上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

因果关系识别

叶南阳

第一节 前门调整

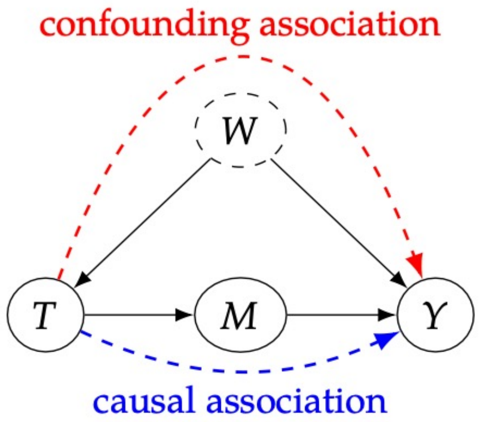
知识回顾

- 上节课中，我们了解到后门调整可以识别因果效应：

- 后门调整：假设变量集合W满足后门准则 ——固定 W 作为条件可以阻塞 T 和 Y 之间的所有后门路径；W不包含 T 的所有子孙结点，那么 T 对 Y 的因果效应可由以下公式计算：

$$P(y|do(t)) = \sum_w P(y|t, w)P(w)$$

- 然而对于含有未观测变量的情况，我们无法block所有后门路径，那么有没有办法可以不满足后门准则也能识别因果效应？

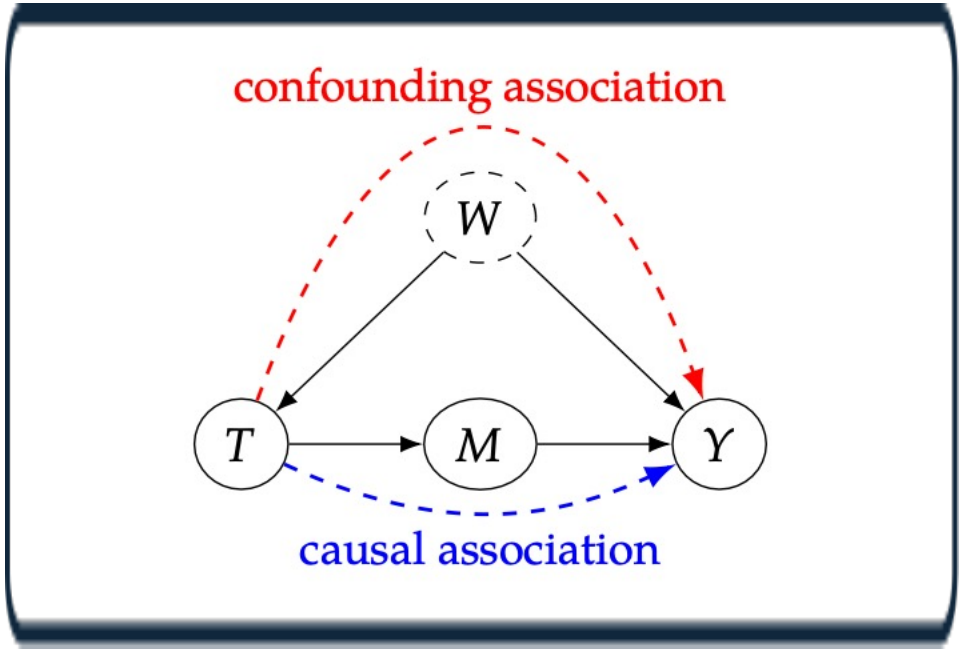


W无法被观察到，如何识别因果效应

知识回顾

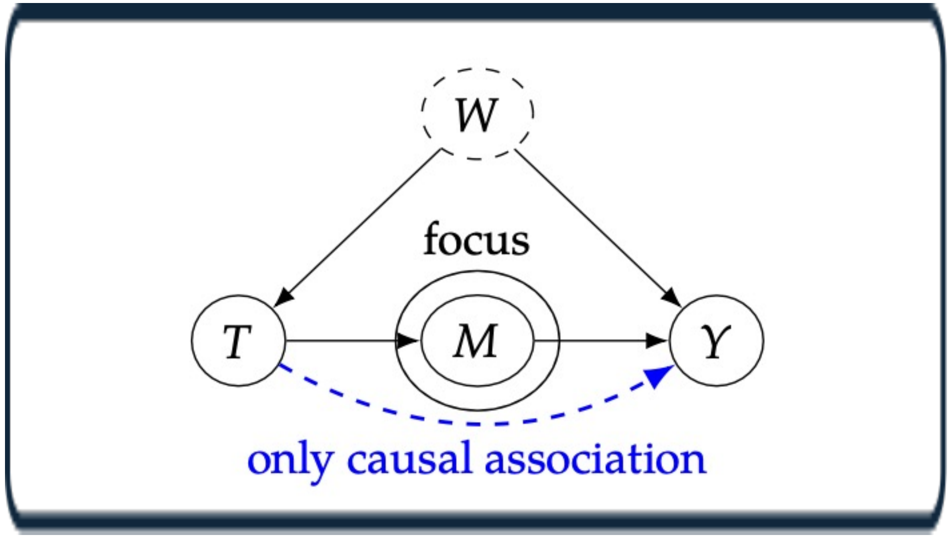
- 例如，右侧因果图中，含有未观测变量 W，我们不能阻塞通过 W 的后门路径和沿此路径的混杂关联。

- 事实证明，使用前门调整可以识别右图的因果效应。我们首先来介绍前门准则和前门调整；然后在第二节介绍 do-calculus，考虑更一般的情况；最后在第三节介绍因果图的可识别性。

