	1	¥8/kg
对照组	0	¥14/kg

表 16.1: 干预组与对照组苹果的平均每公斤采购价格。

计算

将随机对照实验的公式代入这个案例：

- 干预组苹果的平均每公斤采购价格 ($E[PP_i(1)]$) 为 ¥8/kg。
- 对照组苹果的平均每公斤采购价格 ($E[PP_i(0)]$) 为 ¥14/kg。
- 代入公式 16.5: $\widehat{ATE} = 8 - 14 = -6$。

结论

这表明在河南省杞县种植的苹果的平均采购价格比在山东省烟台市种植的苹果低 ¥6/kg。由于随机化，我们可以确信这种差异很可能归因于产地本身，而不是其他混淆因素。

16.8　局限性

随机对照实验被广泛认为是建立因果推断的黄金标准，但它也存在一些研究人员必须意识到的重要局限性。

首先，RCTs 无法推断一个评价对象对另外一个具体影响对象的影响。当影响对象的数目足够多的时候，它能获得两个不同的评价对象对这些影响对象集合的平均影响的差异。

第二，泛化能力（generalizability）（外部有效性）问题是一个重要关切。随机对照实验高度受控的条件和特定的受试者总体可能无法代表真实世界的环境或更广泛的患者总体 [295]。因此，随机对照实验的研究结果虽然在内部有效，但可能无法直接转化为常规

临床实践。

第三，随机对照实验（RCTs）的规范实施通常伴随着高昂的成本和复杂的后勤工作，这往往构成重大障碍。实施随机化、确保对协议（protocol）的依从性，并在足够长的随访期内维持盲法 [295]，需要大量的资金投入、时间成本和操作资源，使得许多研究问题难以开展。

第四，伦理上和实践上的限制 [295] 经常限制了随机对照实验的适用性。将受试者随机分配到已知或普遍认为有害的干预在伦理上是不允许的。此外，对于长期结果或罕见疾病的研究，在有限的时间框架内招募和保留足够的样本量可能在实践中不可行。

最后，随机对照实验（RCTs）如果设计或执行不够严谨，容易受到特定的方法学偏倚影响 [224]。例如，当不同组别的失访率（即退出研究的参与者比例）存在系统性差异时，可能引发失访偏倚（attrition bias）；而盲法实施失败则可能引入实施偏倚和检测偏倚（performance and detection bias）。此外，分析过程必须遵循意向性治疗原则（intention-to-treat principle），即将所有受试者保留在其最初被分配的组别中进行分析，无论其实际接受了何种治疗，以维护随机化的完整性。这一要求使实际干预影响的解释变得复杂。

16.9 总结

本章介绍了随机对照实验（RCTs）的基本概念、原理、方法以及局限性，并提供了一个案例研究。

第十七章

潜在结果理论

本章系统介绍了潜在结果（Potential Outcome，PO）理论的基本概念、问题陈述、核心假设、基本原理、方法体系以及一个示例性案例研究。本章主要贡献者为高婉铃。我从评价学的角度浅显易懂地解释了 PO 可以用来干什么，同时修改了全文。

17.1 潜在结果理论基本概念

PO 概念定义 1（分配机制（Assignment Mechanism））. 分配机制是指："每个单元被分到干预组或对照组的规则或概率结构 *[253]*。在潜在结果框架下，这一规则可泛化地写为 $\Pr(W \mid X, Y(0), Y(1))$，即用协变量与潜在结果来刻画分组如何进行，其中 W 表示干预分配，X 表示协变量，$Y(0)$ 与 $Y(1)$ 分别表示各单元在对照组与干预组下的潜在结果" *[253]*。

PO 用一个泛化的概率表达式，同时描述随机实验中的随机分配，以及观察性研究中的非随机分配。

第 十六章介绍的随机对照实验是 $\Pr(W \mid X, Y(0), Y(1))$ 的特殊情形：分配与潜在结果相互独立，因而具有更明确的因果识别基础。

PO 概念定义 2（协变量（Covariate））. 协变量是指 "在分配干预之前已取值，或不可能受到干预影响的变量，例如人的性别" *[253]*。协变量也常被称为干预前变量（*pretreatment variables*）*[249]*。

PO 概念定义 3（干预后变量（Posttreatment Variable））. 干预后变量是指 "在干预分配之后可能记录的、用于描述实验单元 *(unit)* 的变量" *[249]*。这类变量可能直接受到干预本身的影响，或受到其他干预后因素（*factor*）的影响。

例如，在一项农业实验中，若干预为施用某种特定肥料，则在收获时测量的叶绿素含量或果实大小均属于干预后变量，因为它们是在施肥之后记录的 [249]。

由于缺少评价学里的 EO, AO, COs 等概念，PO 里的协变量和干预后变量定义具有模糊性。协变量相当于评价学里的评价条件概念，但后者定义在自包含系统的概念上，具有严谨的定义。干预后变量类似于 AO 的可测量或者可测试变量。

PO 概念定义 4（中介变量（Mediator））. 中介变量是一类特殊的干预后变量，其不仅发生在干预之后，而且在因果路径中传递了部分干预对结果的因果影响 *[252]*。

对每个单元 i，设 $t, t' \in \mathcal{T}$ 表示干预 T 的可能取值，例如 1 表示干预组，0 表示对照组。定义 $M_i(t)$ 为单元 i 在干预 t 下中介变量 M 的潜在结果。进一步定义潜在结果 $Y_i(t, M_i(t'))$ 为：在干预为 t 的反事实情形下，若中介变量的取值与 $M_i(t')$ 相对应，则结果 Y 将会取的值。这里 t 与 t' 可以相同，也可以不同。当 $t' = t$ 时，简记为 $Y_i(t, M_i(t))$，表示在干预 t 下、且中介变量也按其在该干预下本会出现的取值时的结果 Y 的潜在结果。在二元干预情形下，设 $\mathcal{T} = \{0, 1\}$，其中 0 表示对照组、1 表示干预组。此时，自然直接影响（*Natural Direct Effect, NDE*）与自然间接影响（*Natural Indirect Effect, NIE*）定义为 *[223, 252]*：

$$NDE = \mathbb{E}[Y_i(1, M_i(0)) - Y_i(0, M_i(0))], \tag{17.1}$$

$$NIE = \mathbb{E}[Y_i(1, M_i(1)) - Y_i(1, M_i(0))]. \tag{17.2}$$

因此，中介变量刻画了一条因果路径（causal pathway），通过该路径，干预在超出其自身直接影响之外进一步影响结果 [252]。PO 的中介变量概念借鉴了 SCM 中的因果概念。但 PO 本质上是一种统计学方法。

PO 概念定义 5（潜在结果（PO））. 潜在结果是指在某一特定时间点、单元在采取某一特定行为（即干预或对照）后将会观察到的结果。具体而言，$Y(1)$ 表示在主动干预下的潜在结果，$Y(0)$ 表示在对照干预下的潜在结果。取决于实际接受的干预，现实中每个单元只有一个结果能够被实际观察到 *[253]*。

在评价学理论里，如果一个系统可以完全被复制，即明确定义，所有的结果均可以观察，在这种场景下，不存在潜在结果。

PO 概念定义 6（“反事实”（A Counterfactual））. “反事实”指的是一个单元在另一种干预条件下，其无法被观察到的潜在结果，这个干预条件不同于实际接受的干预条件。由于每个单元只能经历一种干预状态，因此反事实结果始终无法被直接观察，它构成了因果推断的核心挑战——即估计在另一种干预场景下本可能发生的情况 *[224, 133]*。

PO 概念定义 7（因果影响（Causal Effects））. 因果影响被定义为“在同一组单元上，不同干预条件下潜在结果的比较” *[253]*。

17.2　潜在结果理论可以用来干什么？

这一节，我先从评价学的角度浅显易懂地解释 PO 可以用来干什么。

PO 是从干预分配机制的角度上对 RCTs 进行了泛化。如果第 i 个单元的干预分配 T_i 与潜在结果 $(Y_i(1), Y_i(0))$ 相互独立，这就是一个随机实验的情况，PO 等同于 RCTs。而在观察实验情况下，通常干预分配机制受到协变量，也就是干预前变量的影响。

PO 从另一个角度出发，确定同时影响潜在结果和分配机制的干预前协变量集合，并对它们进行控制。若这些协变量足以控制混杂，则在给定这些协变量的条件下，可以认为干预分配与潜在结果相互独立，从而构造具有统计可比性的干预组和对照组，并估计平均干预影响。

17.3　问题陈述

本节首先阐述 PO 框架下因果推断的基本问题，随后给出其形式化的问题陈述。

17.3.1　因果推断的基本问题

该框架的一个关键含义在于：因果推断本质上是一个数据缺失问题 [133]。对于每一个单元，至多只能观测到一个潜在结果，而另一个潜在结果始终处于反事实状态 [224, 133]。这构成了因果推断的核心挑战——在始终存在不能观测的一个结果的情况下，如何估计干预的影响。

从更根本的角度看，在任一给定时间点，单元不可能同时既接受干预又不接受干预，也不可能同时暴露于多个不同的干预下 [224, 133]。现实中只能观察到一条已实现的路径，而所有其他潜在结果均为不可观察的反事实 [224, 133]。

尽管同一单元在不同时间点上可以接受不同干预，但单元在不同时刻的状态并不可简单互换（interchangeable）。后续时间点的结果不仅可能受到当前干预的影响，还可能受到先前干预或单元自身演化特征（evolving characteristics）的影响 [224, 133]。例如，若患者在时间 t 服用某种药物，则在时间 $t+1$ 观察到的影响，可能不仅反映药物的直接因果作用，还包含既往的暴露、累积的生物反应，或基础疾病进展的影响。这些随时间变化的因素构成混杂因素，使得仅将观察差异归因于当前干预变得更加困难 [224, 133]。

17.3.2　问题的形式化表述

问题定义 9 (PO 问题定义)**.** 给定一组观察或实验单元的总体，每个单元 i 可以接受多个干预条件之一 $(T_i \in 0, 1)$，记 $Y_i(1)$ 与 $Y_i(0)$ 分别表示单元 i 在干预与对照条件下的潜在结果。对于每个单元，仅有一个潜在结果能够被观察到，其具体取决于实际分配的干预，另一个潜在结果则不可观察 *[253]*。

未知量正是每个单元的不能观察的潜在结果，这使得无法直接观测单元层面的干预影响 $(Y_i(1) - Y_i(0))$。这一限制定义了潜在结果框架的核心问题：不可能同时观察同一单元在不同干预条件下的结果 *[253]*。

PO 问题可表述为："如何基于可观察数据，包含干预分配、协变量与已实现的结果，推断或估计干预对结果的因果影响。"

该问题以一个核心输入为起点，即可观察数据，其中包括每个单元的已实现结果以及它的干预分配和干预前协变量。

该模型受制于第 *17.4*节中详述的一组已被广泛接受的约束条件 *[241, 245, 241]*，这些约束规范了潜在结果、分配机制与观察数据之间的关系。给定这些要素，潜在结果框架下因果推断的核心任务是：仅利用观测数据以及 *PO* 框架的假设，识别并估计因果估计量（*causal estimands*）*[253]*——例如平均干预影响（*average treatment effect*, *ATE*）或相关的总体层面对比指标 *[133]*。该框架由此提供了一种系统化的方法，明确因果影响中哪些部分是可识别的，并厘清观察数据在多大程度上能够支持对未观察潜在结果的推断。

输入　观察数据 $\mathcal{D}$ 定义为：

$$\mathcal{D} = \{(Y_i^{\text{obs}}, T_i, X_i)\}_{i=1}^{n}, \tag{17.3}$$

其中 Y_i^{obs} 表示每个单元 i 的已实现结果（*realized outcome*），$T_i \in 0,1$ 表示对应的干预分配，X_i 表示干预前协变量。

输出（因果估计量） 目标量是用潜在结果联合分布定义的因果影响，包括：

$$ATE \ \ (Average\ Treatment\ Effect) : \mathbb{E}[Y_i(1) - Y_i(0)], \tag{17.4}$$

$$ATT \ \ (Average\ Treatment\ Effect\ of\ the\ Treated) : \mathbb{E}[Y_i(1) - Y_i(0) \mid T_i = 1], \tag{17.5}$$

以及诸如条件平均干预影响（*Conditional Average Treatment Effect*，*CATE*）等协变量条件下的影响（*covariate-conditional effects*）：

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x]. \tag{17.6}$$

约束（假设） *PO* 框架受到三项关键约束条件的限制，详见第 *17.4* 节。

17.4　基本假设

为了使因果影响能识别和估计，潜在结果（PO）框架依赖于三个基础性假设：稳定单元干预值假说已经在第 十六章假说 1阐述。其他两个假说如下：

PO 假说 1 (可忽略性（Ignorability） / 无混杂性（Unconfoundedness） [241, 245])**.** 在给定观察到的协变量 X 的条件下，干预分配独立于潜在结果 $(Y(0), Y(1))$：

$$(Y(0), Y(1)) \perp\!\!\!\perp T \mid X. \tag{17.7}$$

该假设是从观察数据中建立因果关系的基石。“|” 表示 “在……条件下” 的数学符号，而 “$A \mid B$” 表示 “在 B 条件下的 A”。符号 “⫫” 表示 “相互独立”，其中 “$A \perp\!\!\!\perp B$” 表示 A 与 B 相互独立。

上述公式表明，一旦我们控制了所有同时影响干预分配与潜在结果的干预前协变量 X，决定一个单元 i 接受干预（$T_i = 1$）还是对照（$T_i = 0$）的机制，可以看作 “近似随机” 的，且和两种不同情形下的潜在结果相互独立。因此，在控制 X 后比较两组结果差异，便可将其归因于干预本身，而非由 X 所代表的系统性差异。需要强调的是，该假设要求研究者尽可能识别并控制那些同时影响干预分配与潜在结果的干预前协变量，且其成立依赖于研究设计、领域知识等对协变量选择的支撑。

例如，在一个在线学习平台中评价新推出的高级互动模块相对于标准模块对期末考试成绩（Y）的影响，而学生可自行选择是否参加高级模块。可忽略性假设要求我们能够测量所有影响学生选课决定和最终成绩的协变量 X，例如既往 GPA、平台使用时长、历史测验成绩等。若未控制 X，直接比较参加高级模块与标准模块的考试结果，可能混入选择效应：既往 GPA 更高的学生更倾向于选择高级模块。若满足可忽略性假设，则在这些协变量 X 完全相同的学生之间，是否选择高级模块在统计意义上等同于随机分配，从而允许公平的比较。

PO 假说 2（正值性（Positivity） / 重叠性（Overlap） [241]）. 对于协变量 X 的所有可能取值，接受干预的条件概率严格介于 *0* 与 *1* 之间：

$$0 < P(T = 1 \mid X) < 1. \tag{17.8}$$

该假设保证：在任意协变量取值下，单元既有可能接受干预，也有可能不接受干预，从而使干预组与对照组在协变量空间中均有单元支持干预影响的估计。在在线学习平台的例子中，该假设要求对每一种学生特征组合（例如不同学习基础水平），都存在既有可能选修新模块、也有可能不选修新模块的正概率。如果平台基于某些准则（例如禁止低成绩学生参与）限制访问模块，则在受影响的子群体中该假设被违反，由于缺乏反事实数据，这些亚群（strata）中的因果推断将无法进行。

这些假设提供了将观察数据或实验数据与因果影响联系起来的理论支架。

17.5　基本原理

鲁宾将因果推断问题优雅地重构为：通过精心设计（design）、合理建模（modeling），或二者结合，系统性地恢复或逼近缺失的反事实结果 [133]。

在鲁宾的形式化框架中，研究中的每一个单元 i 都对应一对潜在结果 $(Y_i(1), Y_i(0))$，其中 $Y_i(1)$ 表示单元接受干预时的结果，$Y_i(0)$ 表示未接受干预时的结果。观察到的结果可表示为：

