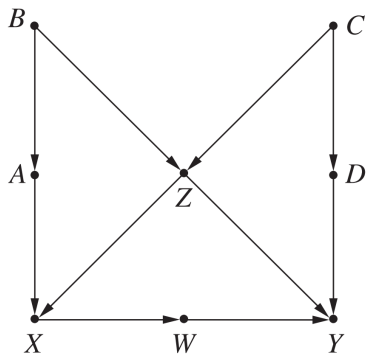
《统计因果推理入门》第三章习题

2021 年 5 月 4 日

1 思考题 3.3.1

考虑如下图所示的图模型



(a) 列出所有满足后门准则的变量集合，以确定 X 对 Y 的因果效应。

要满足后门准则，变量集合中不能包含 X 的后代结点，此图中包括结点 W 和 Y。

其次，变量集合要阻断 X 与 Y 之间的每条含有指向 X 的路径，本图中有四条路径。

首先观察变量 Z。如果 Z 不在变量集合中，则路径 $X \leftarrow Z \rightarrow Y$ 这条路径肯定无法被阻断，因此 Z 必须要在变量集合中。当 Z 在变量集合中时，X 和 Y 之间仅剩 $X \leftarrow A \leftarrow B \rightarrow Z \leftarrow C \rightarrow D \rightarrow Y$ 这一条后门路径，只需要保证 A、B、C、D 中至少一个结点在变量集合中即可满足后门准则。

综上所述，满足确定 X 到 Y 的因果效应的后门准则的变量集合有：

$$
\begin{aligned}
&\{Z, A\}, \{Z, B\}, \{Z, C\}, \{Z, D\}, \\
&\{Z, A, B\}, \{Z, A, C\}, \{Z, A, D\}, \{Z, B, C\}, \{Z, B, D\}, \{Z, C, D\}, \\
&\{Z, A, B, C\}, \{Z, A, B, D\}, \{Z, A, C, D\}, \{Z, B, C, D\}, \\
&\{Z, A, B, C, D\}
\end{aligned}
$$

使用 CEE 进行自动化运算

```go
package main

import (
   "fmt"
```

```
    "cee-lab/graph"
    "cee-lab/utils"
)

func PrintNodeSet(g *graph.Graph, nodeSet []int64) {
    fmt.Print("[")
        for idx, node := range nodeSet {
            if idx == 0 {
                fmt.Print(g.NodeName(node))
                continue
            }
            fmt.Print(", ", g.NodeName(node))
        }
    fmt.Println("]")
}

func main() {
    var g *graph.Graph = graph.NewGraph()
    g.LoadGraph("3.3.1.json")
    fmt.Println(g)
    set := []int64{g.NodeID("A"), g.NodeID("B"), g.NodeID("C"),
                g.NodeID("D"), g.NodeID("W"), g.NodeID("Z")}
    subset := utils.PowerSet(set)
    for _, z := range subset {
        if g.BackDoor(g.NodeID("X"), g.NodeID("Y"), z) {
            PrintNodeSet(g, z)
        }
    }
}
```

(b) 列出所有满足后门准则的最小变量集合，以确定 X 对 Y 的因果效应（即，对于任何一个变量集合，如果从集合中删除任何一个变量，它将不再满足后门准则）

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

(c) 为了识别 D 对 Y 的因果效应，列出需要测量的最小变量集合。如果要识别 $\{W, D\}$ 对 Y 的因果效应，重新回答此问题。

$$\{C\}, \{A, Z\}, \{B, Z\}, \{W, Z\}, \{X, Z\}$$

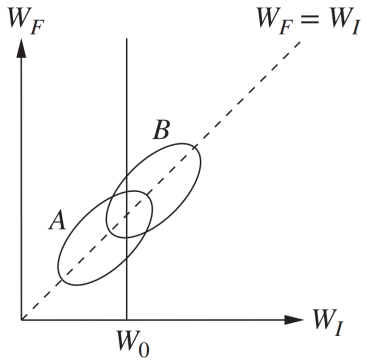
$$\{Z\}, \{C, X\}$$

2 思考题 3.3.2

某年年初，某寄宿学校为学生提供两种用餐方案：方案 A 和方案 B。在年初和年终会记录学生的体重。为了确定每个方案如何影响学生的体重，学校雇用了两位统计学家，但是他们奇怪地得出了不同的结论。第一个统计学家计算每个学生的体重在 6 月（W_I）和 9 月（W_F）之间的差异，发现每个方案的平均增重都为零。

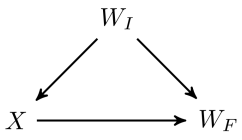
第二个统计学家将学生分为几组，每组对应一个相同的初始体重 W_I。他发现相对于初始体重，方案 B 的最终（平均）体重要高于方案 A 的最终（平均）体重。