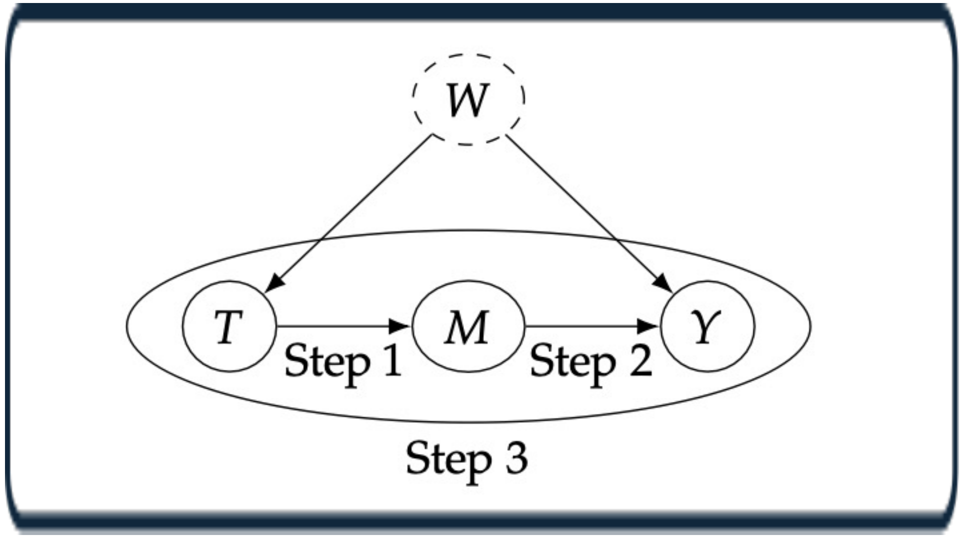


前门调整

- 通过引入中间变量 M， 只关注经过M的因果关联 (沿 T-M-Y 这条有向路径流动的关联)。



- 使用三个步骤来识别 T 对 Y 的因果效应:
- (1) 识别 T 对 M 的因果效应;
- (2) 识别 M 对 Y 的因果效应;
- (3) 结合(1) 和 (2)，识别 T 对 Y 的因果效应。

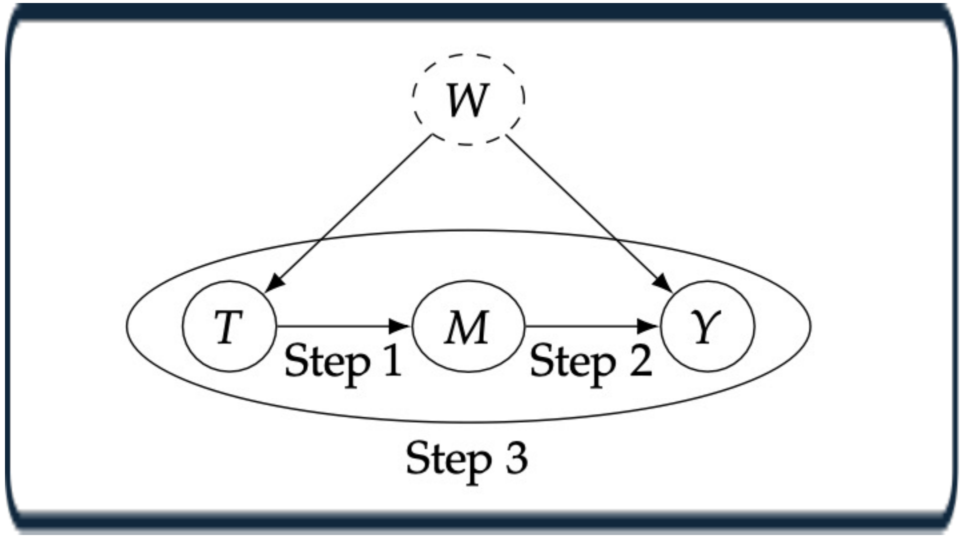


前门调整

- (1) 识别 T 对 M 的因果效应 $P(m|do(t))$:

- 因为 Y 是 W-M 的对撞机（collider），所以Y 阻塞 T-M 的后门路径。即经过 T-M 的唯一关联是因果关联。因此根据后门调整可得：

$$P(m|do(t)) = P(m|t)$$

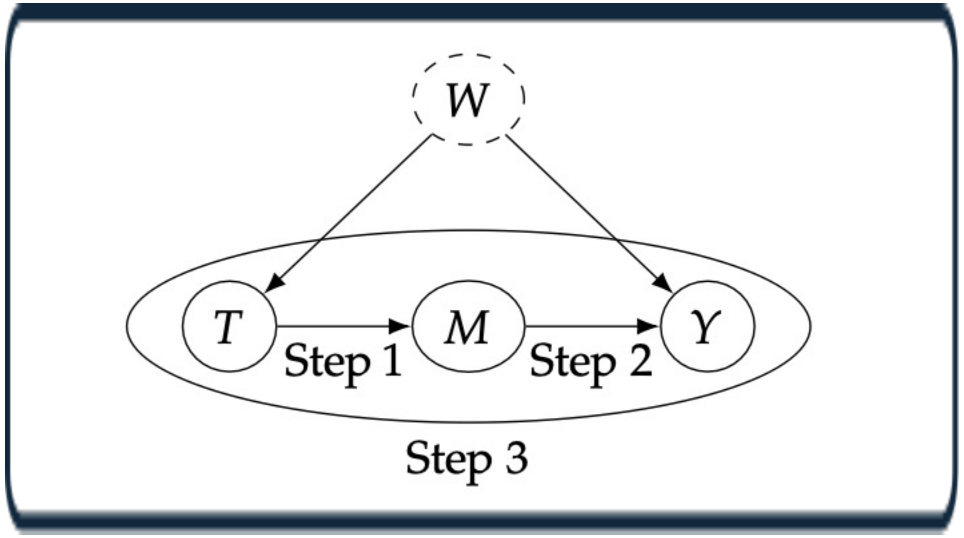


前门调整

- (2) 识别 M 对 Y 的因果效应 $P(y|do(m))$:

- 因为 T 阻塞了 $M \leftarrow T \leftarrow W \rightarrow Y$ 的后门路径，运用后门调整，得到：

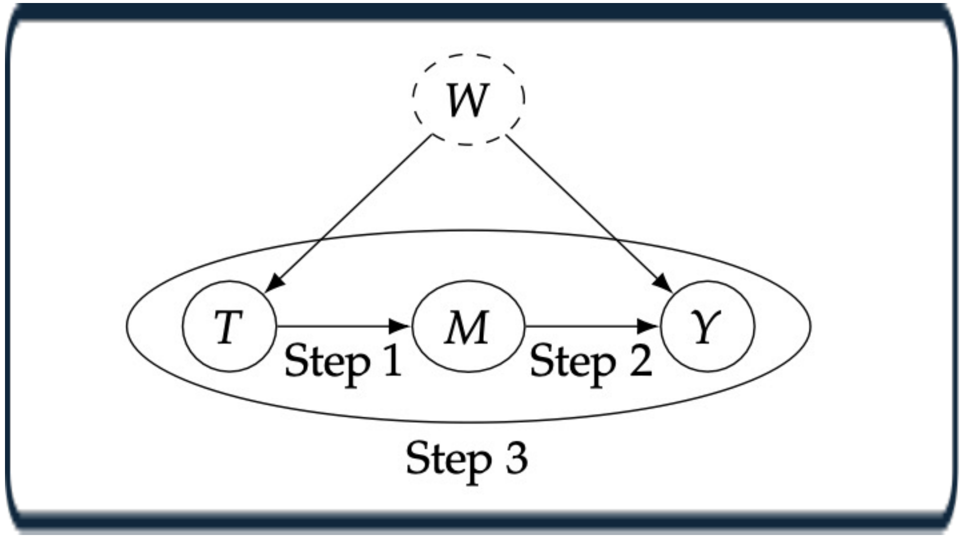
$$P(y|do(m)) = \sum_t P(y \mid m, t)P(t)$$



- (3) 结合(1) 和 (2)，识别 T 对 Y 的因果效应 $P(y \mid do(t))$：

- 由(1)，我们得到了 T 对 M 的影响，由(2)，我们得到 M 对 Y 的影响，结合两者，我们可以得到 T 对 Y 的影响：

$$P(y \mid do(t)) = \sum_m P(m \mid do(t)) P(y \mid do(m))$$



- 其中，右边第一个因子对应将 T 设置为 t 时，M 的结果。第二个因子对应将 M 设置为 m 下，Y 的结果。这里 $P(M \mid \mathrm{do}(t))$是干预分布，对随机变量分布$P(M \mid \mathrm{do}(t))$的所有可能的取值 $P(M=m \mid \mathrm{do}(t))$求和。

前门调整

- 前门准则：

我们说变量集 M 关于 T 和 Y 满足前门准则，若满足以下条件：

- 1.M 完全中介了 T 和 Y，即所有从T到Y的因果路径都经过M；
- 2.从 T 到 M 没有未被阻断的后门路径；
- 3.所有从M到Y的后门路径被 T阻断。

- 定理：前门调整 （**Frontdoor Adjustment**）

如果 （T, M, Y）满足前门准则，并且满足非负性假设，则：

$$P(y \mid do(t)) = \sum_m P(m \mid t) \sum_{t'} P(y \mid m, t')P(t')$$

前门调整

- 我们基于截断式因子分析和贝叶斯网络给出前门调整定理的 证明，证明：

$$P(w,t,m,y) = P(w)P(t \mid w)P(m \mid t)P(y \mid w,m) \quad \text{(贝叶斯网络)}$$

$$P(w,m,y \mid do(t)) = P(w)P(m \mid t)P(y \mid w,m) \quad \text{(截断式因子分解，移除}T\text{)}$$

$$\sum_m \sum_w P(w,m,y \mid do(t)) = \sum_m \sum_w P(w)P(m \mid t)P(y \mid w,m) \quad \text{(对m和w求和)}$$

$$P(y \mid do(t)) = \sum_m P(m \mid t) \sum_w P(y \mid w,m)P(w)$$

我们的目标是把未观测变量 w 消去，只留下 t，m，y，在上式中，如果能将$P(w)$转化为$P(w|\mathrm{m})$，就能刚好消去 w，由 w 与 m 关于 T 条件独立（Chain结构）可得 $P(w|\mathrm{t}) = P(w|\mathrm{m},\mathrm{t})$，于是上式可化为：

$$P(y \mid do(t)) = \sum_m P(m \mid t) \sum_w P(y \mid w,m)P(w)$$

$$= \sum_m P(m \mid t) \sum_w P(y \mid w,m) \sum_{t'} P(w \mid t')P(t')$$

$$= \sum_m P(m \mid t) \sum_w P(y \mid w,m) \sum_{t'} P(w \mid t',m)P(t')$$

$$= \sum_m P(m \mid t) \sum_{t'} P(t') \sum_w P(y \mid w,m)P(w \mid t',m)$$

前门调整

- 为了结合 $P(y \mid w, m)$ 和 $P(w \mid t', m)$，还需要对 $P(y \mid w, m)$ 引入条件 t'

$$P(y \mid w, m) = P(y \mid w, m, t')$$ （因为 *W* 和 *M* 已经 d-分离 *T* 和 *Y*）

- 因此，

$$P(y \mid do(t)) = \sum_m P(m \mid t) \sum_{t'} P(t') \sum_w P(y \mid w, m) P(w \mid t', m)$$

$$= \sum_m P(m \mid t) \sum_{t'} P(t') \sum_w P(y \mid w, t', m) P(w \mid t', m)$$

$$= \sum_m P(m \mid t) \sum_{t'} P(t') \sum_w P(y, w \mid t', m)$$

$$= \sum_m P(m \mid t) \sum_{t'} P(t') P(y \mid t', m)$$

- 将 (1) 和 (2) 的结论 $P(m|do(t)) = P(m|t)$和 $P(y \mid do(m)) = \sum_t P(y \mid m, t) P(t)$ 代入上式得到：