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0), \tag{17.9}$$

其中 $T_i \in 0, 1$ 为处理分配指示符。

该表达 (representation) 显式 (explicit) 地表明，对每个单元而言，两种潜在结果中只有一个能够被实际观察。个体层面的因果影响定义为个体干预影响（*Individual Treatment Effect*）：

$$\tau_i = Y_i(1) - Y_i(0), \tag{17.10}$$

而总体层面的影响，例如平均干预影响（*Average Treatment Effect, ATE*） [133]，定义为：

$$\tau = \mathbb{E}[Y(1) - Y(0)]. \tag{17.11}$$

由于个体影响通常不可识别，大多数分析聚焦于总体平均影响（如 ATE）或子群体层面的影响（如条件平均干预影响，CATE） [133]。

17.6　方法学

在 PO 框架下，因果影响的估计源于一个根本性问题：反事实结果的缺失——每个单元 i 只能观察到 $Y_i(1)$ 或 $Y_i(0)$ 中的一个 [133]。因此，方法论的核心在于构建有效的比较 (valid comparison)，以逼近缺失潜在结果的分布。

用于估计因果影响的方法 [249, 255, 245, 287] 大体可分为两类情形：随机实验与观察性研究。

本章节在潜在结果（PO）框架下关注干预对结果变量的整体因果影响 (total causal effect)，即平均干预影响 $E[Y(1)-Y(0)]$，这里的整体因果影响不同于我在第十二章定义的总体影响（overall effect）和第 十四章定义的聚集效应（aggregate effect）。在该框架中，干预也可能通过干预后的中介变量（mediator）影响结果，从而在概念上允许将整体影响分解为自然直接影响（NDE）与自然间接影响（NIE） [223, 252]。然而，由于中介分析涉及跨世界反事实（cross-world counterfactuals），通常依赖强于可忽略性条件的额外假设 [210]。鉴于此，本章节的方法学重点在于整体因果影响（total causal effect）的定义与估计；基于中介变量的分解（Mediation-based decompositions）仅作为潜在扩展，而不构成核心估计目标。

17.6.1　随机实验

在随机实验（Randomized Experiments） [295, 224, 133] 中，干预分配 T_i 与潜在结果 $(Y_i(1), Y_i(0))$ 相互独立，通过构造（construction），该设计确保了可忽略性条件成立：

$$(Y(1), Y(0)) \perp\!\!\!\perp T. \tag{17.12}$$

在此情形下，平均干预影响（ATE） [133] 的无偏估计可直接通过样本均值实现：

$$\hat{\tau}_{ATE} = \bar{Y}_{T=1} - \bar{Y}_{T=0}, \tag{17.13}$$

其中 $\bar{Y}_{T=1}$ 与 $\bar{Y}_{T=0}$ 分别为干预组与对照组的观察结果均值。随机化确保了样本均值差一致地估计 $E[Y(1) - Y(0)]$，无需额外调整。其抽样方差可采用独立样本的标准方法估计，允许通过置信区间和假设检验推测。

17.6.2　观察研究

当随机分配不可行时，干预分配可能依赖于干预前协变量 X [241, 249]。在这种设置下，无偏估计需要对 X 进行控制（conditioning on X），以恢复干预与潜在结果之间的独立性：

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid X. \tag{17.14}$$

对具有相同或相似协变量值的干预组与对照组单元进行比较，从而推进估计过程。

若干经典方法均基于这一思想：

(a) 协变量调整

协变量调整（covariate adjustment）方法 [133, 241] 通过对结果变量建立显式的统计模型，在给定干预状态与协变量的条件下直接估计潜在结果的条件期望。其核心思想并非在样本层面构造可比单元，而是在模型层面刻画结果生成机制。

具体而言，设定如下结果模型：

$$\mathbb{E}[Y \mid T, X] = \alpha + \tau T + f(X), \tag{17.15}$$

其中 $f(X)$ 用于捕捉协变量对结果的系统性影响，α 表示截距项，与 $f(X)$ 共同决定 $T=0$ 时的条件期望。在可忽略性假设

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid X$$

以及模型设定正确的前提下，模型可用于预测每个单元在两种干预状态下的潜在结果 $\mathbb{E}[Y(1) \mid X]$ 与 $\mathbb{E}[Y(0) \mid X]$，从而通过对预测差异取平均得到平均干预影响。

因此，协变量调整方法通过显式建模结果变量来补全缺失的反事实，其有效性依赖于结果模型的正确或近似正确设定。

(b) 匹配估计

匹配估计（matching estimation） [250, 247, 133] 通过在样本层面构造协变量分布相似的干预组与对照组单元，从而在数据层面近似随机实验所产生的可比性。该方法不对结果变量建立显式模型，而是通过样本重构来逼近反事实结果。

具体而言，对于每一个接受干预的单元 i，在对照组中寻找一个或多个在协变量 X 上与其相似的单元 $j(i)$。在可忽略性假设成立的前提下，可将匹配对照单元的观察结果 $Y_{j(i)}$ 视为干预单元在未接受干预条件下潜在结果 $Y_i(0)$ 的近似，从而构造单元层面的反事实比较。

平均干预影响可由匹配样本中结果差的平均值估计：

$$\hat{\tau}_{\text{match}} = \frac{1}{N_T} \sum_{i \in T=1} \left(Y_i - Y_{j(i)} \right), \tag{17.16}$$

其中 N_T 为干预组样本量。

匹配估计的核心特征在于，其主要作用发生在研究设计阶段：通过构造协变量平衡的比较样本来减少混杂偏倚，而因果效应的估计本身通常采用简单的均值差或低维回归完成。

(c) 倾向得分方法

在观察性研究中，干预分配通常并非随机，而是受到干预前协变量 X 的影响。若进一步满足可忽略性假设

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid X,$$

则原则上可以通过控制 X 来恢复组间可比性，从而估计平均干预影响 $\mathbb{E}[Y(1) - Y(0)]$。但在实际应用中，X 往往维数较高，或既包含连续变量也包含离散变量，直接对高维 X 进行精确匹配、分层或回归调整都较为困难。倾向得分方法正是为缓解这一问题而提出的。鲁宾和罗森鲍姆 [241] 提出了倾向得分（*propensity score*）的概念，定义为在给定协变量条件下单元接受干预的概率：

$$e(X) = P(T = 1 \mid X). \tag{17.17}$$

直观地说，倾向得分表示：在协变量取值为 X 时，一个单元进入干预组的概率。它把“是否接受干预”这一分配机制，压缩为一个一维数值，从而用较低维度的信息概括 X 对分

配的影响。倾向得分的关键性质在于：若在给定 X 后，干预分配与潜在结果相互独立，则在给定倾向得分 $e(X)$ 后，这一独立性仍然成立 [241]：

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid e(X). \tag{17.18}$$

这意味着，与其直接在高维空间中对 X 本身进行控制，不如对 $e(X)$ 进行控制。只要两个单元具有相同或相近的倾向得分，就可以认为它们在干预分配意义上是“可比的”；在此基础上再比较其结果差异，便更接近随机实验中的组间比较。

据此，平均干预影响可通过多种方式估计，包括：

- 分层（*stratification*） / 子分类（*subclassification*） *[241]*：先根据估计得到的倾向得分 $e(X)$ 将样本划分为若干层，例如按十分位或固定区间分组；然后在每一层内分别计算干预组与对照组结果的均值差，再对分层结果进行加权汇总。其基本思想是：在倾向得分相近的单元之间进行比较，从而近似实现“在相同 X 下随机分配”的效果。

- 加权（*weighting*） *[241, 239]*：对每个单元赋予与倾向得分相关的权重，使加权后的干预组与对照组在协变量分布上更为接近。最常见的做法是逆概率加权（inverse probability weighting，IPW）[239, 126]：干预组单元权重约为 $1/e(X)$，对照组单元权重约为 $1/(1-e(X))$。直观地说，对那些“本不太容易接受干预、但实际接受了干预”的单元赋予更高权重，对那些“本较容易接受干预、但实际未接受干预”的单元也赋予更高权重，从而修正非随机分配带来的失衡。

- 匹配（*matching*） *[241]*：在估计得到的倾向得分基础上，为每个干预组单元在对照组中寻找一个或多个倾向得分最接近的匹配对象。匹配完成后，仅在成功配对的样本上比较结果差异。其目标是人为构造出一组在协变量意义上更接近随机分配的“可比样本”。

需要强调的是，倾向得分本身并非一个独立的因果估计量，而是一种用于实现协变量平衡的工具。它并不直接给出 $\mathbb{E}[Y(1)-Y(0)]$，而是先把“谁与谁在分配意义上可比”这一问题处理好，再与匹配、加权、分层或回归等估计策略结合，用于识别和估计因果影响。在倾向得分模型正确设定且可忽略性、正值性等假设成立的前提下，上述方法均可用于估计平均干预影响；其中，分层与匹配侧重构造可比子样本，加权则通过对全体样本重新赋权来实现组间平衡。

17.6.3　推断与不确定性估计

因果影响估计的不确定性，至少包括两个来源：一是由随机分配或随机抽样带来的波动；二是在观察性研究中，由于每个单元只能观测到一个潜在结果，需要通过匹配、加权、回归或插补等方法估计反事实结果，从而引入的额外不确定性 [224, 133]。标准误差通常用于刻画第一种波动，其经典计算常建立在单元相互独立等假设之上。唐纳德 · 鲁宾提出，可采用重复抽样逻辑（repeated sampling logic）[251]，即基于随机化或抽样机制构造估计量的重复抽样分布，以量化这种不确定性；也可采用贝叶斯方法 [249, 253]，将潜在结果与参数视为随机变量，由因果影响的后验分布直接给出不确定性的概率化表述。

17.6.4　删失与主分层

在许多实证研究中，尤其是在生物医学、社会科学和经济学研究中，因果推断常常受到这样一个问题的影响：对于部分单元，某些干预条件下的结果变量 Y 根本无法定义，而不仅仅是“未被观察到”[254]。

这种情况通常被称为删失（censoring）或截断（truncation）。其典型原因是：在干预之后、结果测量之前，出现了某些中间事件（intermediate events），例如死亡、失访、无应答或其他吸收性事件（absorbing events）[254]。一旦发生这些事件，研究者便无法继续测量原本感兴趣的结果变量。例如，若结果变量定义为“治疗后一年的生活质量”，而某单元在某种干预下会在一年内死亡，则该干预下的潜在结果 Y 对该单元而言并不存在。此类情形违反了潜在结果框架中的标准设定，即默认对每个单元 i，$Y_i(1)$ 与 $Y_i(0)$ 都是明确定义的 [254]。

面对这一问题，若仍直接讨论总体平均干预影响 $\mathbb{E}[Y(1)-Y(0)]$，可能会遇到概念上的困难：因为至少对某些单元而言，$Y(1)$ 或 $Y(0)$ 本身未定义。换言之，问题不只是估计困难，而是因果量本身可能不再适用于全体总体。为应对这一困难，鲁宾提出了主分层（principal stratification）框架 [254] 。其核心思想是：不再先对全体人群定义单一的 Y 影响，而是先根据“某单元在两种干预下，中间事件将如何发生”将其归入不同亚群，再在这些亚群内讨论因果影响。

具体地，设 $S_i(1)$ 与 $S_i(0)$ 分别表示单元 i 在干预组和对照组条件下某一中间变量的潜在结果，则有序对 $(S_i(1), S_i(0))$ 定义了该单元所属的主分层（principal stratum）。它刻画的是：该单元在两种干预下，中间事件各自是否会发生。例如，当中间变量为生存状态时，可得到如下分层 [254, 90]：

$$\begin{aligned}
&\text{始终存活者（Always-survivors）：} && S_i(1)=1,\, S_i(0)=1,\\
&\text{受保护者（Protected）：} && S_i(1)=1,\, S_i(0)=0,\\
&\text{受损者（Harmed）：} && S_i(1)=0,\, S_i(0)=1,\\
&\text{始终未存活者（Never-survivors）：} && S_i(1)=0,\, S_i(0)=0.
\end{aligned} \tag{17.19}$$

其中，“始终存活者”指无论接受干预还是对照都会存活；“受保护者”指只有接受干预才会存活；“受损者”指只有不接受干预才会存活；“始终未存活者”指无论接受干预还是对照都无法存活 [254, 90]。

一般而言，只有当 $Y_i(1)$ 与 $Y_i(0)$ 在某个主分层内同时具有定义时，该分层内的结果因果影响才是明确定义的。在生存示例中，这对应于始终存活者分层，因为只有在干预和对照下都存活时，生活质量等干预后结果才有意义。此时，可在该分层内定义分层特定因果效应（stratum-specific causal effect） [254, 90]：

$$\tau_{AS} = \mathbb{E}[Y_i(1) - Y_i(0) \mid S_i(1) = S_i(0) = 1]. \tag{17.20}$$

其含义是：在那些无论接受干预还是对照都会存活的单元中，干预对结果 Y 的平均影响。

需要强调的是，主分层成员资格 $(S_i(1), S_i(0))$ 永远无法被完全观察，因为每个单元实际上只经历一种干预状态。我们最多只能知道 $S_i(1)$ 或 $S_i(0)$ 中的一个，而无法同时知道两者。因此，对主分层内因果影响的估计通常需要额外假设、联合建模或辅助信息；若存在可预测主分层归属的协变量，也可结合贝叶斯或似然方法进行推断。实践中，还常通过敏感性分析考察结论对不同未观察分层设定是否稳健 [254, 90]。

该框架中的两个关键考量包括：

- 分层特定的估计量（*Stratum-specific estimands*） *[254, 90]:* 当总体中部分单元的干预后结果未定义时，总体平均干预影响（ATE）可能不再具有明确意义。此时，研究者应首先明确真正感兴趣的主分层，例如始终存活者。
- 影响分离（*Separating effects*） *[254, 90, 252]:* 干预既可能通过改变中间事件（如生存或删失）影响结果的可定义性，也可能在结果可定义时直接影响 Y。主分层在概念上有助于分离这两类影响，尽管其识别通常需要较强假设。

总而言之，由干预后中间事件所导致的删失，会对标准潜在结果框架下的因果推断构成挑战，因为它使某些潜在结果本身不再定义。主分层方法提供了一条连贯路径：先依据中间变量的联合潜在结果划分亚群，再在 Y 同时有定义的亚群内定义和估计因果影响。它在保留潜在结果逻辑的同时，明确承认：当存在干预后截断时，可识别性受到限制，因果问题必须被更精确地表述。

17.6.5　小结

总体而言，潜在结果（PO）方法在明确定义的潜在结果之间比较来定义因果影响，并在随机实验和非随机研究设计下提供了一个连贯一致的（coherent）估计框架。通过控制(conditioning)、匹配或加权等方法，该框架在显式假说下重构缺失的反事实结果，从而得到具有可解释性且符合统计原理的因果影响估计。其核心优势在于概念上的清晰性：它严格区分因果影响的定义与因果影响的估计，确保所有实证分析均建立在透明的反事实逻辑之上。

通过将因果推断表述为一个受明确假说约束的数据缺失问题，PO 框架构建了一个统一的统计范式。随机对照实验（RCTs）在设计上天然满足可忽略性假设，而观察研究则需要通过匹配、回归、倾向得分或工具变量等调整策略（工具变量见第 二十 章）来近似随机化。该框架的这种多功能性（versatility）解释了为何潜在结果（PO）框架已成为现代因果推断的主流语言，它既在概念推理上提供了清晰性，也在统计估计上实现了严谨性。

17.7　示例：推断苹果产地对采购价格的因果影响

本示例在 Neyman（1923）与 Rubin（1974）提出的潜在结果框架下，对苹果采购价格问题进行重新表述。研究目标是估计苹果产地（OR）对采购价格（PP）的因果影响，并进一步分析该影响是否通过关键中介变量 – 甜度（SW）、果径大小（SI）和保质期（SL）– 直接或者间接产生。