因此，第一个统计学家认为饮食对增重没有影响，第二个统计学家则得出结论：有影响。



上图展示了一个能导致两个统计学家得出矛盾结论的数据集，其中横轴表示学生初始体重，纵轴表示学生最终体重。统计学家 1 检查了体重增量 $W_F - W_I$，对于每一个学生，该增量由图中点到 45° 线的竖直距离表示。确实，这两个组分别关于零增重线（$W_F - W_I$）对称，因此每种饮食方案的平均增重是零。另一方面，统计学家 2 比较了入校时有相同体重 W_0 的学生在方案 A 的最终体重和方案 B 的最终体重，正如图中竖线表明：沿着这条竖直线，方案 B 学生平均最终体重高于方案 A 学生平均最终体重。无论选取哪条竖直线 W_0，结果都一样。

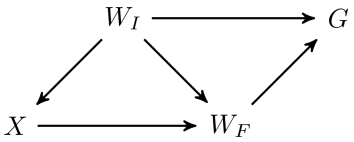
(a) 给出一个表示题目中情况的因果图。



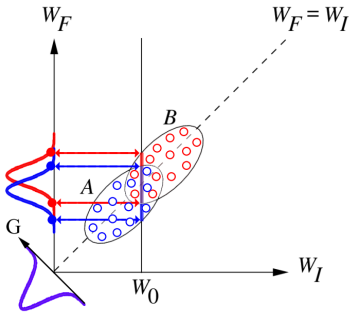
(b) 确定哪个统计学家是正确的。

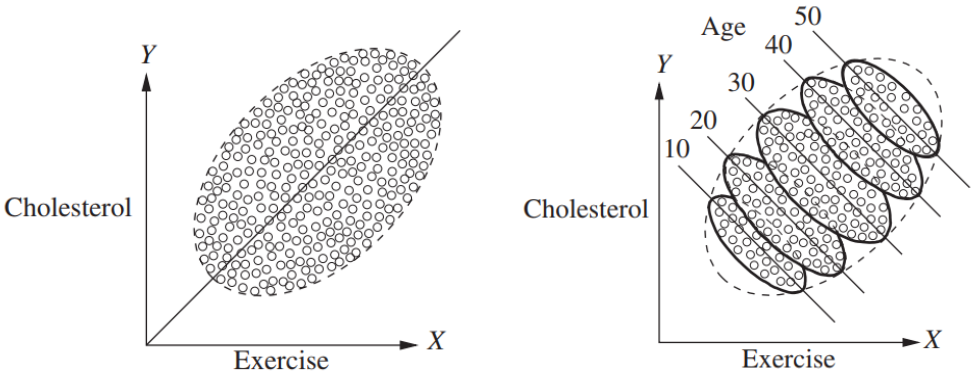
统计学家 2 是正确的。他的研究对应了（a）中的因果图。要确定 X 对 W_F 的因果效应时，存在后门路径 $X \leftarrow W_I \rightarrow W_F$，需要控制变量 W_I。统计学家 2 控制变量了 W_I，得出了正确的结论。

统计学家 1 的研究对应了以下的因果图，其中 G 表示体重增量。要确定 X 对 G 的因果效应时，存在后门路径 $X \leftarrow W_I \rightarrow W_F \rightarrow G$ 和 $X \leftarrow W_I \rightarrow G$，需要控制变量 W_I。统计学家 1 未控制变量 W_I，产生了错误。



(c) 这个例子与辛普森悖论有什么关系？



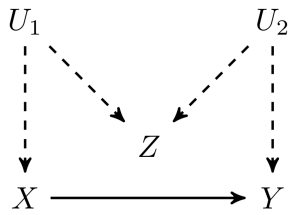


辛普森悖论中的因果图结构与这个例子的因果图结构相同。但本例子中的混杂因子（Confounder）导致了结果从相等变为不相等，辛普森悖论中的混杂因子导致了结果的完全相反。

3 思考题 3.3.3

为评估一种新药的疗效，进行了一项随机试验。总体来说，50% 的患者被分配服用新药，50% 的患者接受安慰剂。在试验的前一天，某位护士给一些表现抑郁的患者分发棒棒糖，他们中的大多数是被指定第二天接受治疗的（也就是说，护士这轮查房正巧经过诊疗病房）。奇怪的是，试验数据揭示了一个辛普森悖论：尽管证明药物对总体受试者来说是有益的，但在得到棒棒糖的亚群中和没有棒棒糖的亚群中服用药物的患者比不服用药物的患者更不容易痊愈。假设吃棒棒糖本身对痊愈没有任何影响，回答以下问题：