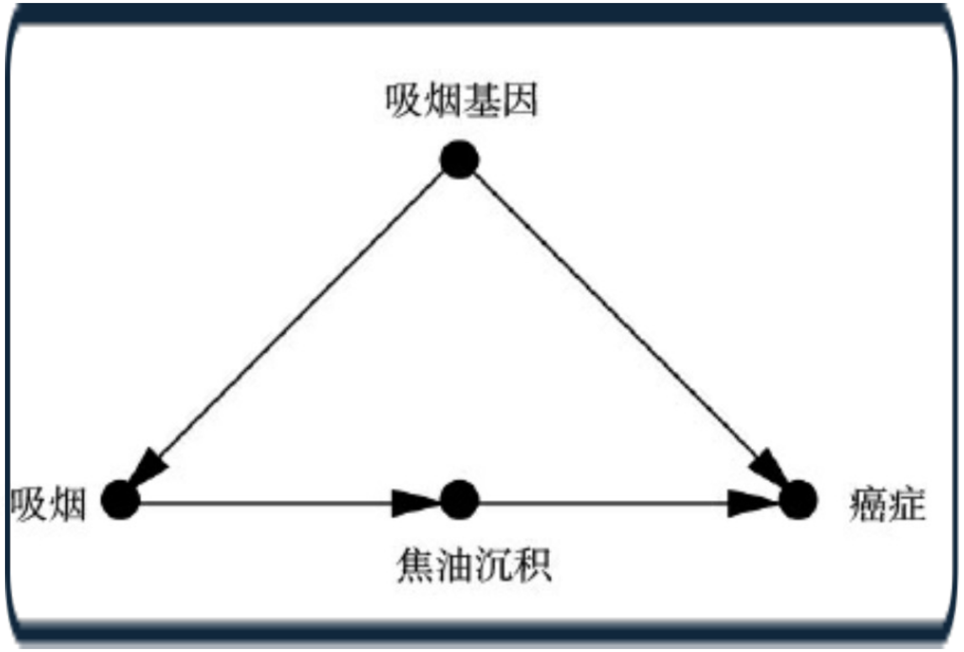
$$P(y \mid do(t)) = \sum_m P(m \mid t) \sum_{t'} P(y \mid m, t') P(t') = \sum_m P(y|do(m)) P(m|do(t))$$

证毕。

前门调整实例

- **经典的吸烟的因果效应悖论：**

- 早在20世纪50年代，焦油沉积的形成就被怀疑是肺癌发展的一个可能的中间阶段。



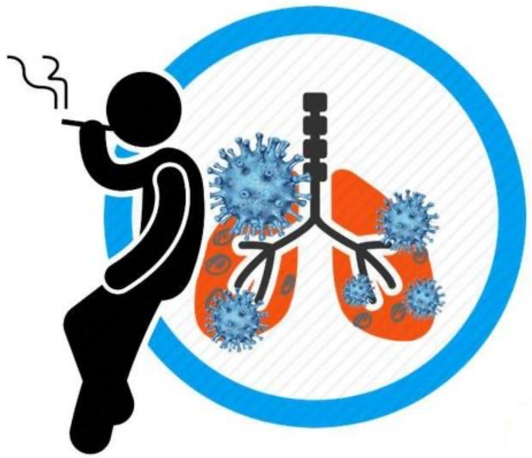
- 假设研究人员可以测量吸烟者肺部的焦油沉积量，吸烟基因是吸烟行为和肺癌的混杂因子。

- 上图包含了两个非常重要的假设：
- (1) **吸烟基因对焦油沉积物的形成没有影响，焦油沉积只与香烟烟雾的物理作用有关。**这一假设以“吸烟基因”和“焦油沉积”之间没有箭头来表明；不过它并不能排除与“吸烟基因”无关的其他随机因素对“焦油沉积”的影响。
- (2) **只有通过焦油沉积的积累，“吸烟”才会导致“癌症”。**因此，我们假设从“吸烟”到“癌症”之间没有直接箭头，也没有其他间接路径。

前门调整实例

	Group Type	$P(x, z)$ Group Size (% of Population)	$P(Y = 1 \mid x, z)$ % of Cancer Cases in Group
$X = 0, Z = 0$	Nonsmokers, No tar	47.5	10
$X = 1, Z = 0$	Smokers, No tar	2.5	90
$X = 0, Z = 1$	Nonsmokers, Tar	2.5	5
$X = 1, Z = 1$	Smokers, Tar	47.5	85

- 从数据中来看，似乎吸烟对肺癌有显著影响，但烟草公司会从不同的角度争辩，从而给出不同的答案——若只看非吸烟者，体内有焦油的人群患癌率从10%降到了5%；若只看吸烟者，体内有焦油的人群患癌率从90%降到了85%，以此来争辩焦油有防护作用。

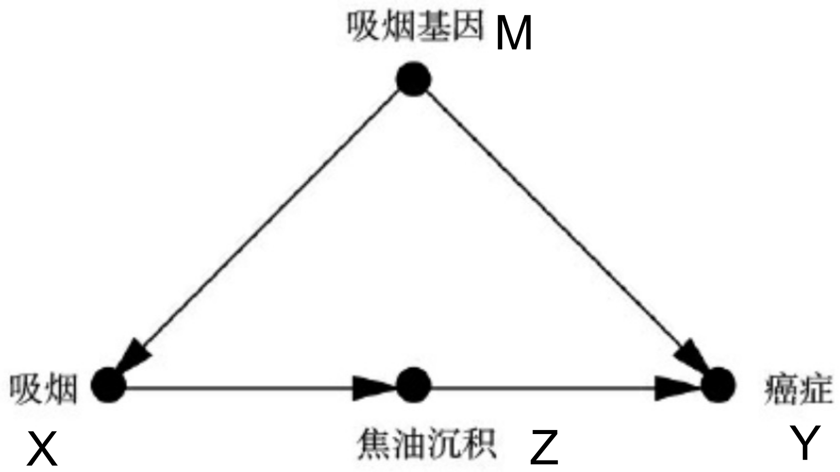


前门调整——练习

	Group Type	$P(x, z)$ Group Size (% of Population)	$P(Y = 1 \mid x, z)$ % of Cancer Cases in Group
$X = 0, Z = 0$	Nonsmokers, No tar	47.5	10
$X = 1, Z = 0$	Smokers, No tar	2.5	90
$X = 0, Z = 1$	Nonsmokers, Tar	2.5	5
$X = 1, Z = 1$	Smokers, Tar	47.5	85

运用前门调整 $P(y \mid do(x)) = \sum_z P(z \mid x) \sum_{x'} P(y \mid z, x') P(x')$:

计算P(癌症|do(吸烟)) 和P(癌症|do(不吸烟))



前门调整——练习答案

$P(y=1 \mid do(x=0))$

$= \sum_z P(z \mid x=0) \sum_{x'} P(y=1 \mid z, x') P(x')$

$= P(z=0 \mid x=0) \sum_{x'} P(y=1 \mid z=0, x') P(x') + P(z=1 \mid x=0) \sum_{x'} P(y=1 \mid z=1, x') P(x')$