单元，干预及潜在结果

每一批苹果或一次运输批次均视为一个独立的单元，记为 $i \in \{1, 2, \ldots, N\}$。
干预变量 OR_i 表示苹果的产地状态，表示苹果是否来自有利的种植区域（例如具有更优的气候或土壤条件）。为简化起见，假设干预变量为二元：

$$OR_i \in \{0, 1\}, \quad 0 = \text{较不利产地}, \quad 1 = \text{有利产地}.$$

对于每一个单元 i，采购价格的潜在结果定义为：

$$PP_i(OR_i) = \text{在干预 } OR_i \text{ 下可观察到的采购价格}$$

然而，在实际数据中，仅有一个潜在结果可以被观测到：

$$PP_i^{obs} = PP_i(OR_i^{obs}),$$

而反事实结果 $PP_i(1 - OR_i^{obs})$ 不能观察。

中介变量与结果

要捕捉苹果产地影响购买价格结果的机制，我们为每个单元定义了一组中介变量：

$$M_i = (SW_i, SI_i, SL_i),$$

分别对应苹果的甜度、果径大小和保质期。
每一个中介变量本身也受到干预变量的影响，因此均具有其各自的潜在结果：

$$SW_i(OR_i), \quad SI_i(OR_i), \quad SL_i(OR_i).$$

因此，采购价格的潜在结果可进一步表示为：

$$PP_i(OR_i, M_i(OR_i)),$$

表明，苹果产地既可能直接（经由 OR_i），也可能通过中介变量的潜在结果（经由 $M_i(OR_i)$），对采购价格产生影响。

因果估计量

在潜在结果框架下，苹果产地对采购价格的平均干预影响（*Average Treatment Effect, ATE*）定义为：

$$ATE = \mathbb{E}[PP_i(1) - PP_i(0)],$$

该量表示从较不利产地转变为有利产地时，采购价格的期望因果影响。
为将该总体影响分解为直接影响与间接影响，定义自然直接影响（*Natural Direct Effect, NDE*）与自然间接影响（*Natural Indirect Effect, NIE*）：

$$NDE = \mathbb{E}[PP_i(1, M_i(0)) - PP_i(0, M_i(0))],$$

$$NIE = \mathbb{E}[PP_i(1, M_i(1)) - PP_i(1, M_i(0))].$$

其中：

- 自然直接影响度量干预改变时价格如何变化，但中介变量被固定为较不利产地下的反事实取值；

- 自然间接影响度量仅由干预导致的中介变化引起的价格变化，干预水平保持不变。

根据定义，有：

$$ATE = NDE + NIE.$$

协变量与分配机制

记 X_i 为干预前协变量向量（如生产规模、采摘周期、市场环境等），这些变量可能混杂干预与结果之间的关系。干预分配机制 $P(OR_i = 1 \mid X_i)$ 描述了在给定协变量条件下干预分配的概率。在标准假设下，对 X_i 进行适当调整可以保证因果估计量的可识别性。

识别所需的假设

对 ATE、NDE 与 NIE 的识别依赖于以下潜在结果假设：

- 稳定单元干预值假设 *(SUTVA):* 每个单元的潜在结果仅依赖于其自身的干预状态与中介变量，不存在单元间干扰，且不存在多版本干预：

$$PP_i(OR_i, M_i(OR_i)) \text{ 与} M_i(OR_i) \text{ 对所有单元} i \text{ 均明确定义.}$$

- 可忽略性*:* 在给定协变量 X_i 的条件下，干预分配与潜在结果及中介变量相互独立：

$$\{PP_i(1), PP_i(0), M_i(1), M_i(0)\} \perp\!\!\!\perp OR_i \mid X_i.$$

- 正值性*:* 在任意协变量取值下，各干预水平均以正概率出现：

$$0 < P(OR_i = 1 \mid X_i) < 1.$$

数值示例 Numerical Illustration

假设一个基于农业数据校准的的二元处理情景：

$$\mathbb{E}[PP_i(1)] = 6.8 \text{ ¥/kg}, \quad \mathbb{E}[PP_i(0)] = 6.1 \text{ ¥/kg}.$$

则有：

$$ATE = 6.8 - 6.1 = 0.7 \text{ ¥/kg}.$$

假设进一步分解得到：

$$NDE = 0.25 \text{ 元/千克}, \quad NIE = 0.45 \text{ 元/千克}.$$

由此可见，约 64% 的总影响通过甜度、果径大小与保质期的改善实现，而其余 36% 体现为产地所带来的直接市场优势。

结果解释 Interpretation

潜在结果框架明确区分了潜在结果与观察结果，从而支持对反事实进行形式化推理，例如：若某一批苹果产自不同地区，其采购价格本应为何值。
上述因果分解表明：

- 直接影响刻画了有利产地所带来的区域或品牌层面的优势；
- 间接影响反映了更优种植条件所诱导的品质提升，包括更高甜度、更大果径以及更长保质期。

总体而言，即便在市场条件完全一致的情况下，转向有利产地仍可使苹果的平均采购价格提高约 0.7 元/千克，其主要来源为中介变量所驱动的品质影响。

17.8　局限性

尽管潜在结果框架为因果影响的定义与估计提供了严谨基础，但仍需正视其若干局限。

首先，该框架依赖一系列关键的可识别性假设，如可忽略性、正值性以及稳定单元处理值假设（SUTVA），而这些假设在实际研究中通常无法直接检验。若存在隐藏的混杂或单元间干扰，因果估计可能产生偏差。

其次，基于 PO 的分析通常是数据密集的，对数据质量和样本规模要求较高，反事实量的有效估计往往需要大量且平衡的数据，尤其当干预或中介变量是连续或高维时。

再次，虽然该框架在概念上区分了干预、结果与协变量，但其本身并不能确定正确的因果结构；若协变量或中介变量设定不当，可能导致因果影响估计失真。间接（中介）影响的估计还引入了额外的假设，诸如序列可忽略性（sequential ignorability）等更强的假设 [132]，更难实证地验证。

最后，潜在结果框架虽然为因果推理提供了精确的语言，但在缺乏强假设或实验控制的情况下，无法单独实现因果识别。因此，因果推断在潜在结果（PO）框架下，始终需要以实质性理论、领域知识和敏感性分析作为支撑。

17.9　本章总结

潜在结果框架是定义与估计因果影响的基础性方法。它通过比较不同干预条件下本应观察到的结果，将因果推断形式化为一个数据缺失问题。潜在结果方法之所以尤为强大，是因为它将因果推理与明确的假设联系起来，例如一致性、SUTVA（稳定单元处理值假设）和可忽略性，这些假设是从现有数据推断未观察到的反事实所必需的。当谨慎应用时，该方法使研究者能够量化因果影响、评价干预策略，并理解不同总体间的异质性。然而，在此框架下进行严谨的因果推断，高度依赖于研究设计、假设的合理性以及观察协变量的质量。诸如未测量的混杂因素、干扰或选择偏倚等问题，都可能破坏对估计因果影响的解释。因此，尽管潜在结果框架为因果分析提供了坚实的理论基础，但其实际成功离不开严格的执行过程和对方法学细节的细致把握。

第十八章

准实验

本章将介绍准实验（Quasi-experiments）的基本概念、问题定义、假说、原理、方法学以及一个案例研究。李鸿霄是本章的主要贡献者。我从评价学的角度浅显易懂地解释了准实验可以用来干什么，同时修改了全文。

18.1 准实验基本概念

本章将复用第 十六章定义的大多数概念，本章只介绍准实验中与 RCTs 中不同的概念。

准实验概念定义 1 (自然分组)**.** 自然分组（*natural group*）是指在现实世界（*real world*）中已经存在的组织结构 *[191]*。

准实验概念定义 2 (比较组（comparison groups）)**.** 比较组是准实验里的中心概念，它利用自然或现实世界中的变化（*variation*）来构造不同的分组。比较组与随机对照实验或者 *PO* 里的干预组 *(treatment group)*/对照组 *(control group)* 概念完全不同。因为无法保证干预随机分配到不同分组以及不同分组的队列（*cohort*）长度有限，视为干预组或者对照组的比较组通常是不等价的。根据 *Campbell* 等人的观点 *[43]*，“由于实验单位没有被随机分配到至少两种‘试验’条件，达不到实验控制的全部要求，因此对数据使用实验模式的分析与解释不适用。” [1]

准实验概念定义 3 (原因（cause），影响（effect）)**.** 在准实验场景下里，原因（*cause*）指的自然分组或者比较组的差异，影响指的是比较组在结果上的差异。

18.2 准实验可以用来干什么？

这一节，我先从评价学的角度浅显易懂地解释准实验可以用来干什么。由于受试者随机分配到干预组或者对照者不可行或者不道德，准实验引入了比较组的概念。比较组是利用自然或现实世界中的变化（variation），通常基于自然分组，来构造不同的分组。

[1] 根据原作者的意思，“两种试验条件”应该是干预组和对照组。该段文字是詹剑锋的意译，原文为 “the concept of the application of an experimental mode of analysis and interpretation to bodies of data not meeting the full requirements of experimental control because experimental units are not assigned at random to at least two ‘treatment’ conditions.”。

因为无法保证干预随机分配到不同分组以及不同分组的队列（cohort）长度有限，不同比较组通常是不等价的。因此，准实验通过观察近似等价的分组结果的差异来推测干预的近似影响。

准实验是一种经验方法，缺少严谨的方法学基础。

18.3 问题定义

当随机分配参与者到干预组或者对照组不可行或不道德时，准实验设计 [240] 是一种可以用于估计因果关系的研究设计。与 RCTs 不同，它缺乏对变量的完全控制，但仍然比较暴露在不同条件下的组。

问题定义 10 (准实验问题定义)**.** 准实验问题可以定义为“当随机机制（随机化）不可用时，如何估计平均干预影响（*average treatment effect*，*ATE*)” *[42]*。特别地，准实验关注一段时间内的结果。

与 *RCTs* 一样，准实验的输入是参与者、干预组、对照组和每个参与者的观察结果。输出是干预的近似影响，而非因果影响。

18.4 基本假说

准实验假说 1 (自然分组或者比较组假说)**.** 准实验的基础假说是隐含的，可以通过比较自然分组或者有意地构造不同的比较组，通过观察结果的差异来推测原因对象的影响 *[42]*。

18.5 基本原理

当随机化不可行或不道德时，准实验提供了一个结构化的框架来近似估计影响。准实验原理通过以模拟随机化的方式精心识别变化来源（sources of variation），显式区分因果关系与关联关系。这些框架融入了先验的领域知识，设计并论证比较组（comparison groups）的选择，确保其能够近似反事实情景。准实验通过利用自然或现实世界中的变化（variation），基于自然分组或者在其基础上构造比较组来估计干预措施的影响。

18.6 方法学

准实验设计 [240] 从简单的前后设计（before-and-after design）演变而来，遵循以下五个主要策略，每种均针对特定的内部有效性威胁。

1. 添加非随机对照组（*non-randomized control group*）[240]。这种方法类似于 RCTs。然而，完全随机分组有时是不可能的。因此，采用准实验设计。

在本章里，我们使用符号 O 来表示测量，而符号 X 来表示干预。如果有两行符号，则代表有两个分组。对应于一组中的符号 X，另一组中的空白表示不干预。该方法可以表示为：

O X O
O O

示例 [240]：为了降低驯鹿牧民的高伤害率，设置了三个地理分组：A 组获得预防措施信件，B 组从训练有素的卫生人员那里获得信息，对照组什么都不获得。

在试验前（1985 年）和试验后（1987 年），所有分组的事故率都下降了：A 组：18.7→15.1；B 组：21→14.9；控制组：19.2→14.6。根据标准差阈值要求，未出现显著组间差异，认定干预无效。

2. 进行更多测量 [240]。与单一的前后测量不同，多次基线测量（baseline measurement）用于建立干预前的状态，多次干预后测量用于建立干预后的状态。基于更多测量进行评价，结果更为可靠。

该方法可以表示为：

$$\mathrm{O\ O\ O\ X\ O\ O\ O}$$

为单个测量前后的示例，而

$$\mathrm{O\ O\ O\ X\ O\ O\ O}$$
$$\mathrm{O\ O\ O\quad O\ O\ O}$$

显示了多个基线测量的示例。

示例 [240]：一家食品厂对两个部门进行了安全培训。首先花了 3 ~ 4 周的时间对两个部门进行安全行为基线检查（根据检查表）。接着，在干预后（包括教育、安全清单和反馈），检查再次进行。结果显示，各科室的安全行为在干预后发生了变化，加强了干预与行为之间有关联 (link) 的可能。

3. 延迟引入干预 [240]。“延迟” 指所有分组最终都接受了干预，但干预时间不同，每一组都作为其他组的对照。这最大限度地减少了时间顺序影响（chronological effects）：偶然事件与某一次干预同时发生是可能的，但多次巧合是小概率。

例如，针对不同群体的薪酬改善干预措施在不同的时间实施。因此，薪酬改善的影响差异较少受到偶发事件的干扰，例如自然灾害或经济波动。该方法可以表示为：

$$\mathrm{O\ O\ O\ X\ O\ O\ O\quad O\ O\ O}$$
$$\mathrm{O\ O\ O\quad O\ O\ O\ X\ O\ O\ O}$$

示例 [240]：同一家食品厂首先在打包部门开展安全培训，然后在配料部门开展。每个部门在培训前都有基线检查。延迟干预可以让部门进行比较，减少外部事件（extraneous events）影响结果的可能。

4. 逆转干预 [240]。比较干预的结果和逆转干预的结果。如果逆转后的结果恢复到基线水平，则推定干预有因果影响。这种方法消除了测试过程中涉及的时间顺序和偶然因素威胁和误差，但只适用于干预可逆转的场景。符号 $-X$ 表示逆转干预，该方法可以表示为

$$\mathrm{O\ O\ O\ X\ O\ O\ O\ {-X}\ O\ O\ O}$$

示例 [240]：某公司计划开展参与式的人体工学项目，包括任务修改、新设备和工人教育。为了测量其影响，协调员将比较项目前后的症状和伤害，同时跟踪设备购买情况，以及对任务和压力源的自我报告，以排除其他因素。

5. 测量多个结果 [240]。有两种方法。第一个是增加对干预结果的测量。跟踪实施情况以及短期或中期影响，以区分无效干预措施和有缺陷的实施。例如，“训练教练” 项目没有减少伤害归因于执行不力，而不是无效的技术。其次，添加相关但非直接目标的结果测量。测量与主要目标相似但不受干预影响的结果。该方法可以表示如下，其中 $\mathrm{O_1/O_2}$ 表示两个测量的量：

$$\mathrm{O_1/O_2\ X\ O_1/O_2}$$

示例 *1* [240]：增加干预测量（intervening measures）：一家公司的人体工程学项目跟踪最终结果（症状/损伤），并增加了设备购买记录和工作任务报告等检查。这有助于判

断失败是来自糟糕的项目还是不佳的实现。

示例 2 [240]：增加非目标性测量：石油平台跟踪与大钳相关的损伤（新设备的目标）和非大钳损伤（非目标）。非大钳损伤无变化，说明大钳损伤的变化可能是由器械引起的。一项人体工程学干预发现，目标（颈部/背部）不适有所减轻，但非目标（手臂/手腕）不适没有减轻，从而提高了评价的有效性（validity）。

18.7　案例：推测苹果原产地对苹果采购价的影响

准实验

案例研究

本例使用第一种策略。为了研究苹果原产地如何影响其市场售价，在两个原产地（指定为对照组和试验组）进行实验。然而，由于所选原产地的样本量不足（由于政策或实际条件限制），从原产地本身及其邻近地区收集补充样本数据。该设计利用了自然特征：对于样本量小的原产地，地理上相邻的区域通常具有相同的气候、土壤和农业实践，可以作为补充。准实验设计近似保证了两个分组之间唯一的系统差异是“原产地标签（OR）”，而混杂因素（SW, SI, SL）保持平衡，其中 SW 代表甜度（12、14、16 Brix），SI 代表苹果尺寸（小、中、大），SL 代表保质期（3 天、一周、两周）。