(a) 给出一个刻画该故事的因果图。



图中 X 表示是否服用新药，Y 表示治疗的效果，Z 表示是否拿到了棒棒糖，U_1 和 U_2 是两个外生变量。吃棒棒糖与是否服用新药没有直接因果关系，对痊愈也没有任何影响。

(b) 应用后门准则确定必须校正哪些变量。

根据后门准则，确定 X 对 Y 的因果效应时，不存在后门路径（路径 $X \leftarrow U_1 \rightarrow Z \leftarrow U_2 \rightarrow Y$ 被对撞因子 Z 所阻断）。因此，无需校正任何变量。

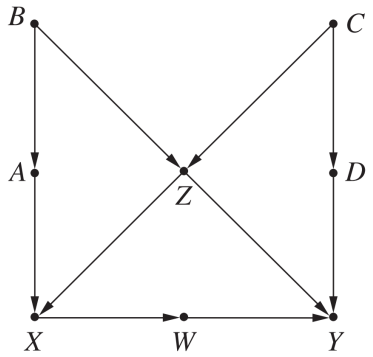
(c) 写出药物对恢复效果的校正公式。

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

(d) 假设护士在研究结束第二天给了棒棒糖，相比于接受安慰剂的那些患者，护士仍然优先选择接受治疗的患者，在此情形下，重新回答问题 (a) ∼ (c)

给棒棒糖的时间并没有影响因果图的结构，因此问题（a）∼（c）的答案无需改变。

4 思考题 3.4.1



假设在上图中只能测量 X，Y 和一个附加变量。哪个变量可用于确定 X 对 Y 的因果效应？请找出并计算 X 对 Y 的因果效应。

要确定 X 对 Y 的因果效应时，图中存在后门路径 $X \leftarrow Z \rightarrow Y$，只能尝试控制变量 Z。

但控制变量 Z 会打开新的后门路径 $X \leftarrow A \leftarrow B \rightarrow Z \leftarrow C \rightarrow D \rightarrow Y$，根据题目要求又不能控制更多的变量，因此无法使用后门准则的方法确定因果效应。

尝试使用前门准则，发现 X 到 Y 的有向路径仅有 $X \rightarrow W \rightarrow Y$，尝试控制变量 W。发现满足了前门准则的全部三个要求：

1. W 切断了所有 X 到 Y 的有向路径，
2. X 到 W 没有后门路径，
3. 所有 W 到 Y 的后门路径都被 X 阻断。

因此可以使用前门校正公式计算 X 对 Y 的因果效应：

$$P(y \mid do(x)) = \sum_{w} P(w \mid x) \sum_{x'} P(y \mid x', w) P(x')$$

5 思考题 3.4.2

假设你去一家药店买一种药，发现不同价格的两个品种：一种标价为 1 美元，另一种标价为 10 美元。你问店员："有什么区别?"店员说："10 美元那瓶是新进的，而 1 美元那瓶已经在货架上 3 年了。但是，数据显示，买便宜药的有更高的痊愈率。很神奇吧?"你追问这些陈旧的药物是否经过测试。店员说："是的，这甚至更神奇，95% 的陈旧药物失去了活性成分，而仅有 5% 的新进药物失去了活性成分，但那些购买旧药的患者仍然比购买了新药的患者的痊愈率高"

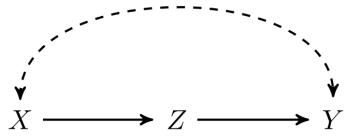
在买陈旧的药物之前，你出于好奇查看了一下数据。数据记录了购买信息，包括购买的药物类型（陈旧或新进）、药物活性成分浓度（高或低）、顾客是否从疾病中痊愈等。数据完美地证实了店员的说法。但是经过一番额外的计算后，你还是决定购买昂贵的新药，甚至没有测试它的活性成分。你确认新药会为普通患者提供更高的痊愈率。

基于两个非常合理的假设，数据明确地表明新药更加有效，这两个假设如下：

（1）顾客不知道他们要买的药中化学成分，他们的选择仅受价格和上架时间影响；

（2）药物对任何个体的疗效仅取决于它的化学成分，而不是上架时间（新进或陈旧）。

(a) 确定问题的相关变量，并以一个因果图来描述该问题。



图中 X 表示购买的药物类型，Z 表示药物活性成分浓度，Y 表示治疗效果。X 和 Y 之间存在一些未观测到的外生变量，例如一个人的经济情况可能既导致了其对不同价格药物的选择，又导致了其治疗恢复的状况不同。

(b) 构建与题干适配的数据集。

复用书中表 3.2 中的数据，得到：

治疗效果		便宜旧药		昂贵新药		合计	
		低活性	高活性	低活性	高活性	低活性	高活性
		380	20	20	380	400	400
痊愈	人数/1000	323	18	1	38	324	56
	比例/%	85	90	5	10	81	14
未痊愈	人数/1000	57	2	19	342	76	344
	比例/%	15	10	95	90	19	86