$= P(z=0 \mid x=0)[P(y=1 \mid z=0, x'=0) P(x'=0) + P(y=1 \mid z=0, x'=1) P(x'=1)] +$
$P(z=1 \mid x=0)\,[P(y=1 \mid z=1, x'=0) P(x'=0) + P(y=1 \mid z=1, x'=1) P(x'=1)]$

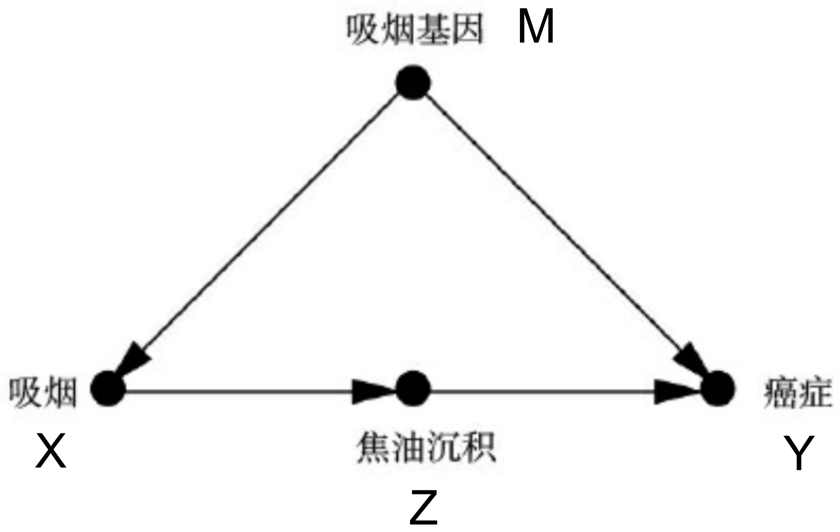
$= 0.95 * (0.10 * 0.5 + 0.9 * 0.5) + 0.05 * (0.05 * 0.5 + 0.85 * 0.5) = 0.4975$

同理： $P(y=1 \mid do(x=1)) = 0.4525$

前门调整实例——练习答案

$$P(Y=1|do(X=0))=0.4975, P(Y=1|do(X=1))=0.4525$$

这个结论基于前面提到的两个假设时成立，**如果认为吸烟基因对焦油沉积有因果效应，则前门调整失效。**



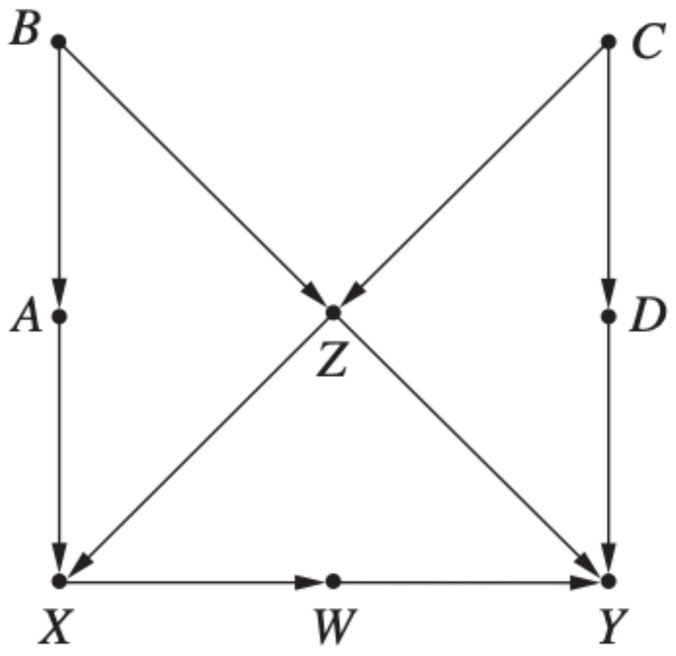
因果图的正确性，决定了因果推理的正确性

后门调整——练习

考虑右侧的因果图：

（a）列出所有满足后门准则以确定X对Y的因果关系的变量集。

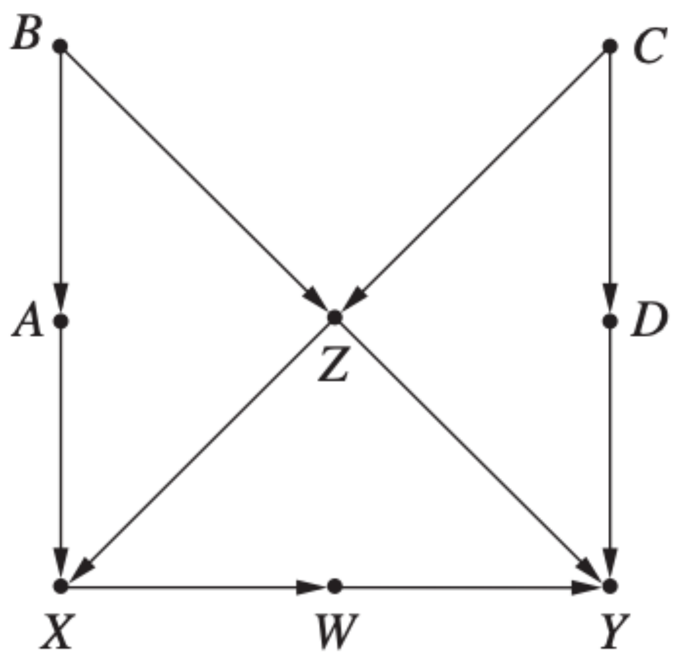
（b）列出所有满足确定 X 对 Y 的因果影响后门准则的最小变量集，(即从集合中删除其中任何一组变量，它将不再满足准则)。



后门调整——练习答案

（a）列出所有满足后门准则以确定X对Y的因果关系的变量集。

1. *Sets of 2 nodes:* $\{Z, A\}, \{Z, B\}, \{Z, C\}, \{Z, D\}$
2. *Sets of 3 nodes:* $\{Z, A, B\}, \{Z, A, C\}, \{Z, A, D\}, \{Z, B, C\}, \{Z, B, D\}, \{Z, C, D\}$
3. *Sets of 4 nodes:* $\{Z, A, B, C\}, \{Z, A, B, D\}, \{Z, A, C, D\}, \{Z, B, C, D\}$
4. *Sets of 5 nodes:* $\{Z, A, B, C, D\}$