1. 试验组
   - 范围：距离原产地 1 不超过 5 公里的农户。
   - 标签：苹果销售时带有目标原产地认证（$OR = 1$）。

2. 对照组
   - 范围：距离原产地 2 不超过 5 公里的农户。
   - 标签：苹果销售时带有目标原产地认证（$OR = 2$）。

表 18.1是两个原产地每公斤苹果的价格。

分组	T_i	$E[Y_i(T_i)]$
试验组	1	¥8/kg
对照组	0	¥14/kg

表 18.1: 两次试验中每公斤苹果的平均收购价格。

计算

分析价格的影响前需要验证 A, B, C 在两个分组之间是均衡的，以排除非原产地因素对价格的影响。

1. 数据收集
   - 在每个区域采样，收集 50 个苹果样本。
   - 测量指标：甜度（SW，通过 Brix 折射仪 ），尺寸（SI，通过单个水果的克重），保质期（SL，正常储存的最大天数）。

2. 平衡测试方法（*Balance Test Methods* ）
   - 描述性统计量：计算三组 SW、SI、SL 的平均值、标准差和中位数。
   - 统计显著性检验：采用独立样本 t-检验，验证三组间 SW、SI、SL 的平均值差异不具有统计显著性（标准阈值：p-值 <0.05）。
   - 分布可视化：绘制 (SW, SI, SL) 的核密度图（kernel density plots）；重叠曲线表示分布均衡。

确定混杂因素的均衡后，通过直接比较计算原产地（X）对购买价格（Y）的影响：

1. 价格数据收集
   - 跟踪同一批商户，收集两个分组苹果的采购价格。
   - 记录 PP 为符合相同基本品质标准的苹果的单价。

2. 影响计算
   1）计算干预组（$\bar{PP}_1$）和对照组（$\bar{PP}_0$）的平均购买价格。
   2）计算影响:Effect $= \bar{PP}_1 - \bar{PP}_0 = 8 - 14 = -6$¥.
   解释：该值代表“目标产地标签”带来的额外价格，混杂因素(SW, SI, SL)通过地理匹配（geographic matching）均衡。

结论

结果显示，原产地 1 种植的苹果的平均收购价为 ¥6/kg，低于原产地 2 种植的苹果。因此，推测价格的差异可能因为原产地而不是其他混杂因素。

18.8　局限性

准实验的主要局限性是缺乏内部有效性，更不用说外部有效性 [240, 42]。内部有效性缺乏的原因如下。

1. 缺少随机化：在准实验中，对象不是随机分配到干预组和对照组的。相反，分组是

基于预先存在的条件作为弱替代（weak substitute）。缺少随机化导致了选择偏差，组间的系统性差异可能独立于试验组和对照组而存在。

2. 混杂因素的影响：如果不能完全控制实验环境，其他混杂因素可能同时干扰自变量和因变量，掩盖或放大自变量的真实影响。

3. 历史性影响：由于准实验经常在现实场景中实施，过程中发生的一些外部事件可能会影响因变量，与实验干预无关。这些外部事件会对结果产生混杂。例如，在特定时期的政策干预可能会受到同时发生的经济变化的干扰。

18.9 总结

准实验是一种经验方法，它通过利用比较自然分组或者有意构造不同的比较组，通过观察结果的差异来近似地推断干预的影响。尽管准实验在社会科学及其他一些领域中具有重要应用价值，但其内在有效性和外在有效性都难以保证。

第十九章

结构因果模型

本章将介绍结构因果模型（Structural Causal Model，SCM）的基本概念、问题定义、假说、原则、方法学以及案例研究。王晨曦、王磊、李鸿霄是本章的主要贡献者，康国新参与了本章的修改。我从评价学的角度浅显易懂地解释了 SCM 可以用来干什么，同时修改了全文。

19.1 结构因果模型基本概念

SCM 概念定义 1（内生（Endogenous）变量）. 内生变量是模型内所关注的可观察变量，受所建立的模型内部其他变量的影响 *[221]*。

SCM 和 IV 概念定义 1（外生（Exogenous）变量）. 外生变量来自于模型之外，通常作为外部输入，它不受模型内部任何其他变量的影响 *[221]*。

SCM 概念定义 2（干预（Intervention））. 干预在数学上形式化地表达为 $\mathbf{do}(X = x)$ 运算符，它指的是将干预变量 X 设置为其值范围内的固定值的物理操作，这里的固定值用 x 表示 *[221, 224]*。设置固定值的本质是对 X 进行控制，从而阻断所有指向干预变量 X 的有向边，排除混杂影响。

朱迪亚 · 珀尔等人 *[221]* 在因果贝叶斯网络（*Causal Bayesian Network*）的定义中首次形式化了这一概念。结构因果模型（*SCM*）本身是一种数学模型，它通过结构方程刻画变量之间的因果影响；而 $\mathbf{do}(X = x)$ 运算符则是在该数学模型中形式化表示干预的操作符。从数学模型的角度来看，一次干预的实际效果如下：原始 *SCM*，记为 M，删除 M 中 X 的结构方程，将其替换为带有固定值的方程 $X = x$，产生修改的子模型 M_x。

SCM 里的干预是评价学中存在算子的一个特例，它以对量（可计数属性）操作的形式来体现。

SCM 和 IV 概念定义 2（混杂变量（confounding variable））. 混杂变量是同时影响原因变量和结果变量的变量 *[224]*。*SCM* 和 *IV* 的混杂变量是 *PO* 里影响结果的协变量，反之则不一定。

SCM 概念定义 3（变量 X(因) 对变量 Y（果）的因果影响）. 变量 X(因) 对变量 Y（果）的因果影响定义为 $\mathbb{P}(Y = y|\mathbf{do}(X = x))$，即 Y 在子模型 M_x 中的分布，其中 Y

为单元（*Unit*）的观察变量，X 是干预变量 *[221, 224]*。$\mathbb{P}(Y = y|\mathbf{do}(X = x))$ 不同于 $\mathbb{P}(Y = y|X = x)$，后者包含变量 X(因）对变量 Y（果）的因果影响以及混杂变量的影响。

SCM 概念定义 4 (因果结构方程的噪声项)**.** 因果结构方程的噪声项是指捕获了未观察的变量影响和随机变化的项 *[221]*。

SCM 和 IV 概念定义 3 (中介/调节变量(mediators/moderators))**.** 中介/调节变量(*mediators/moderators*）是指在连续的因果层次中传递或修改因果影响的变量 *[221, 224]*。

SCM 概念定义 5 (有向无环图(DAG))**.** 有向无环图是图论中使用的一种表示，在 *SCM* 里作为因果结构的表示方法，其节点表示变量，一条有向边 $X \to Y$ 表示因果关系，由因 X 指向果 Y *[309, 19]*。*SCM* 要求图必须是无环的，也就是说没有箭头路径允许被指向已经包含在该路径中的节点。

DAG 本质上表示的是一种经验假设，例如，两个变量之间没有箭头，表明它们之间没有因果关系，这就是经验假设。

SCM 概念定义 6 (接合（junction）)**.** 接合是复杂 *DAG* 的基本构造体。*DAG* 中存在三种基本的接合（*junction*)：链（*Chain*)、叉（*Fork*）以及碰撞（*Collider*） *[224]*。

SCM 概念定义 7 (链（Chain）)**.** 链是 *DAG* 的一个基本接合，以 $A \to B \to C$ 来描述三个节点 A、B、C 之间有向边连接的接合形式 *[224]*。在不控制节点 B 的条件下，链接合是连通的，其观察数据会显示节点 A 和节点 C 是统计上相关的。而在控制节点 B 的条件下会阻断此接合，即节点 A 和节点 C 是统计上独立的。

SCM 概念定义 8 (叉（Fork）)**.** 叉是 *DAG* 的一个基本接合，以 $A \leftarrow B \to C$ 来描述三个节点 A、B、C 之间有向边连接的接合形式 *[224]*。在不控制节点 B 的条件下，叉接合是连通的，其观察数据会显示节点 A 和节点 C 是统计上相关的。而在控制节点 B 的条件下会阻断此接合，节点 A 和节点 C 是统计上独立的。

SCM 概念定义 9 (碰撞（Collider）)**.** 碰撞是 *DAG* 的一个基本接合，以 $A \to B \leftarrow C$ 来描述三个节点 A、B、C 之间有向边连接的接合形式 *[224]*。只有控制节点 B 的条件下，碰撞接合才是连通的，其观察数据会显示节点 A 和节点 C 是统计上相关的。而在不控制节点 B 的条件下此接合是阻断的，节点 A 和节点 C 是统计上独立的。

SCM 概念定义 10 (路径（path）)**.** 路径是指 *DAG* 中一个由节点与有向边交替构成的有序序列 *[309, 19, 224]*。该序列严格以根节点起始，以叶节点终止，其间顺次经过零个或多个内部节点。

SCM 概念定义 11 (前门路径（front-door path)，后门路径（back-door path）)**.** 基于 *DAG* 研究变量 X（因）对变量 Y（果）的影响时，前门路径是指由所有起点为 X 同时方向指向 Y（终点）的有向边形成的路径，即 $X \to \cdots \to Y$，前门路径是没有混杂变量影响的因果路径；后门路径是由起点 X 到终点 Y 并且第一条有向边指向起点的路径，即 $X \leftarrow Z \ldots Y$，未被阻断的后门路径引入了混杂变量 Z 的影响 *[224]*。

SCM 概念定义 12 (直接原因 (direct cause)). 直接原因由直接因果关系 $X \to Y$ 定义: 如果 X 是因果图中 Y 的父节点, 并且作为 Y 结构方程中的一个参数出现, 那么 X 是 Y 的直接原因。这种关系量化了 X 对 Y 的因果关系, 表明无论模型中其他变量取何值, X 的变化造成 Y 的相应变化。在这个上下文里, 变量 Y 分布的变化被称为 X 的影响 *[221, 224]*。

在实际进行因果推理的时候, 难以获得完整先验知识, 即 DAG 图未知, 为此朱迪亚 · 珀尔定义了潜在原因。

SCM 概念定义 13 (潜在原因 (potential cause)). 如果满足以下条件, 则可以从观察联合分布 $\hat{P}$ 中推断出变量 X 对变量 Y 具有潜在因果影响 *[221]*。

1. X 和 Y 在每个 "上下文" (*context*) 中都是不独立的。这里的 "上下文" 是指一组被固定为特定取值的变量。也就是说对这些变量进行控制, X 和 Y 仍然不独立。
2. 存在一个变量 Z 和一个上下文 S, 使得:
   (i) 给定 S 时, X 和 Z 相互独立, 即 $X \perp\!\!\!\perp Z \mid S$.
   (ii) 给定 S 时, Z 和 Y 不独立, 即 $Z \not\perp\!\!\!\perp Y \mid S$.

19.2 结构因果模型可以用来干什么?

这一节, 我先从评价学的角度浅显易懂地解释结构因果模型 (SCM) 可以用来干什么。

在研究原因变量 X 对结果变量 Y 的因果影响时, SCM 的前提是能用 DAG 全面、准确地表达相关变量之间的影响关系。

在充分利用 DAG 因果结构特征的前提下, SCM 通过直接或者间接控制符合某种特定准则—例如前门准则、后门准则或者 do 演算—的变量, 隔离混杂变量的影响, 从而准确地估计原因变量的因果影响。

19.3 问题定义

SCM [133, 224] 是一个基础的统计和概念框架, 使用因果图和结构方程来表示和推断变量之间的因果关系。该方法广泛应用于流行病学、经济学、计算机科学和社会科学等领域, 用于因果发现 (causal discovery)、中介分析 (mediation analysis) 和政策评价 (policy evaluation) 等任务。

问题定义 11 (SCM 问题定义). *SCM* 的问题定义如下: 如何对系统背后潜在的因果机制进行建模? 如何超越数据中观察到的简单的相关模式 (*correlation pattern*), 估计因果影响? *[224]*"

该问题的输入包括对关注的变量的因果结构进行编码的 *DAG* 和外生噪声变量的分布。输出为对这些关注的变量之间因果影响的估计。约束条件是图必须是有向无环图, 每个变量有唯一的结构方程, 外生噪声项满足给定的独立性假说。

19.4　基本假说

结构因果模型方法依赖于以下三个假说。

SCM 假说 1（因果有向无环图假说 [221]）. 假设变量之间的因果关系可以用 *DAG* 表示，并且图是无环的，即没有变量可以通过因果路径影响其自身。

SCM 用有向无环图（*DAG*）表示，其中：

- 节点表示变量。
- 有向边表示因果关系从原因变量指向结果变量。

SCM 假说 2（结构方程（Structural Equations）假说 [221]）. 假设每个变量 V_i 被由其直接原因（*DAG* 中的父变量 $\mathbf{Pa}(V_i)$）和表达外生变量的独立噪声项 U_i 的函数决定：

$$V_i := f_i(\mathbf{Pa}(V_i), U_i), \tag{19.1}$$

其中：

- $\mathbf{Pa}(V_i)$ 表示 *DAG* 中 V_i 的直接原因（父节点）。
- f_i 是一个确定的函数，描述 V_i 如何依赖于它的父节点。
- U_i 是捕获外生影响的外生变量或噪声项。

SCM 假说 3（因果 Markov 条件（Causal Markov Condition）假说 [113, 221]）. 在一个表达因果关系的有向无环图中，在给定其父变量的条件下，每个变量 V_i 与其非后代变量V_j是条件独立的。

$$V_i \perp\!\!\!\perp V_i\text{的非后代节点}V_j \mid \mathbf{Pa}(V_i), \tag{19.2}$$

该假设允许基于局部因果关系对全局概率分布进行因式分解（*factorized*）。

19.5　基本原理

SCM 提供了一种形式化语言，用于编码因果假设、推测并量化干预的影响。与纯粹的统计关联不同，SCM 通过一组结构化方程及相应的图形表示，明确地融入了因果知识。

该形式化语言明确区分因果关系和相关关系（associational relationship），将领域内先验知识编码到因果图的结构中，其中边表示直接因果影响，通过干预（模拟变量的外部操作）来定义和计算反事实，并回答“如果... 会发生什么（what if）”问题 [224]。

这些框架通常涉及一组结构方程，这些方程量化了每个变量是如何由其直接原因和未观察的噪声项决定的，同时方程也区分了混杂变量和中介/调节变量（mediators/moderators）[224]。

19.5.1 SCM 的能力

SCM 使用三元组 (V, U, F) 形式化，其中：

- V 是一组内生变量。
- U 是一组外生变量，这些变量不由 V 中的任何变量所引起。
- F 是一组函数 $f_1, f_2, \cdots, f_n$，这些函数根据其因果父变量 $Pa(V_i)$ 的值和外生变量 U_i，赋予 V 中的每个变量 $V_i \in V$ 一个值，即 $V_i = f_i(Pa(V_i), U_i)$。

SCM 提供以下能力。

- 明确的因果假设：图模型要求研究人员提供透明化和可检验的因果假设。这些假设的含义（例如条件独立性）可以根据数据进行检验。
- 混杂处理：SCM 为定义、识别和调整混杂变量影响提供了一个清晰的框架，这些变量是观察性研究（observational studies）中常见的偏差（bias）来源。
- 统一概念：SCM 将统计学和图论的概念统一到一个统一的框架中，促进对因果关系的更深入理解。
- 反事实推断：它为回答反事实问题提供了一种形式化语义，这是解释和有效推理（valid reasoning）的核心。

19.6 方法学

19.6.1 因果影响识别与 DAG

在 SCM 中，首先要对所有内生变量$V = \{V_i\}$ 使用 DAG 进行因果建模，这个 DAG 称为因果图 [224]。因果图的建模首先将所有内生变量以节点形式画在图上，然后对于图上任意两个节点，确定它们之间是否存在表示因果关系的有向边。对于图上两个确定节点所代表的内生变量，如果研究人员能够确定其中一个变量（原因变量）是另一个变量（结果变量）的直接原因，那么在因果图上就用一条从原因变量节点指向结果变量节点的有向边表示这种因果关系。因果图建模了所有内生变量之间的因果关系。

SCM 概念定义 14. 因果影响识别 *(Identification of Causal Effects)* 是 *SCM* 中的核心问题：在给定因果图结构和观察数据分布的前提下，判断某个变量 X 对结果变量 Y 的因果影响 $\mathbb{P}(Y|\mathbf{do}(X))$ 是否能够被唯一确定 *[224]*。

对于因果图中任意两个节点，若这两个节点之间存在路径，则需要分辨这些路径中的因果路径和非因果路径（后门路径）。其中，因果路径代表根节点（原因）变量和叶节点（结果）变量之间真正的因果影响机制；而由于系统中往往存在未观察的混杂变量，这些混杂变量出现在根节点变量和叶节点变量之间的后门路径中引入混杂干扰，并不代表根节点变量和叶节点变量之间存在因果关系。

在 SCM 中，阻断因果图中根节点到叶节点的所有后门路径是至关重要的一步，实现这一步是对真正的因果影响进行求解计算的常用方法，即后门准则（*backdoor criteria*）

[224]。除此之外，在 SCM 中还可以通过前门准则 *(frontdoor criteria)*、*do* 演算 *(do-calculus)* 等方法求解因果影响，从而实现因果影响识别 [224]。第 19.6.3节将详细介绍这些求解方法。

19.6.2 基本方程

SCM 的核心是结构方程。对于每个内生变量 V_i，有：

$$V_i := f_i(\mathbf{Pa}(V_i), U_i), \tag{19.3}$$

其中 $\mathbf{Pa}(V_i)$ 是图中 V_i 的父变量，即 V_i 的直接原因；U_i 是外生变量或噪声项。

19.6.3 求解方法

基础工具：do 演算

do 演算（do-calculus）[224] 是一套基于因果图拓扑结构的代数公理系统，也是一套数学法则，它将因果图的结构特征转化为严格的符号推导条件，从而指导对概率表达式（含干预算子）的等价变换与化简。do 演算是 SCM 的基础方法。它包含三条基本规则，其核心目的是在满足特定图论条件时，对含有 $\mathbf{do}(\cdot)$ 的概率表达式进行等价变换，最终实现因果影响的可识别。

- *1*：观察的插入/删除规则 *(Insertion/deletion of observations)*

  具体含义：如果我们在已知变量集合 Z 的条件下，观察到一个与 Y 无关的变量 W，那么 Y 的概率分布不会改变。

  公式：

  $$\mathbb{P}(Y|\mathbf{do}(X), Z, W) = \mathbb{P}(Y|\mathbf{do}(X), Z), \tag{19.4}$$

  基于 DAG，这个规则可以解释如下。

  条件：在删除了所有指向 X 的有向边后，变量集合 Z 能够阻断从 W 到 Y 的所有路径。

  如果满足以上条件，则公式 19.4成立。

- *2*：干预/观察的等价交换规则 *(Action/observation exchange)*

  含义：如果变量集合 Z 阻断了从 X 到 Y 的所有后门路径，那么在给定 Z 的条件下，“干预 X” 与 “被动观察 X” 是完全等价的。完全地控制了混杂变量集合后，剩余的相关性就是真实的因果影响。

  公式：

  $$\mathbb{P}(Y|\mathbf{do}(X), Z) = \mathbb{P}(Y|X, Z), \tag{19.5}$$

  基于 DAG，这个规则可以解释如下。

  条件：Z 满足相对于 (X, Y) 的后门准则，后门准则的详细介绍见章节 19.6.3。

  如果满足以上条件，则公式 19.5成立。

- *3*：干预的插入/删除规则 *(Insertion/deletion of actions)*

  含义：如果系统内不存在从 X 指向 Y 的因果路径，那么对 X 实施的任何强制干预都不会影响 Y 的概率分布。

  公式：

  $$\mathbb{P}(Y|\mathbf{do}(X)) = \mathbb{P}(Y), \tag{19.6}$$

  基于 DAG，这个规则可以解释如下。

  条件：不存在从 X 到 Y 且仅包含从 X 指向 Y 有向边的因果路径。

  如果满足以上条件，则公式 19.6成立。

这三条规则在句法上赋予了明确的操作权限：规则 1 允许增加或删除观察变量；规则 2 允许在干预和观察之间进行等价替换；规则 3 允许增加或删除干预操作。所有这些代数变换的合法性，都严格依赖于具体因果图（DAG）表达的因果关系假设的验证：如果假设验证为真，则变化合法。

混杂可控的求解方法：后门准则

SCM 中的一个核心问题是能否从观察数据和给定的因果图中确定干预影响 $\mathbb{P}(Y|\mathbf{do}(X = x))$。后门准则 [224] 是解决这一问题的通用且关键的方法，主要用于处理可观察的混杂变量（混杂可控）。

后门准则的核心思想是：通过阻断干预变量 X 与结果变量 Y 之间的非因果路径（即后门路径），来消除混杂偏倚（Bias）。所谓“后门路径”，是指任何从 X 连接到 Y，且以指向 X 的有向边作为起点的路径（例如，在路径 $X \leftarrow Z \rightarrow Y$ 中，Z 就是混杂变量）。为了对 X 和 Y 进行去混杂，我们需要阻断它们之间所有的后门路径，同时不能阻断或干扰任何真实的因果路径。如果变量集合 Z 满足以下两个条件，则称 Z 相对于变量对 (X, Y) 满足后门准则：

- **非后代条件**：Z 中不包含任何 X 的后代节点。这避免将中介变量或其他处理后的变量纳入调整集，从而阻断真实的因果传导路径。
- **阻断条件**：Z 阻断了从 X 到 Y 的所有后门路径。这切断了导致 X 和 Y 产生虚假相关性（Spurious correlation）的通道。

当变量集合 Z 满足后门准则时，X 对 Y 的真实因果影响可以通过以下后门调整公式（Back-door Adjustment Formula）进行精确识别与计算 [224]：

$$\mathbb{P}(Y|\mathbf{do}(X = x)) = \sum_{z} \mathbb{P}(Y|X = x, Z = z)\mathbb{P}(Z = z), \tag{19.7}$$

公式 19.7可以解释为在混杂变量 Z 的所有可能取值上进行加权求和，它由两部分组成：

- $\mathbb{P}(Y|X = x, Z = z)$：这是在给定 $X = x$ 且混杂变量 $Z = z$ 的特定“切片”或“分层”下，Y 的观察条件概率。

- $\mathbb{P}(Z = z)$：这是变量 Z 在没有任何干预时的自然概率分布，它作为每个分层的权重。

在实际的数据分析中，计算方法通常遵循“分层与加权”的步骤：

1. **寻找后门调整集**：首先利用因果图识别出满足后门准则的变量集合 Z。
2. **分层估计（Stratification）**：将数据按照变量 Z 的不同水平（例如，如果 Z 是性别，则分为男性组和女性组）进行分层。在每一个特定的分层内，由于后门已经被 Z 阻断，观察到的 X 与 Y 的关系能够反映该分层下的因果关系。
3. **加权平均（Weighted Average）**：计算 Z 的每个分层在总体人群或样本中出现的比例，用这些比例作为权重，对各分层得到的结果进行加权平均，从而恢复总体层面的因果效应。

通过这种方式，只要有足够的数据支持，并且能够观察到足以阻断所有后门路径的变量，后门准则就能将复杂的因果推断转化为逐层计算观察数据的统计问题。

后门调整公式可以完全由 do 演算进行推导，相关推导如下 [224]：

$$
\begin{aligned}
P(Y \mid do(X = x)) &= \sum_z P(Y \mid do(X = x), Z = z) P(Z = z \mid do(X = x)), && \text{概率公理} \\
&= \sum_z P(Y \mid X = x, Z = z) P(Z = z \mid do(X = x)), && \text{规则 2} \\
&= \sum_z P(Y \mid X = x, Z = z) P(Z = z). && \text{规则 3}
\end{aligned}
$$

典型案例：教育年限与收入的因果推断 在社会科学研究中，后门准则常用于解决混杂变量导致的因果推断偏差 [221]。一个经典案例是评估教育年限（干预变量 X）对个人收入（结果变量 Y）的真实影响。假设存在一个关键的混杂变量——家庭背景（Z），它同时影响教育选择（X）和收入水平（Y）。例如，家庭经济条件优越的个体更可能接受高等教育（X），同时也更可能凭借家庭资源获得高收入（Y），从而在 X 与 Y 之间形成虚假关联。

为应用后门准则，首先绘制因果图：X(教育年限) $\to Y$(收入)，但图中同时存在后门路径 $X \leftarrow Z$(家庭背景) $\to Y$。我们将“家庭背景（Z）”作为控制变量。由于 Z 不是 X 的后代，且在给定 Z 后，唯一连接 X 和 Y 的后门路径被彻底切断，因此 Z 满足后门准则。在实际计算中，我们使用后门调整公式，将观察样本按家庭背景（Z）分层，分别计算同一背景阶层内不同教育年限带来的收入差异，最后再按照各阶层在总人口中的真实比例进行加权平均。

混杂不可控的求解方法：前门准则

当存在未观察到的混杂变量 U，该混杂不可控制，导致后门路径无法被直接阻断时，无法直接使用后门调整公式识别因果影响。此时，如果干预变量 X 和结果变量 Y 之间存在一个合适的中介变量 M，我们就可以利用前门准则（*Front-door Criterion*）[224] 来识别因果影响。

前门准则的核心思想是利用完全的中介传导机制来“绕开”不可观察的混杂。如果变量（或变量集合）M 满足以下三个条件，则称 M 相对于变量对 (X, Y) 满足前门准则：

- *条件 1*：M 阻断了所有从 X 指向 Y 的因果路径。这意味着 X 对 Y 的所有因果影响必须且只能通过 M 来传递，不存在绕过 M 的因果路径。
- *条件 2*：从 X 到 M 之间没有任何未被阻断的后门路径。这意味着不可观察的混杂变量 U 不能直接干扰 M。
- *条件 3*：所有从 M 到 Y 的后门路径都可以被 X 阻断。

当满足前门准则时，X 对 Y 的真实因果影响可以通过以下前门调整公式（Front-door Adjustment Formula）进行精确计算，该公式完全不需要观察到混杂变量 U：

$$\mathbb{P}(Y|\mathbf{do}(X=x)) = \sum_{m} \mathbb{P}(M=m|X=x) \sum_{x'} \mathbb{P}(Y|X=x', M=m)\mathbb{P}(X=x').$$

前门公式本质上是两段因果影响的组合：

- $\mathbb{P}(M=m|X=x)$：这代表因果链条的第一步（$X \to M$）。由于该路径没有混杂，观察概率直接等于因果影响。
- $\sum_{x'} \mathbb{P}(Y|X=x', M=m)\mathbb{P}(X=x')$：这代表因果链条的第二步（$M \to Y$）。在这里，$X$（记为 x'）被当作了混杂变量进行后门控制，从而算出了 M 对 Y 的没有被混杂变量干扰的影响。

前门准则的计算过程本质上是“拆解传导，分段计算，最后合并”的逻辑：

1. *估计 $X \to M$ 的影响*：因为 X 到 M 之间是不受混杂干扰的，我们直接在观察数据中统计，当干预变量 $X=x$ 时，中介变量 M 呈现特定状态的概率。
2. *估计 $M \to Y$ 的影响*：虽然不可观察的 U 同时影响了 X 和 Y，但在计算 M 对 Y 的影响时，我们可以利用已知的 X 作为“虚拟的”后门混杂变量，通过对其进行分层控制，阻断 $M \leftarrow X \leftarrow U \to Y$ 的干扰。
3. *合并两段影响*：将第一步（X 导致特定 M 的概率）与第二步（该特定 M 导致 Y 的概率）相乘，并遍历 M 的所有可能状态进行求和。

前门调整公式可以完全由 do 演算进行推导，相关推导如下 [224]：

$$
\begin{aligned}
P(Y \mid do(X=x)) &= \sum_m P(Y \mid do(X=x), M=m) P(M=m \mid do(X=x)), && \text{概率公理} \\
&= \sum_m P(Y \mid do(X=x), do(M=m)) P(M=m \mid do(X=x)), && \text{规则 2} \\
&= \sum_m P(Y \mid do(X=x), do(M=m)) P(M=m \mid X=x), && \text{规则 2} \\
&= \sum_m P(Y \mid do(M=m)) P(M=m \mid X=x), && \text{规则 3} \\
&= \sum_{x'} \sum_m P(Y \mid do(M=m), X=x') P(X=x' \mid do(M=m)) P(M=m \mid X=x), && \text{概率公理} \\
&= \sum_{x'} \sum_m P(Y \mid M=m, X=x') P(X=x' \mid do(M=m)) P(M=m \mid X=x), && \text{规则 2} \\
&= \sum_{x'} \sum_m P(Y \mid M=m, X=x') P(X=x') P(M=m \mid X=x). && \text{规则 3}
\end{aligned}
$$

典型案例：吸烟与肺癌的因果推断　在研究吸烟与肺癌时候，吸烟（X）导致肺癌（Y）存在某种未知的“致病基因（U），既让人易染烟瘾又易诱发肺癌。此时存在后门路径 $X \leftarrow U \rightarrow Y$。由于基因 U 在当年研究时候的不可观察，后门准则完全失效。科学家通过引入肺部焦油沉积（M）作为中介变量。M 满足前门准则的三大条件：

- **条件 1**：吸烟必须通过产生焦油等物质的沉积来引发肺癌。
- **条件 2**：未知基因不会凭空在肺部产生焦油，焦油只能因吸烟吸入（$X \rightarrow M$ 无混杂）。
- **条件 3**：焦油到肺癌（$M \rightarrow Y$）虽受基因干扰，但可通过可控制的吸烟行为（X）来阻断。

利用前门公式，分析被拆解为两步：先统计吸烟产生焦油的直接概率，再在分别控制吸烟/不吸烟人群的前提下，统计焦油引发肺癌的概率。最后将两段概率相乘并求和，从而绕过了不可观察的基因干扰。

19.6.4　进阶推断：反事实推断

正如朱迪亚 · 珀尔在其著作《为什么》[224] 中所指出的，反事实分析利用观察数据和实验数据来推断假设情境下的结果。在朱迪亚 · 珀尔的因果阶梯（Ladder of Causation）中，反事实处于最顶层。我们在前文讨论的 **do** 演算、后门与前门准则，本质上都是在计算群体级别的平均干预影响。然而，在真实场景中，我们需要回答的是针对特定个体或单次事件的追问：“已知现象已经发生，如果当初条件不同，结果会改变吗?”这种超越了群体平均影响，实现对特定单一事件历史回溯与逻辑推演的工具，正是反事实推断。

因果关系的三个维度　为了在科学研究（如气候异常归因）和司法实践中做出更精确的定量声明，反事实分析要求我们在词汇上严格区分三种不同强度的因果关系。设 X 和 Y 为二值变量，其中 $X=1$ 表示事件 X 发生，$X=0$ 表示事件 X 未发生；$Y=1$ 表示结果 Y 发生，$Y=0$ 表示结果 Y 未发生。

1. **必要因果关系（Necessary causation）**：对应于法律上的“若非（but-for）”因果关系，这主要用于事后的“责任归属”判定。刻画的是在已经观察到 $X = 1$ 且 $Y = 1$ 的真实世界条件下，若反事实地令 $X = 0$，则反事实结果为 $Y = 0$ 的概率 [222]。

2. **充分因果关系（Sufficient causation）**：这主要用于事前的“政策有效性”评估。刻画的是在已经观察到 $X = 0$ 且 $Y = 0$ 的真实世界条件下，若反事实地令 $X = 1$，则反事实结果为 $Y = 1$ 的概率 [222]。

3. **充要因果关系（Necessary-and-sufficient causation）**：兼具必要性与充分性。刻画的是 Y 会随 X 的发生而发生、并随 X 的不发生而不发生的概率 [222]。

严格的形式化定义与三步计算法则　反事实探讨的是在特定的个体背景特征（外生变量 $U = u$）下，假设对该个体实施了与现实相反的干预 $\mathbf{do}(X = x)$，其结果变量 Y 会取什么值 [224]。在 SCM 中，这被记为 $Y_x(u)$。与群体干预分布 $\mathbb{P}(Y|\mathbf{do}(X = x))$ 不同，反事实 $Y_x(u)$ 的计算强依赖于底层具体的结构方程 F，而不仅仅是因果图的箭头走向。只要拥有完全指定的结构因果模型（即已知结构方程 F 和外生变量的先验分布），任何反事实问题都可以通过以下标准的三步法进行精确计算：