（b）列出所有满足确定 X 对 Y 的因果影响后门准则的最小变量集，
(即从集合中删除其中任何一组变量，它将不再满足准则)。

$\{Z, A\}, \{Z, B\}, \{Z, C\}, \{Z, D\}$

总结

- 后门调整公式和后门标准就像硬币的正反面。后门标准告诉我们哪些变量集可以用来去除数据中的混杂。
- 遗憾的是，如果我们因缺乏必要的数据而无法阻断某条后门路径，后门调整公式就会完全失灵。然后我们通过引入中间变量，使用前门调整，只关注经过中间变量的因果关联，来完成因果推断。

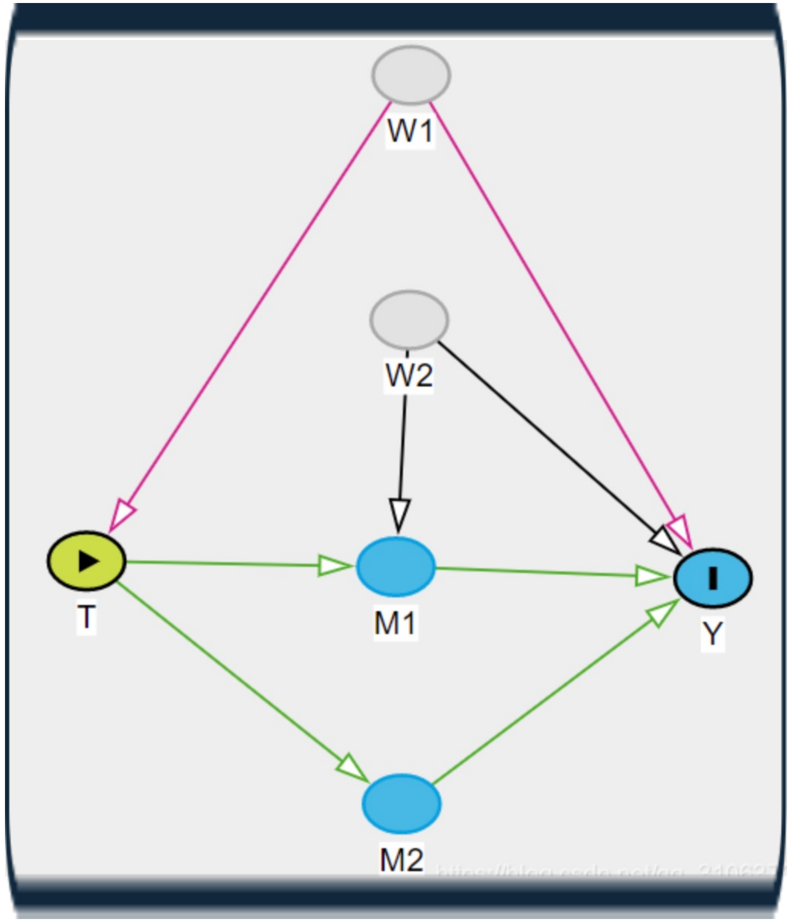
第二节 do-calculus

知识回顾

- 上节中，我们了解了通过前门调整识别因果效应。那么当 *T* 到 *Y* 之间既不满足前门准则，也不满足后门准则的时候，怎么识别因果效应呢?

- 如右图，前后门准则都无法使用。

- 因为 *W1* 和 *W2* 都是无法观测的，无论使用前门准则还是后门准则，*T* 对于 *Y* 的因果效应总是无法正确估计。对于这种情况，我们可以使用unconfounded children criterion或者 do-calculus 来正确估计。本节首先介绍 do-calculus 理论。



知识回顾

- 事实上，前门调整公式和后门调整公式的最终目标是根据 P(Y | X，A，B，Z，...) 此类不涉及do算子的数据估算干预的效果，即 P(Y | do(X))。

- 如果我们成功消除了计算过程中的do概率，就可以利用观测数据来估计因果效应，这样一来，我们就从因果关系之梯的第一层级（相关性估计）踏上了第二层级（因果性估计）。

符号定义

- 在介绍 do-算子 之前，我们先给出符号说明：
- $G_{\overline{X}}$：移除 G 中所有指向 *X* 的边，即 *X* 没有双亲结点；
- $G_{\underline{X}}$：G 中所有被 *X* 指向的边，即 *X* 没有孩子结点；
- $G_{\overline{z(w)}}$：移除 Z 指向 W 的所有边，即 Z 不是 W 的双亲结点;
- $\perp\!\!\!\perp_G$：d分离

do-calculus 准则

- do-calculus 三公理：

给定因果图*G*， 关联分布*P*， 不相交的随机变量 *Z, T, W, Y*，则有以下准则：

（1） 增添或删除观察 （ Ignoring observations）

$$P(y \mid do(t), z, w) = P(y \mid do(t), w) \quad \text{if} \quad Y \perp\!\!\!\perp_{G_{\overline{T}}} Z \mid T, W$$

（2） 干预与观察交换 （Action / Observation exchange）

$$P(y \mid do(t), do(z), w) = P(y \mid do(t), z, w) \quad \text{if} \quad Y \perp\!\!\!\perp_{G_{\overline{T}\underline{Z}}} Z \mid T, W$$

（3） 增添或删除干预 （Ignoring actions）

$$P(y \mid do(t), do(z), w) = P(y \mid do(t), w) \quad \text{if} \quad Y \perp\!\!\!\perp_{G_{\overline{T}\,\overline{z(w)}}} Z \mid T, W$$

do-calculus 准则

- do-calculus 三公理的直观解释：

（1）增添或删除观察 （ Ignoring observations）

- 如果移除干预 $do(t)$，得到：

$$P(y \mid z, w) = P(y \mid w) \quad \text{if} \quad Y \perp\!\!\!\perp_G Z \mid W$$

- 上式正是 d-分离 （d-separation） 的定义，即准则（1）是 d-分离 到干预分布的泛化。

（2）干预与观察交换 （Action / Observation exchange)

- 如果移除干预 $do(t)$，得到：

$$P(y \mid do(z), w) = P(y \mid z, w) \quad \text{if} \quad Y \perp\!\!\!\perp_{G_{\underline{Z}}} Z \mid W$$

- 上式是后门调整，即准则（2）是后门调整到干预分布的泛化。

do-calculus 准则

- do-calculus 三公理的直观解释：

（3）增添或删除干预 （Ignoring actions）

- 如果移除干预 $do(t)$，得到：

$$P(y \mid do(z), w) = P(y \mid w) \quad \text{if} \quad Y \perp\!\!\!\perp_{G_{\overline{Z(W)}}} Z \mid W$$

- 观察条件 $Y \perp\!\!\!\perp_{G_{\overline{Z(W)}}} Z \mid W$， $G_{\overline{Z(W)}}$ 表示移除 Z 指向 W 的所有边。当 Z 没有指向 *W* 的边时，给定 *W*，*Y* 和 Z 是 d-分离的，那么这种情况下干预 $do(z)$ 对 y 没有影响（Modularity）。

do-calculus 准则

- do-calculus 三条基本准则的完备性：

- 是否可能存在不能用 do-calculus 准则来识别的可识别因果图?