1. **溯因（Abduction）**：利用现实世界中已观察到的个体数据（例如事实上的 $X = x'$ 和 $Y = y'$），代入结构方程进行反推，更新并算出该个体专属的底层背景噪声（即外生变量 $U = u$ 的后验分布）。

2. **干预（Action / Intervention）**：在数学模型中实施反事实的假设干预。即删去原 SCM 中关于 X 的结构方程，替换为假设的常量方程 $X = x$，从而生成一个反事实子模型 M_x。

3. **预测（Prediction）**：将第一步算出的个体专属背景变量 $U = u$，代入第二步生成的修改后子模型 M_x 中，重新正向计算出 Y 的值。这就是反事实结果 $Y_x(u)$。

通过区分必要与充分因果关系，并辅以三步计算法，SCM 能在分析复杂系统时，能够将宏观概率映射到微观个体，给出具有确定性的反事实量化结论。

19.6.5　流程

使用结构因果模型的评价需遵循以下六个步骤，其中第五步至关重要。

1. 定义问题：定义研究问题、变量和变量之间的因果假设。

2. 构造因果图：使用 DAG 来表示变量之间的因果关系，标记混杂变量。

3. 确定可识别性：检查因果关系是否可以估计。

4. 数据处理：收集数据，处理缺失值、异常值等。

5. 因果推断：使用结构公式计算因果影响，例如 19.6.3或 19.6.3。

6. 验证和解释：验证结果的可靠性并得出结论。

19.7　案例：推断苹果产地对苹果采购价的影响

结构因果模型案例分析

SCM 应用背景

我们利用结构因果模型 (SCM) 评估苹果产地 (OR) 对收购价格 (PP) 的真实因果影响。
混杂难题: 存在一个未观察的外生变量 C：经销商档次。例如高端经销商倾向于从产地 1 进货，且无论质量如何都愿意支付更高价格。这通过后门路径 $OR \leftarrow C \rightarrow PP$ 在 OR 和 PP 之间创建了非因果关联。
前门识别: 我们使用一个内生中介变量 SW：甜度。SCM 假设产地 OR 决定了甜度 SW，而甜度决定收购价格 PP。

SCM 变量与结构方程

变量定义:

- OR (内生输入)：苹果产地 ($0 =$ 产地 0，$1 =$ 产地 1)。
- PP (内生结果)：收购价格 (¥)。
- SW (内生中介)：甜度 ($0 =$ 含糖量低，$1 =$ 含糖量高)。
- C (外生混杂)：经销商档次（未观察）。

Structural Equations (结构方程组):

$$\begin{cases} OR = f_{OR}(C, \epsilon_{OR}) & \text{（产地选择受经销商档次影响）} \\ SW = f_{SW}(OR, \epsilon_{SW}) & \text{（甜度仅由产地决定，} C \notin f_{SW}\text{）} \\ PP = f_{PP}(SW, C, \epsilon_{PP}) & \text{（价格由甜度和经销商档次共同决定，} OR \notin f_{PP}\text{）} \end{cases} \tag{19.8}$$

Note: $\epsilon_{OR}, \epsilon_{SW}, \epsilon_{PP}$ 为相互独立的误差项。

SCM 观察数据表

产地 (OR)	甜度 (SW)	平均价格 (¥)	组内占比
$OR = 0$	$SW = 0$	5.0	0.80
	$SW = 1$	10.0	0.20
$OR = 1$	$SW = 0$	8.0	0.30
	$SW = 1$	15.0	0.70

表 19.1: 观察数据。注意：由于 C 的影响，相同甜度下产地 1 的价格更高。

假设边界: $P(OR = 0) = P(OR = 1) = 0.5$. 在我们的观察样本中，产地 0 和产地 1 的苹果数量各占一半（50%）

SCM 识别条件验证（前门准则）

如图 19.1所示，该模型可通过前门准则识别，原因如下：

1. 条件 *1*：在因果图中，OR 仅通过 SW 影响 PP。
2. 条件 *2*：OR 到 SW 之间不存在后门路径。
3. 条件 *3*：OR 阻断了 SW 到 PP 的后门路径。

SCM 因果推断计算

利用 SCM 的结构特性计算干预期望 $E[PP \mid do(OR = 1)]$：
步骤 *1*：确定中介变量的因果影响值（$SW \to PP$）

$$\begin{aligned} E[PP \mid do(SW = 0)] &= \sum_{or'} E[PP \mid SW = 0, OR = or'] P(OR = or') \\ &= 5.0 \times 0.5 + 8.0 \times 0.5 = \mathbf{6.5}\ ¥ \\ E[PP \mid do(SW = 1)] &= \sum_{or'} E[PP \mid SW = 1, OR = or'] P(OR = or') \\ &= 10.0 \times 0.5 + 15.0 \times 0.5 = \mathbf{12.5}\ ¥ \end{aligned}$$

步骤 *2*：通过 *SCM* 计算总影响（$OR \to PP$）

$$\begin{aligned} E[PP \mid do(OR = 1)] &= \sum_{sw} E[PP \mid do(SW = sw)] \cdot P(SW = sw \mid OR = 1) \\ &= 6.5 \times 0.30 + 12.5 \times 0.70 = \mathbf{10.7}\ ¥ \\ E[PP \mid do(OR = 0)] &= \sum_{sw} E[PP \mid do(SW = sw)] \cdot P(SW = sw \mid OR = 0) \\ &= 6.5 \times 0.80 + 12.5 \times 0.20 = \mathbf{7.7}\ ¥ \end{aligned}$$

SCM 结论分析

1. 观察层面相关性 *(Observational Association):*
根据观察数据计算两个产地的平均售价（包含混杂偏差）：

- 产地 1 的平均观察价格：$E[PP|OR = 1] = 8.0 \times 0.30 + 15.0 \times 0.70 = \mathbf{12.9}\ ¥$
- 产地 0 的平均观察价格：$E[PP|OR = 0] = 5.0 \times 0.80 + 10.0 \times 0.20 = \mathbf{6.0}\ ¥$
- 观察差异分析 (Observed Difference)：$12.9 - 6.0 = \mathbf{6.9}\ ¥$

2. 因果影响与洞察 *(Causal Insight):*

- 平均处理影响 *(Average Treatment Effect, ATE)*:

$$ATE = E[PP \mid do(OR = 1)] - E[PP \mid do(OR = 0)] = 10.7 - 7.7 = \mathbf{3.0}\ ¥$$

- 结论分析与因果洞察: SCM 揭示了原始数据中产地 1 的 6.9 元价格优势中具有误导性。在这之中，只有 3.0 元是由产地通过影响甜度带来的因果贡献，剩余的 3.9 元差异并非由产地决定，而是源于未观察到的经销商档次 C 产生的选择偏差（高端经销商更倾向于选择产地 1，并支付了渠道溢价）。

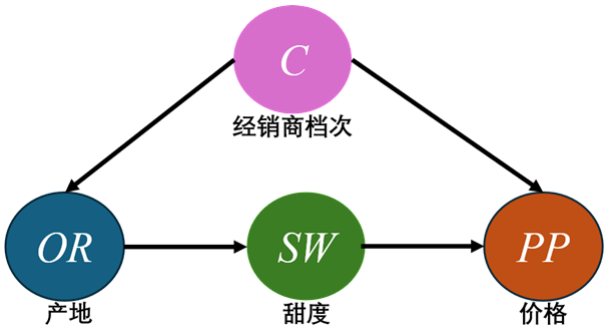


图 19.1: 本示例中多个变量之间的因果关系模型。

19.8　局限性

SCM 是用于理解和分析因果关系的稳健框架。然而，SCM 有以下限制。

首先，它完全依赖于 DAG 因果假设的正确性。如果假设不正确或不可信，则得出的结论可能无效或具有误导性。

其次，SCM 因果结构的提出依赖于足够的背景知识，其合理性需要检验。

第三，SCM 通常需要大数据集，尤其是具有许多变量的复杂模型。测量误差或未观察的混杂变量会进一步损害 SCM 结果的有效性，因为框架假设所有相关变量都被正确测量和包含。

最后，当处理高维数据或试图估计复杂 DAG 中的影响时，计算将会非常昂贵，甚至不可行。

19.9　总结

结构因果模型是描述和估计变量影响的综合框架。通过结构方程系统和相应的图模型引入一个框架来表示因果关系，扩展了传统的统计关联分析。该框架被应用于流行病学、经济学、计算机科学和社会科学等领域，用于因果发现、中介分析和干预影响评价等任务。结构因果模型的局限性是数据收集成本高，假设和因果关系图可能不可靠。

第二十章

工具变量

本章介绍了工具变量（Instrumental Variables，简称 IV）的基本概念、问题定义、假说、原理、方法学以及案例研究。本章主要贡献者为王磊。我从评价学的角度浅显易懂地解释了 IV 可以用来干什么，同时修改了全文。

20.1 工具变量基本概念

IV 概念定义 1 (结果变量（Outcome Variable）). 结果变量（Y）是与所研究的内生变量（*endogenous variable*）X 存在因果关系的因变量（*dependent variable*） *[329]*。

IV 概念定义 2 (内生变量). 内生变量（X）是被假定对结果变量 Y 存在因果影响的自变量（*independent variable*） *[329]*。

IV 概念定义 3 (普通回归方法). 普通回归方法是一种统计方法，通过最小化结果变量的观察值与预测值之间的平方差之和，来拟合解释变量（*explanatory variable*）与结果变量之间的线性关系 *[78]*。普通回归通常通过普通最小二乘法（*ordinary least squares*，*OLS*）进行估计 *[78]*。

在工具变量（IV）模型中，内生变量 X 是解释变量，也是自变量，Y 是结果变量，也是因变量。

IV 概念定义 4 (内生性（Endogeneity）). 内生性指的是统计模型中的解释变量 X 与误差项（*error term*）相关 *[329]*，或者与未观察到的混杂变量相关，从而导致内生性 *[224]*。

工具变量（Z）用于分离出内生变量 X 中对于结果变量 Y 来说是外生（exogenous）的变化成分 (外生变量的影响), 并且该工具变量还必须与任何未观察混杂变量独立 [224], 这里的未观察变量可以是不可观察变量或者遗漏变量。

20.2 工具变量可以用来干什么？

这一节，我先从评价学的角度浅显易懂地解释工具变量（IV）可以用来干什么。

IV 可以认为结构因果模型的一个应用特例。在观察研究里，研究原因变量 X 对结果变量 Y 因果影响时，某些混杂变量 U 无法观察或者本身就是遗漏变量，无法对它进行控制，但它同时影响 X 和 Y，成为了混杂变量。同时也没有适合的中介变量可以用来识别 X 对 Y 的因果影响。

如果存在一种情况：除了 U 的影响外，X 和 Y 的变化完全被动地由外生的工具变量集驱使，而这些工具变量不直接影响 Y，同时和 U 独立。在这种情况下，可以借助工具变量分离出 X 对 Y 的因果影响。

20.3 问题定义

IV（工具变量）是一种广泛应用于计量经济学与因果推断的统计方法，可用于解决内生性问题：存在未观察的混杂因素，无法对它进行控制，因果影响无法通过标准识别策略（如后门准则或前门准则）识别，当引入的工具变量 IV 满足相关性与外生性条件时，可以实现对因果影响的识别 [329, 237, 266, 224, 11]。

问题定义 12 (IV 问题定义)**.** *IV 问题可以表述为：“存在内生性问题时，如何估计变量 X 对变量 Y 的因果影响，形式上表达为 $Y = \alpha + \beta X + \epsilon$。” 当 X 与误差项 ϵ 相关时，就会产生内生性问题，从而导致 β 的估计存在偏差和不一致性 [266]。*

IV 问题的输入包括内生变量 X、结果 Y 以及建议的工具变量 Z 的观察数据。一个有效的解决方案依赖于 Z 是否满足三个关键约束，这些约束在第 20.4节中有详细介绍。当这些约束得到满足时，IV 过程的输出就是因果参数 β 的一致估计值。

20.4 基本假设

IV 假说 1 (相关性（Relevance）)**.** *一个工具变量 Z 必须与内生变量 X 相关（correlated)。也就是说，工具变量必须对 X 有非零影响（工具变量必须显著地影响 X） [11]。*

IV 假说 2 (外生性（Exogeneity）)**.** *一个工具变量 Z 必须与误差项不相关。这确保了 Z 对 X 引起的变化是外生的，并且不会受到与内生变量 X 相同的偏差影响 [11]。*

IV 假说 3 (排除限制（Exclusion Restriction）)**.** *一个工具变量 Z 必须只能通过影响 X 来影响结果 Y，并且不应对 Y 有直接影响。如果 Z 直接影响 Y，那么因果影响的估计是有偏的（biased） [11]。*

20.5 基本原理

IV（工具变量）为解决因果推断中的内生性问题提供了一个强有力的解决方案。内生性问题出现在回归模型中，当解释变量 X 与误差项相关时，估计其对结果 Y 的因果影响存在偏差。内生性的主要来源是存在不可观察的混杂变量 U 同时影响 X 和 Y，产生了虚假的关联 [224]。由于 U 无法被控制，无法适用后门准则，同时也不存在一个适合的中介变量可以利用前门准则来识别因果影响。此时 IV 方法通过引入一个工具变量 Z 来

解决这一问题，Z 作为外生的变化源。如图 20.1所示，核心思想是使用一个与内生变量 X 相关但与混杂变量 U 独立的工具变量 Z，且 Z 仅通过 X 影响结果 Y [224]。

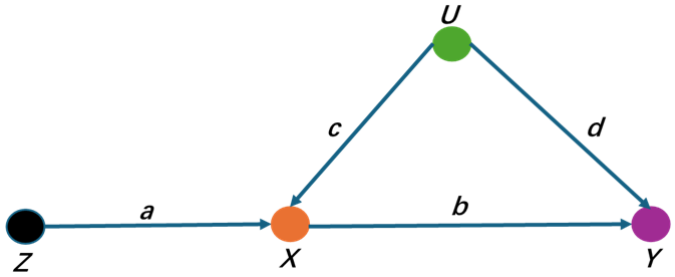


图 20.1: IV 框架：工具变量 Z 分离出内生变量 X 中对结果变量 Y 外生（exogenous）的变化成分（外生变量的影响），从而解决内生性问题。Z 与混杂变量 U 独立 [224]。

IV 方法的最常见实现是*两阶段最小二乘法*，该过程包括两个阶段 [11]：

- *第一阶段*：将内生变量 X 对工具变量 Z（可以是一组变量）进行回归，以获得其预测值 $\hat{X}$。
- *第二阶段*：将结果变量 Y 对预测值 $\hat{X}$ 进行回归。这一步分离出由外生工具变量驱动的 X 的部分，从而消除由混杂因素 U 引起的偏差。

从约翰 · 斯诺（John Snow）对霍乱传播的调查到菲利普 · 赖特（Philip Wright）对供给曲线的分析，历史上的应用表明，选择恰当的工具变量（IV）能够揭示因果关系 [224]。总之，工具变量框架为非实验环境（non-experimental settings）下的因果影响估计提供了一种严谨的方法。然而，其有效性在很大程度上依赖于满足三个核心假设：工具变量的相关性、外生性以及排除约束。这些假设若被违背，可能导致误导性结论，因此需要谨慎的设计与检验。

20.6　方法学

IV 的基本模型是一个回归模型。假设我们有一个回归模型，其中 Y 是因变量，X 是内生解释变量，Z 是工具变量。我们希望估计的基本模型是：

$$Y = \beta_0 + \beta_1 X + \epsilon. \tag{20.1}$$

- Y：结果变量。
- X：内生解释变量，原因变量。
- β_0：截距项，表示当 $X = 0$ 时 Y 的期望值。
- β_1：X 的系数，表示 X 每增加一个单位对 Y 的影响。
- ϵ：误差项，表示所有未被 X 解释的因素或随机误差。

然而，由于内生性，即 X 与误差项 ϵ 相关，我们无法直接估计 β_1。

为了解决这个问题，我们引入一个工具变量 Z，它与 X 相关，但与 ϵ 不相关。IV 的标准估计程序是*两阶段法*。

阶段 1：对内生解释变量进行回归 [11] 在第一阶段，我们使用 Z 来预测 X，即进行如下回归：

$$X = \pi_0 + \pi_1 Z + u. \tag{20.2}$$

- X：内生解释变量。
- Z：工具变量（IV）。
- π_0：截距项，表示当 $Z = 0$ 时 X 的期望值。
- π_1：工具变量 Z 的系数，表示 Z 增加时对 X 的影响。
- u：误差项，表示 X 中未被 Z 解释的部分。

这个回归的目标是通过工具变量 Z 来解释 X 的变动。

阶段 2：利用预测值进行回归 [11] 在第二阶段，我们将 X 替换为第一阶段回归得到的预测值 $\hat{X}$（它是 Z 的线性组合），然后估计 Y 和 $\hat{X}$ 之间的关系：

$$Y = \beta_0 + \beta_1 \hat{X} + \nu. \tag{20.3}$$

- Y: 结果变量。
- $\hat{X}$: 从第一阶段回归得到的 X 的预测值，它是 Z 的线性组合。
- β_0: 截距项，表示当 $\hat{X} = 0$ 时 Y 的预期值。
- β_1: 新的回归系数，表示在第二阶段回归中，预测的 $\hat{X}$ 对 Y 的影响。
- ν: 新的误差项，表示第二阶段回归中的随机误差成分，它与 ϵ 无关。

由于 $\hat{X}$ 是使用 Z 进行预测的，并且 Z 与 ϵ 不相关，因此这个第二阶段回归提供了 β_1 的一致估计。

20.7 IV 和 SCM 的关系

IV 可以认为是 SCM 的一个特例。本节由王晨曦撰写。

在 SCM 的前门准则中，基础的因果模型如图 20.2所示：

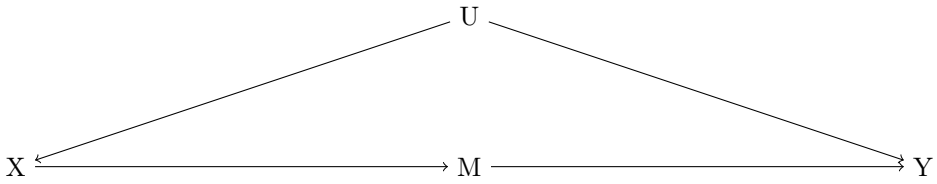


图 20.2: 结构因果模型的前门准则对应的基础因果模型。

需要研究原因变量 X 对结果变量 Y 的因果影响，对应于因果路径 $X \to M \to Y$。在这个因果模型下，原因变量 X 对结果变量 Y 的影响还包括由混杂因子 U 引入的混杂影响，对应于后门路径 $X \leftarrow U \to Y$。这个混杂因子 U 无法进行控制，也就是说后门路径 $X \leftarrow U \to Y$ 无法通过后门准则阻断。此时，可以通过分别求解 X 对 M 的因果影响和 M 对 Y 的因果影响来计算 X 对 Y 的因果影响。求解 X 对 M 的因果影响时，后门路径 $X \leftarrow U \to Y \leftarrow M$ 被其中的碰撞 $U \to Y \leftarrow M$ 所阻断。求解 M 对 Y 的因果影响时，后门路径 $M \leftarrow X \leftarrow U \to Y$ 可以通过控制变量 X 来阻断。

工具变量（Instrumental Variables，IV）建立局部因果识别模型。它通过工具变量识别内生变量对结果变量的影响，工具变量要求满足相关性、外生性和排除限制假设。满足这些条件时，IV 可以识别因果影响。但它不是完整的因果模型。对于 IV，基础的因果模型如图 20.3所示：

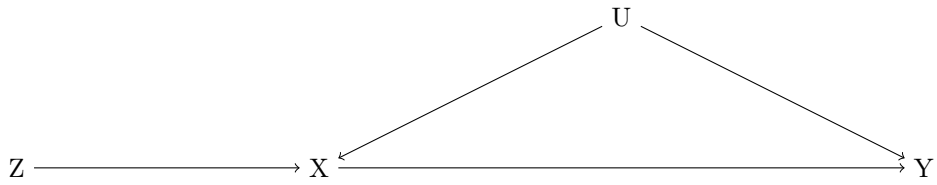


图 20.3: 工具变量对应的基础因果模型。

需要研究原因变量 X 对结果变量 Y 的因果影响，对应于因果路径 $X \to Y$。在这个因果模型下，原因变量 X 对结果变量 Y 的影响还包括由混杂因子 U 引入的混杂影响，对应于后门路径 $X \leftarrow U \to Y$。这个混杂因子 U 无法进行控制，也就是说后门路径 $X \leftarrow U \to Y$ 无法通过后门准则阻断。此时，可以通过引入一个不受混杂因子影响的工具变量 Z，这个工具变量需要满足对 X 有直接因果影响且对 Y 的因果影响是通过 X 间接影响的（即 Z 与 Y 之间没有有向边）。引入工具变量 Z 后，可以分别求解 Z 对 Y 的影响和 Z 对 X 的因果影响来计算 X 对 Y 的因果影响。求解 Z 对 Y 的因果影响时，路径 $Z \to X \leftarrow U \to Y$ 被其中的碰撞 $Z \to X \leftarrow U$ 所阻断。求解 Z 对 X 的因果影响时，除前门路径 $Z \to X$ 外节点 Z 和 X 之间无其他路径，可以直接求解。

20.8 示例：推断苹果产地对购买价格的影响

这个例子展示了如何应用工具变量（IV）方法来解决由于“未观察的混杂因素”导致的内生性问题。我们旨在估计“苹果产地”对购买价格的因果影响。内生性的主要来源是“未观察的消费者认知”。例如，消费者认为本地苹果更新鲜或质量更高。这种认知同时影

响了苹果是否会标定为本地生产以及消费者愿意支付的价格。由于我们无法直接测量这种认知，因此我们使用工具变量来隔离产地的外生性变化。

变量

- PP：购买价格（每公斤，单位：¥）。
- OR：苹果产地（二元变量，0 = 其他地区，1 = 本地产地）。这是内生变量。
- ZZ：农业补贴政策（二元变量，0 = 无政策，1 = 有政策）。这是工具变量。
- U：未观测到的消费者对本地产品质量的认知。

假设数据（Hypothetical Data）

我们收集了 1000 批苹果的样本数据。

- PP：苹果的购买价格在 5.00 至 15.00 ¥/公斤之间。
- OR：苹果产地为 0（其他地区）或 1（本地）。
- ZZ：农业补贴政策为 0（无政策）或 1（有政策）。

假设以下样本数据：500 个样本来自其他地区的苹果（$OR = 0$）；500 个样本来自本地种植的苹果（$OR = 1$）；60% 的样本（600 批次）产自实施农业补贴政策的地区（$ZZ = 1$）。

问题设定：未观察的混杂因素

我们希望估计如下的因果模型：

$$PP = \beta_0 + \beta_1 OR + \epsilon$$

其中误差项 ϵ 包含了未观察的消费者认知 U。这会产生内生性问题，因为 U 同时影响 OR（苹果产地）和 PP（购买价格），导致 OR 和 ϵ 之间存在相关性。例如，具有较强正向认知（U）的地区会有更多的本地苹果（OR）和更高的价格（PP），从而产生伪相关。如果对 PP 与 OR 进行普通最小二乘回归，将得到 β_1 的有偏估计。

使用工具变量解决方法

我们引入一个工具变量 ZZ，即农业补贴政策（政府通过降低本地农业生产成本以鼓励本地农业）。ZZ 的有效性基于两个假设：

- 相关性：政策 ZZ 直接影响苹果产地 OR。补贴降低了本地生产成本，使得本地苹果的生产和销售更为可能（$OR = 1$）。因此，ZZ 与 OR 相关。
- 排除限制：政策 ZZ 对购买价格 PP 的影响，必须仅通过其对产地 OR 的苹果供应量的作用而产生。政策本身与未观察到的消费者认知 U 无关，并且除了通过改变产地构成之外，不会直接影响市场价格。

工具变量提供了 OR 的外生性变化来源，这些变化独立于混杂因素 U。

两阶段最小二乘法（Two-Stage Least Squares）实现

第一阶段：隔离苹果产地的外生性变化
在第一阶段，我们将内生变量 OR 对工具变量 ZZ 进行回归。这将隔离由外生政策驱动的苹果产地变化：

$$OR_i = \pi_0 + \pi_1 ZZ_i + u_i,$$

其中 OR_i 表示苹果产地是否为本地（1 为本地，0 为非本地），ZZ_i 表示是否实施了政策（1 为有政策，0 为无政策）。该回归使用普通最小二乘法估计，得到拟合值 $\hat{OR}_i$，它代表由政策驱动的苹果产地变化，因此对于未观察到的混杂因素 ϵ 是外生的。
第二阶段：估计购买价格的因果影响
在两阶段最小二乘法程序的第二阶段，我们将购买价格 PP_i 对第一阶段预测的（外生的）苹果产地值 $\hat{OR}_i$ 进行回归：

$$PP_i = \beta_0 + \beta_1 \hat{OR}_i + \nu_i.$$

该阶段得出的 β_1 是一致的因果参数估计，因为它仅使用工具变量 ZZ 引发的 OR 的外生性变化。
示范性数值结果
假设第一阶段的普通最小二乘回归结果为：

$$OR_i = 0.20 + 0.65 * ZZ_i + u_i$$

这表明，该政策使苹果产自本地的概率提高了 65 个百分点。
将 $\hat{OR}_i$ 代入第二阶段回归得到两阶段最小二乘法估计：

$$PP_i = 8.50 + 2.10 * \hat{OR}_i + \nu_i.$$

这意味着本地种植苹果的因果影响是使价格平均增加 *2.10* ¥/公斤。

结论与解读

两阶段最小二乘法程序提供了苹果产地对购买价格的因果影响的一致估计。第一阶段隔离了由于农业政策导致的苹果产地的外生性变化，第二阶段则修正了内生性问题并得出了稳定的参数估计。在工具变量的相关性和排除限制假设下，系数 $\hat{\beta}_1 = 2.10$ 表明本地种植的苹果价格比非本地苹果高出约 *2.10* ¥/公斤。
然而，需要注意的是，这一估计代表了局部平均处理效应（Local Average Treatment Effect,LATE），适用于那些产地受农业政策影响的苹果。这意味着估计的影响仅反映了在农业政策改变了的苹果产地的地区中，本地种植和非本地种植苹果之间的价格差异。换句话说，这一结果并非对所有本地苹果的普遍估计，而是专门针对那些生产决策受到政策影响的苹果。因此，估计值 $\hat{\beta}_1 = 2.10$ 反映了在政策影响生产决策的地区，本地种植对苹果价格的因果影响。

上述例子展示了如何使用工具变量解决内生性问题，并估计苹果产地对购买价格的因果影响，同时也验证了农业政策对购买价格的间接影响。通过该方法我们也能够就农业政策对苹果定价等市场结果的影响得出更可靠的结论。

20.9 局限性

尽管 IV 方法具有强大的功能，但也存在一些重要的局限性 [329, 237, 266, 224, 11]。首先，找到一个有效的工具变量是困难的，因为在实践中，很少有变量能够同时满足相关性、排除限制和外生性。其次，弱工具变量仅与感兴趣的变量（X）弱相关，这可能导致有偏的估计和较大的标准误差（standard error）。第三，IV 估计通常只识别受工具变量影响的亚群（subpopulation）的局部平均处理效应（LATE），这限制了结果的广泛性。最后，IV 估计对模型设定较为敏感，错误的模型或不恰当的协变量可能会削弱估计的有效性。

20.10 总结

工具变量（IV）方法在计量经济学中广泛用于解决内生性问题，并在无法进行控制实验时估计因果关系。通过使用工具变量——即与内生解释变量相关但与误差项不相关的变量——IV 方法帮助隔离解释变量对结果变量的因果影响。然而，IV 方法也存在一些局限性，包括找到有效工具变量的挑战、对模型设定的敏感性，以及弱工具变量可能导致有偏估计的风险。

第二十一章

不同方法的总结与对比

本章总结不同的评价方法学，并从多种的视角对这些方法进行比较。非常遗憾的，目前仍然缺少系统的评价方法的验证工作，正如第 十二章指出的，这是评价学的基本问题之一。

21.1 基础假说的差异

评价学 是一个综合性的方法。它的基础假说是因果律：除非对象及其影响机制具有真随机性，否则对象的影响是确定的。

实验设计（DoE） 是一个模型方法。它基础假说主要包括 线性、效应可加性、随机化、独立同分布、同方差。

结构方程模型（SEM） 是一个模型方法。它的基础假说主要包括测量误差独立性、多元正态分布、线性、结构残差与外生变量独立性和模型可识别性。

随机对照实验（RCTs） 是基于随机化机制的物理方法。它的基础假说主要包括稳定单元干预值假设、可忽略性假设和一致性假设。稳定单元干预值假设要求一个单元的潜在结果不受其他单元干预分配的影响，并且不存在隐藏的干预版本；可忽略性假设认为干预分配与潜在结果相互独立；一致性假设要求观察到的结果等于该单元在实际接受干预水平下的潜在结果。

潜在结果（PO） 是结合随机化机制和模型的方法。它的基础假说主要包括稳定单元干预值假设、可忽略性假设和正值性假设。稳定单元干预值假设要求潜在结果只依赖于单元自身接受的干预状态，并且干预被明确定义；可忽略性假设要求在控制给定协变量 (X) 后，干预分配 (T) 与潜在结果 ((Y(0),Y(1))) 条件独立，这实际是一个隐含的模型；正值性假设要求每类协变量取值下都存在接受干预和不接受干预的可能性。

准实验 是观察的方法。它的基础假说是隐含的：基于自然分组或者有意构造的比较组，通过观察结果的差异可以推测干预的影响。

结构因果模型（SCM） 是一种模型的方法。它的基础假说主要包括因果有向无环图假说、结构方程假说和因果 Markov 条件假说。DAG 假说认为变量之间的因果关系可以用有向无环图表示；结构方程假说认为每个变量由其直接原因和外生扰动共同决定；因果马尔可夫条件根据图结构表达条件独立关系。

工具变量（IV） 是一种模型的方法。基础假说主要包括相关性、外生性和排除限制。相关性要求工具变量 (Z) 与内生变量 (X) 相关；外生性要求 (Z) 与影响结果变量 (Y) 的误差项或未观察混杂因素不相关；排除限制要求 (Z) 只能通过 (X) 影响 (Y)，而不能直接影响 (Y)。

21.2 原因和影响定义的差异

评价学 以存在算子来定义原因对象的影响。影响可以是对象结构、属性、机制和行为的变化，对象的消失或新对象的产生。

实验设计（Design of Experiments，DoE） DoE 把因子水平的改变作为原因侧的变化，把相应的平均响应变化定义为**因子效应（*factor effect*）** [197]。因子对应评价学里对象的属性。

结构方程模型（SEM） SEM 用**路径系数（*path coefficient*）**以及直接、间接和总路径影响表示变量之间的关系 [330, 32]。变量可以表示对象的属性，但变量之间的路径并不自动等同于对象之间的因果影响；只有当模型方向具有因果含义且相应假设成立时，路径系数才能解释为因果影响。

随机对照实验（RCTs） RCT 通过随机分配形成干预组与控制组，干预是原因，干预组与控制组间的结果差异表示干预影响 [295]。

潜在结果（PO） PO 把影响定义为同一单元在不同干预状态下潜在结果的差异 [248, 253]，干预即原因。

准实验（Quasi-experiments） 将不同比较组下观察结果的差异定义为影响。比较组的差异即为原因。

结构因果模型（SCM）和工具变量（IV） SCM 用 do 算子来定义影响，被操作的变量即为原因。工具变量可以认为是结构因果模型的一个特例。

21.3 方法学上的差异

评价学通用方法 （第 十二章阐述）是一种“条条道路通铁林寨”的方法。评价学通用方法的首要原则是命题或者模型需要通过检验，在满足这个原则的前提下，它接受猜想、观察、实验或者推理得出的任何命题或者模型。