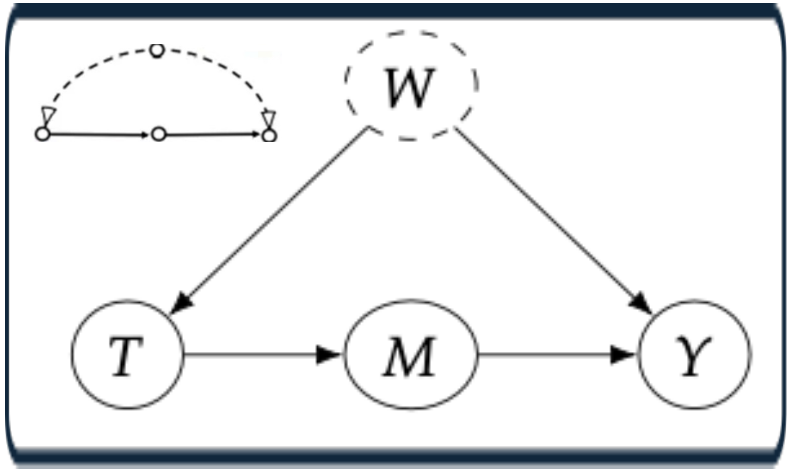
- 答案是否定的，Shpitser和Pearl 等证明了 do-calculus 是完备的，即 do-calculus 的三个准则可以识别所有可识别的因果推断。

- do-calculus三条基本准则证明：

  参考文献：https://arxiv.org/pdf/1305.5506.pdf

do-calculus 实例

- 上节我们基于截断式因子分析和贝叶斯网络给出前门调整的证明，本节我们基于do-calculus 的三个准则给出 $P(y \mid do(t)) = \sum_m P(m \mid t) \sum_{t'} P(y \mid m, t')P(t')$ 的证明，并结合吸烟与肺癌的例子，以此解释三公理：



$P(y \mid do(t)) = \sum_m P(y \mid do(t), m)P(m \mid do(t))$ （全概率公式）

$= \sum_m P(y \mid do(t), m)P(m \mid t)$ （公理（2）干预观察交换 ）

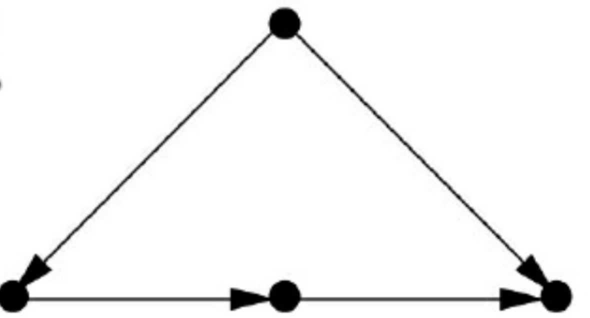
$= \sum_m P(y \mid do(t), do(m))P(m \mid t)$ （公理（2）干预观察交换 ）

$= \sum_m P(y \mid do(m))P(m \mid t)$ （公理（3）删除干预）

$= \sum_m P(m \mid t) \sum_{t'} P(y \mid do(m), t')P(t' \mid do(m))$ （全概率公式）

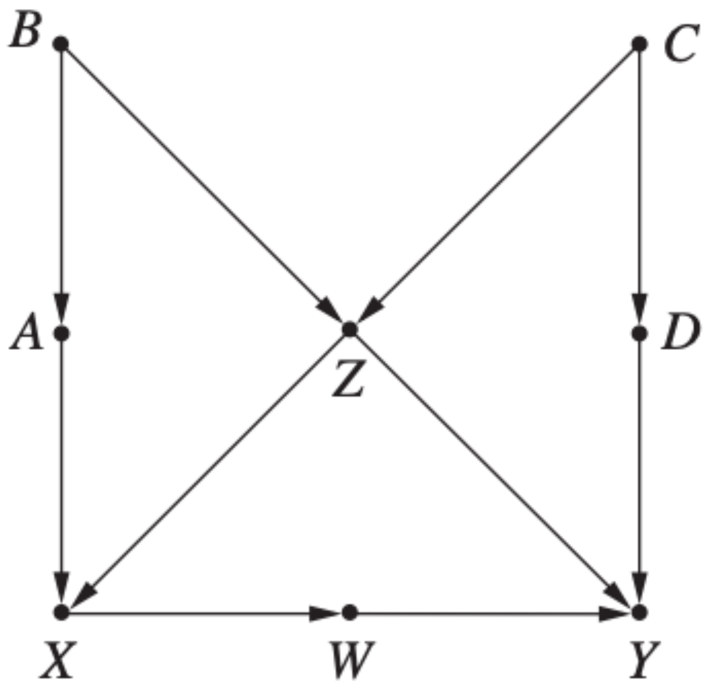
$= \sum_m P(m \mid t) \sum_{t'} P(y \mid m, t')P(t' \mid do(m))$ （公理（2）干预观察交换 ）