评价学通用方法实际上是三个方法的综合体：基于现实世界系统的观察方法，基于自包含系统的实验方法，基于对象因果模型的模型方法。

基于现实世界系统的方法可以结合观察与建模。基于自包含系统的方法，可以采纳三种基础的物理实验方法。评价学可以使用随机化，但显式地承认它的限制，无法获得一个具体的原因对象对一个具体的影响对象的影响。评价学显式地使用隔离方法，在物理手段上减少混杂对象的影响。在实验过程中，可以使用存在算子，来模拟原因对象的存在与否。隔离与存在算子是互补的物理实验方法，前者针对混杂对象，后者针对原因对象。

基于自包含系统的方法也可以与建模的方法相结合。

对象因果模型是基于模型的方法，它可以在特定的条件下简化为其他方法，例如结构因果模型。do-演算作为一个虚拟化的隔离手段，也可以用于对象因果模型。评价学也可以采用其他虚拟化的实验方法。

DoE 是一个模型的方法，需要接受检验，但并没有显式地纳入检验的方法。DoE 是一个线性模型，可以用于推测现象背后的原因。

SEM 是一个模型的方法，需要接受检验，但并没有显式地纳入检验的方法。SEM 通常是一个线性模型，建立的是隐藏变量与可观察变量以及隐藏变量之间关系的模型。它可以用于推测现象背后的原因。

RCTs 是一个基于随机化机制的实验方法。

PO 是结合随机化机制和模型的方法。在随机化机制方面，它是对 RCTs 的泛化。

PO 的无混杂性假说实际上是一个模型，它假设能识别影响潜在结果和分配机制的协变量，这实际上建立了一个模型，也就是哪些协变量影响分配机制，哪些协变量影响潜在结果。在这个条件下，对这些协变量进行控制，能保证潜在结果与分配机制独立，从而在观察研究中模拟随机实验。

我认为如果能确定影响潜在结果的协变量，直接建立模型，是更为直接的方法。

PO 的倾向性得分更难以在物理世界实施。它的基本思路是说控制某个协变量集合，从而能固定原因对象分配到比较组的概率，在这个前提下，分配机制与潜在结果独立。

PO 并没有引入检验机制来检验模型的正确性。

SCM 和 IV 都是模型的方法，通过DAG显式表达，它们实际上是虚拟化的隔离方法，对混杂变量的影响传递路径进行控制。它的困难在于，在物理世界难以实施，另外，完全依赖于模型的正确性。

准实验 是一种经验性的观察方法。它充分利用自然分组的特性，或者有意地构造比较组，可以认为它经验性的使用了存在算子，在不同的分组里删除或者插入某个对象（类似于存在算子），用这种方法近似地获得原因对象的影响。

21.4 模型的差异

评价学的对象因果模型 是一个建立在对象基础上的通用的模型方法。对象是具有属性、结构、功能和行为的一类实体。

其他方法均不是建立在对象基础之上。

结构因果模型（Structural Causal Model，SCM） 建立明确的变量因果模型。它用有向无环图表示因果结构。它用结构方程描述变量之间的因果影响。它用 do 算子表达干预。因此，SCM 是典型的因果建模方法。

工具变量(IV) 是类似于结构因果模型的方法。它的基本组成是工具变量、内生变量和结果变量等，用于解决因果推断中的内生性问题。

实验设计（Design of Experiments，DoE） 建立的是因子和响应变量之间的线性模型，能够计算不同因子对响应变量的方差贡献程度以及统计显著性。DoE 本身不描述因果结构，也不解释影响机制。

结构方程模型（SEM） 建立的是隐藏变量与可观察变量以及隐藏变量之间关系的模型，这些关系通常采用线性方程表示。

随机对照实验（RCTs） 通过随机化构造干预组和对照组。它没有显式地构造模型。

潜在结果（PO）理论 是一种建立在潜在结果与反事实比较基础上的因果推断框架。它的基本组成包括潜在结果：$Y(1)$ 与 $Y(0)$、干预分配以及由潜在结果 $Y(1)$ 与 $Y(0)$ 二者之差定义的因果影响。在观察性研究中，PO 通过控制协变量来估计因果影响：既可采用基于模型的方法，如结果回归、倾向得分建模等；也可采用不依赖显式结果模型的方法，如匹配、分层和逆概率加权等。PO 本身不显式建立变量间的结构方程或有向无环图，也不直接描述因果生成机制。

21.5 是否是蛮力法

概念定义 70 (蛮力法)**.** *蛮力法是指单纯通过扩大实验或计算规模、穷举状态组合或实施随机分配——“以量换质”——以获得预期结果的方法。*

在评价领域，蛮力法的核心逻辑是：只要实验覆盖足够完整、样本足够大、即使存在不可观察或者未知的混杂因素，不同求索条件具有统计意义上的等价性。

DoE 与 RCTs 是蛮力法的典型代表，二者共同特征是实施成本高，但识别路径相对直接；相比较之下，SEM、PO、准实验、SCM 与 IV 更强调模型结构、识别条件与场景适应性的论证，实验成本更低，但对理论设定和假说成立的正确性更敏感。

实验设计（Design of Experiments，DoE） 属于典型的蛮力法。全因子设计要求对因子水平组合开展完整实验，才能分别估计主效应与交互效应；若缺少关键因子，模型方程便无法完整求解，效应识别会出现缺口（gap）或混杂。部分因子设计虽可减少实验次数，但本质上仍依赖预先设计的实验覆盖度来换取可估计性，因此实验次数、重复次数和计算开销依然较高。

结构方程模型（Structural Equation Model，SEM） 不属于蛮力法。它通常不通过新增大规模实验或穷举因子组合获得结论，而是利用既有观察数据，通过建立模型进行估计。SEM 的成本主要在模型设定、参数识别、数据收集、模型的拟合与验证，而不是实验执行规模。

随机对照实验（Randomized Controlled Trials，RCTs） 属于典型的蛮力法。它基于随机分配机制使干预组与对照组具有统计意义的等价性，以样本规模消除影响对象的差异，而不必逐一测量并控制所有混杂因素。换言之，RCT 用“随机化”换取统计意义上的因果识别，因此通常需要较大样本量、完整随访和较高实施成本。

潜在结果（Potential Outcome，PO） 本身不是蛮力法，而是一种因果影响定义与估计框架。它通过反事实结果 $Y(1)$ 与 $Y(0)$ 定义平均干预影响，再借助随机化、协变量调整、匹配或倾向得分实现干预影响的识别。在随机实验中，PO 与 RCTs 结合，此时“蛮力”来自随机试验设计；在观察性研究中，PO 更强调可忽略性、协变量平衡和识别条件，而不是扩大实验规模。

准实验（Quasi-experiments） 不属于蛮力法。它通常不依赖大规模随机化，而是利用自然分组或者有意地构造比较组。准实验的成本更多体现在比较组的设计与实现，以及自然分组的混杂控制上，而不是单纯扩大样本或穷举实验组合。相较 RCTs，它实施门槛较低，但也更依赖研究者对时间趋势、选择机制和不可观察混杂的论证。

结构因果模型（Structural Causal Model，SCM） 不属于蛮力法。它并不依赖新增实验估计因果影响，而是充分利用专业知识，定义因果图，识别后门路径、前门路径与可调整变量集合，并用 Do 演算或后门调整公式计算 $\mathbb{P}(Y \mid \mathbf{do}(X))$。因此，SCM 的核心在于因果结构是否被正确设定，而不是实验次数本身。

工具变量（Instrumental Variables，IV） 不属于蛮力法。它通过寻找一个只影响干预、不直接影响结果的变量，在存在不能观察混杂时识别因果影响，因而更依赖“好工具”的存在，而不是大规模实验的覆盖度。

21.6 是否是确定性方法

评价学概念定义 16 (确定性方法，随机方法)**.** *确定性方法是指对影响结果的因素进行显式的建模，通过明确定义的方程直接计算影响的方法；随机方法则是依赖于实现真随机性的机制通过随机分配实现求索条件等价性的方法。*

DoE、SEM 与 SCM 属于确定性方法：DoE 通过因子设计与方差分析将响应分解为主效应、交互效应和误差；SEM 通过模型刻画变量关系；SCM 则进一步以 DAG、结构方程和 Do 演算对因果影响进行建模和计算。相比较之下，RCTs 与 PO 属于随机方法：前者依赖随机化构造组间可比较性，后者以潜在结果定义因果影响，并借助随机化或协变量调整实现因果识别。

21.7　是否适用于观察

评价学方法、SCM 以及 PO 都认为在基于观察也可以获得有效的结论。准实验则强调近似实验方法的有效性有限。其他方法均不认可观察方法。

实验设计（DoE） 不认可观察方法。DoE 要求研究者主动设定因子水平、安排实验组合并实施重复测量，通过可控实验系统分离主效应、交互效应和随机误差。若仅依赖自然状态下形成的观察数据，而未按设计开展实验，便无法完整识别因子效应，也难以进行规范的方差分析。因此，DoE 本质上属于实验驱动的方法，而不是观察驱动的方法。

结构方程模型（SEM） 不将观察直接视为充分的因果证据来源。SEM 虽然通常利用既有数据进行估计，但其重点在于通过事先设定的模型表达变量关系，而不是把自然观察到的相关关系直接等同于因果影响。SEM 常用于分析观察数据，但并不把观察到的相关关系直接等同于因果影响。

随机对照实验（RCTs） 不认可观察方法。RCT 的核心在于研究者通过随机分配构造干预组与对照组，使组间比较能够反映干预影响，而不是事后根据已有样本进行被动比较。观察性数据即使样本量很大，也不能替代随机化所带来的组间可比较性。因此，RCT 明确属于实验方法，而不是观察方法。

潜在结果（PO） 认可在特定条件下可基于观察获得有效结论。PO 框架虽然以潜在结果定义因果影响，但并不要求所有研究都必须实施随机实验。在随机实验中，PO 等同于 RCTs；在观察性研究中，只要满足可忽略性等识别条件，便可通过协变量调整、匹配或倾向得分识别因果影响。因此，PO 既可用于实验研究，也可用于观察研究，是少数明确承认观察方法有效的基础框架之一。

准实验（Quasi-experiments） 是一种经验性的观察方法。

结构因果模型（SCM） 认可在识别条件成立时基于观察可获得有效结论。SCM 并不必然要求新的随机实验，而是先根据专业知识建立因果图，再判断在现有观察数据中，某一因果影响是否能够被识别。若满足后门准则或其他识别条件，便可通过调整、分层和加权从观察数据中估计 $\mathbb{P}(Y \mid \mathbf{do}(X))$。因此，SCM 属于少数明确承认观察数据也可用于因果推断的方法，但前提是因果结构和识别规则必须事先明确。

工具变量（IV） IV 认为当存在未观察混杂时，不能将干预组与对照组之间的直接观察比较视为识别因果的充分证据。因为在存在未观察混杂时，简单观察到的相关关系可能是有偏的。IV 要求引入一个只影响干预、不直接影响结果的外部变量，通过两阶段估计恢复因果影响。因此，IV 并不否定观察数据本身，而是否定“仅凭观察比较即可得到因果结论”。

21.8 从评价学视角下看不同的方法

评价学通用方法是一个泛化（generalized）的方法。

整体上来说，评价学通用方法是一种基于对象与对象关系建模的方法。它显式地区分对象的属性、结构、机制和行为，清晰地定义了原因对象、影响对象、混杂对象概念。

它显式地引入了模型的检验，将猜想、观察、实验和推理都认为是合理的方法，其有效性取决于能否通过检验。

其他方法可以认为是评价学通用方法的特例，尽管有些方法的合理性在评价学视角下可能会被质疑。

实验设计 是对对象的量之间的关系进行线性建模的方法。量是对象可计数的属性。

结构模型方程 是对隐藏变量与观察变量关系进行建模的方法。

结构因果模型和工具变量 是对外生和内生的可观察变量之间的关系进行因果建模的方法，它们通过一些规则隔离混杂变量的干扰。

以上四种方法均可以在特定情况下，作为基于对象因果模型的建模方法的简化表示，例如将对象简化为对象的一个或多个量，但有其局限性。

RCTs 可以认为是评价学基于自包含系统的实验方法的一个特例，借助于随机化机制，但没有显式地使用隔离和存在算子等机制。

PO 可以认为随机化机制与模型方法的结合。它和 RCTs 一样，可以视为基于自包含系统的实验方法的特例。在模型方面，它的无混杂假说实际上是对对象的量之间的关系进行建模，确定哪些协变量能决定潜在结果及分配机制。它可以应用于基于自包含系统的评价学方法。

准实验 是一种可以用于现实世界系统的观察方法。

21.9 总结

本章从假说、原因和影响的定义、方法学、模型、是否是蛮力法、是否是确定性方法、是否适用观察这七个角度总结了不同评价方法的差异。同时，从评价学的视角讨论了其他方法与评价学通用方法之间的关系，它们是后者的特例或者补充。

第二十二章

评价学未来工作

本章从不同的角度讨论评价学的未来工作。

22.1 基础理论

独立性、随机性和计算复杂性之间的关系 第一个需要回答的基础理论是独立性、随机性和计算复杂性之间的关系是什么?

我提出了一个有关独立性、随机性和计算复杂性关系的猜想。

猜想 9 (独立性、随机性和计算复杂性关系猜想)**.** *独立性、随机性和计算复杂性是对象或衍生对象不同的量，它们的求索结果的空间具有相似性和等价性。*

我非正式地介绍这个猜想的含义。假如一个主体正在研究对象 A 和对象 B 以及它们的关系，主体同时想建立对象 A 和 B 之间关系的模型。

考虑两个极端的情况。在最简单的情况下，对象 A 决定着对象 B 的行为。从主体来看，A 和 B 的随机性在同一个水平上，B 的独立性由 A 的独立性决定。A 和 B 关系的模型的计算复杂性也由 A 决定。

在最复杂的情况下，对象 A 和对象 B 具有原因独立性，则 A 和 B 的随机性由其自身决定；A 和 B 关系的模型的计算复杂性也由 A 和 B 的随机性共同决定。

因此从求索的角度来看。独立性、随机性和计算复杂性是对象或者衍生对象（对象关系）不同的量，求索结果的空间具有相似性和等价性。

已有 RCTs 和 PO 理论对独立性和随机性这两个概念的关系进行了讨论。例如，在研究两个原因对象 A_1 和 A_2 对一组影响对象集合的影响时，将 A_i 随机分配给干预组和控制组，解释为分配机制与潜在结果独立。

但什么是独立性? 本文提出的因果独立性和它的关系是什么? 进一步，独立性与真随机性、表面随机性、自由意志之间的关系是什么? 它们的物理和数学模型是什么?

我的猜想给出了初步的回答。我认为它们是对象或者衍生对象不同的量，其求索结果的空间具有相似性和等价性。

随机化、隔离和存在算子之间的关系 有没有可能进一步简化为一个或者两个算子，在这个基础发展通用的代数演算理论?

22.2 评价学基础理论

第 十二章提出了评价学的基础问题。解决这些问题，当然能完善评价学基础理论。
但评价学还有一些更基础的理论需要研究。

当对象及对象影响机制具有真随机性时的评价学理论 当前版本的评价学理论基于因果律公理，要求对象及影响机制不具有真随机性。当这个条件不成立时，需要发展新的评价学理论乃至整个求索科学体系。

适合评价学通用方法的建模语言 DAG 适合 SCM 的建模，但显然无法适应评价学通用方法的建模需求。

存在算子的演算 需要发展存在算子和虚拟存在算子的演算理论。

表面随机性的不确定性分析框架 需要为表面随机性建立模型。

基于现实世界系统的观察方法 相对于观察，实验受限，不仅存在伦理性问题，也存在成本问题。然而，现实世界本身存在丰富的差异性，如何利用这种天然的求索条件的差异性，值得进一步发展超越经验的方法。

22.3 总结

本章提出了评价学的未来工作。