$= \sum_m P(m \mid t) \sum_{t'} P(y \mid m, t')P(t')$ （公理（3）删除干预）



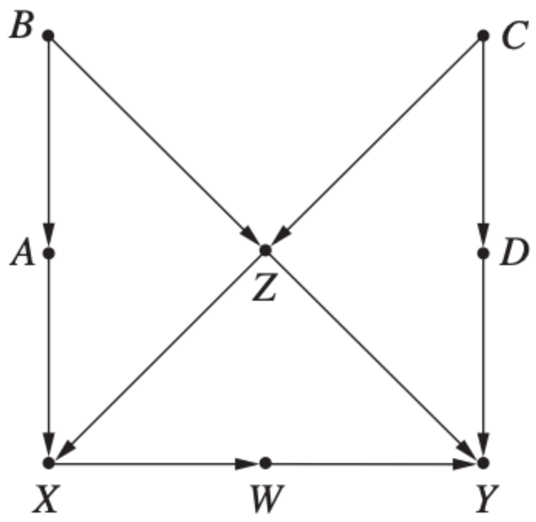
do-calculus 练习1

- 考虑如下因果图，假设只有 X，Y 和 另一个随机变量是可观测的，确定该随机变量，使得 X 对 Y 的因果效应是可识别的，并写出X 对 Y的因果效应。



do-calculus 练习1

- 考虑如下因果图，假设只有 X，Y 和 另一个随机变量是可观测的，确定该随机变量，使得 X 对 Y 的因果效应是可识别的，并写出 X 对 Y 的因果效应。



- W，满足前门准则，$P(y \mid do(x)) = \Sigma_w P(w \mid x) \Sigma_{x'} P(y \mid x', w) P(x')$

do-calculus 练习2

- 药店有两种药：一种定价 1 美元，另一种定价 10 美元。 10美元一瓶的日期新鲜，而1美元一瓶的已经放置了3 年。” 但数据显示，那些买便宜药的康复率要高得多。现对这两种药进行药性测试，测试结果表明 95%的陈年药品和 5%的新鲜药品失去了有效成分，但使用陈年药品的人群的康复率远高于使用新鲜药品人群的康复率。”

- 为了分析药品与康复的因果效应，现在提供一些调研数据。数据包括每位顾客购买的药的类型(陈年的或新鲜的)，药品中有效成分的浓度(高或低)，以及顾客是否从疾病中恢复。

do-calculus 练习2

- 数据完全证实了陈年药品康复率远高于新鲜药品。

	Cheap	Expensive	All Subjects
	400	400	800
	Low | High	Low | High	Low | High
	380 | 20	20 | 380	400 | 400
Recovery	323 | 18	1 | 38	324 | 56
Not Recovery	57 | 2	19 | 324	76 | 344

- 基于两个非常合理的假设，请你分析新鲜药品真的会给普通病人提供更大的康复机会吗？

(i) 顾客对所购买药品的化学含量(高或低)没有任何信息；他们的选择只受价格和日期的影响。

(ii) 药物对任何特定个体的作用只取决于其化学成分，而不取决于药品日期。

do-calculus 练习2

	Cheap	Expensive	All Subjects
	400	400	800
	Low | High	Low | High	Low | High
	380 | 20	20 | 380	400 | 400
Recovery	323 | 18	1 | 38	324 | 56
Not Recovery	57 | 2	19 | 324	76 | 344

- 具体考虑两个问题:
- (a) 确定问题的相关变量，并在因果图中描述这种情况。
- (b) 基于假设 (i) 和 (ii) 以及给出的数据，确定选择新鲜药物和陈年药物对康复率的因果效应。

do-calculus 练习2

（a）X, Y, Z 分别表示药物是否新鲜，药物化学成分浓度，是否恢复。因果图为：

$$\begin{tikzcd} X \arrow[r] \arrow[rr, bend left, dashed, leftrightarrow] & Z \arrow[r] & Y \end{tikzcd}$$

（b）比较两种药物的因果效应。令 $y1$表示恢复，$x1$表示选择新鲜药品，$z1$表示化学成分高。应用前门准则，类似吸烟与肺癌实例的分析，计算 X 对 Y 的因果效应，由计算结果，新鲜药品的恢复效果好于陈年药品。

do-calculus 练习2

$$
\begin{aligned}
P(y_1|do(x_1)) &= \Sigma_z P(z|x_1)\Sigma_{x'}P(y_1|x',z)P(x') \\
&= P(z_1|x_1)[P(y_1|x_1,z_1)P(x_1)+P(y_1|x_0,z_1)P(x_0)] \\
&\quad + P(z_0|x_1)[P(y_1|x_1,z_0)P(x_1)+P(y_1|x_0,z_0)P(x_0)] \\
&= 0.95*(0.1*0.5+0.9*0.5)+0.05*(0.05*0.5+0.85*0.5) \\
&= 0.4975 \\
P(y_1|do(x_0)) &= \Sigma_z P(z|x_0)\Sigma_{x'}P(y_1|x',z)P(x') \\
&= P(z_1|x_0)[P(y_1|x_1,z_1)P(x_1)+P(y_1|x_0,z_1)P(x_0)] \\
&\quad + P(z_0|x_0)[P(y_1|x_1,z_0)P(x_1)+P(y_1|x_0,z_0)P(x_0)] \\
&= 0.05*(0.1*0.5+0.9*0.5)+0.95*(0.05*0.5+0.85*0.5) \\
&= 0.4525
\end{aligned}
$$

$$
\begin{aligned}
ACE &= P(y_1|do(x_1)) - P(y_1|do(x_0)) \\
&= 0.045 > 0
\end{aligned}
$$

do-calculus 练习2

- 药物选择和恢复之间**未被观察到的混杂因素**可能会造成便宜的药物更有效的错觉。比如，**更节俭的顾客同时也有更健康的饮食习惯。**

- 为了消除这种错觉，我们必须评估ACE。首先，我们可以计算药物的有效成分对恢复的因果影响。设 p (p')表示随机选择的人服用活性成分含量高(低)的药物后的恢复概率，即:

$$p = P(Y = recover|do(Z = high))$$
$$p' = P(Y = recover|do(Z = low))$$

- 我们已通过计算得到了成分的因果效应，可以进行进一步分析：如果选择新鲜药品，则有5%的机会得到有效成分高的药，恢复率为 p，有95%的机会得到有效成分低的药，恢复率为 p' 。因此，选择陈年药品的平均恢复率是 $0.05p + 0.95p'$ 。相应新鲜药品的平均恢复率是 $0.95p + 0.05p'$ 。因此，两种选择之间的差别为： $0.05(p - p') + 0.95(p' - p) = 0.9(p' - p)$。 如果 $p' < p$，则新鲜药品更具优势。