(c) 使用假设 (1) 和 (2) 及问题 (b) 的数据，确定选择新进药物相比于陈旧药物的效果。

$$y_0 \to \text{未痊愈；}\ y_1 \to \text{痊愈；}$$
$$x_0 \to \text{便宜旧药；}\ x_1 \to \text{昂贵新药；}$$
$$z_0 \to \text{低活性；}\ z_1 \to \text{高活性；}$$

$$\begin{aligned}
P(y_1 \mid do(x_1)) =& \sum_z P(z \mid x_1) \sum_{x'} P(y_1 \mid x', z) P(x') \\
=& P(z_1 \mid x_1) \left[P(y_1 \mid x_1, z_1) P(x_1) + P(y_1 \mid x_0, z_1) P(x_0)\right] \\
&+ P(z_0 \mid x_1) \left[P(y_1 \mid x_1, z_1) P(x_1) + P(y_1 \mid x_0, z_1) P(x_0)\right] \\
=& 0.95 \times (0.1 \times 0.5 + 0.9 \times 0.5) + 0.05 \times (0.05 \times 0.5 + 0.85 \times 0.5) \\
=& 0.4975 \\
P(y_1 \mid do(x_0)) =& \sum_z P(z \mid x_0) \sum_{x'} P(y_1 \mid x', z) P(x') \\
=& P(z_1 \mid x_0) \left[P(y_1 \mid x_1, z_1) P(x_1) + P(y_1 \mid x_0, z_1) P(x_0)\right] \\
&+ P(z_0 \mid x_0) \left[P(y_1 \mid x_1, z_1) P(x_1) + P(y_1 \mid x_0, z_1) P(x_0)\right] \\
=& 0.05 \times (0.1 \times 0.5 + 0.9 \times 0.5) + 0.95 \times (0.05 \times 0.5 + 0.85 \times 0.5) \\
=& 0.4525
\end{aligned}$$

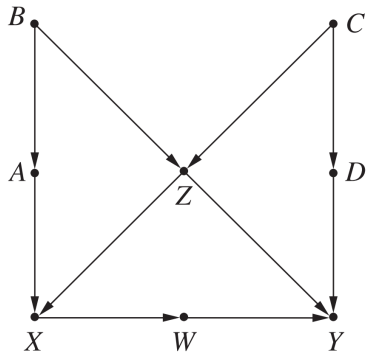
由此可得

$$\text{ACE} = P(y_1 \mid do(x_1)) - P(y_1 \mid do(x_0)) = 0.045 > 0$$

因此可以认为新进药物比陈旧药物效果更好，可以更好地促进痊愈。

6 思考题 3.5.1

考虑下图的因果模型



(a) 找到对于 Y 的 X 的 c-特定效应的表达式（注：干预后变量 C 取 c 时的特定效应）。

令可测变量集合 $S = \{Z\}$，可以验证集合 $S \cup C = \{Z, C\}$ 满足了后门准则，因此满足了书中的规则 3.5.1。因此 Y 的 X 的 c-特定效应可以表达为:

$$P(Y = y \mid do(X = x), C = c) = \sum_z P(Y = y \mid X = x, Z = z, C = c)P(Z = z \mid C = c)$$

(b) 确定一个需要测量四个变量的集合，以便估计 X 对 Y 的 z-特定效应，并给出表示该效应的值的表达式。

类似上一小问，令可测变量集合 $S = \{C\}$，则测量的四个变量为 $\{X, Y, Z, C\}$，该效应的值的表达式为:

$$P(Y = y \mid do(X = x), Z = z) = \sum_c P(Y = y \mid X = x, Z = z, C = c)P(C = c \mid Z = z)$$

(c) 使用（b）的答案，确定在 Z-依赖策略下 Y 的期望值，其中当 $Z \leq 2$ 时 X 被设置为 0，当 $Z > 2$ 时，X 被设置为 1（假设 Z 取 $1 \sim 5$ 的整数值）。

由描述可得

$$X = g(Z) = \begin{cases} 0 & Z = 1, 2 \\ 1 & Z = 3, 4, 5 \end{cases} \tag{1}$$

因此根据书中公式 (3.17) 可得

$$\begin{aligned} P(Y = y \mid do(X = g(Z))) = &\sum_z P(Y = y \mid do(X = g(Z)), Z = z)P(Z = z) \\ = &P(Y = y \mid do(X = 0), Z = 1)P(Z = 1) \\ &+ P(Y = y \mid do(X = 0), Z = 2)P(Z = 2) \\ &+ P(Y = y \mid do(X = 1), Z = 3)P(Z = 3) \\ &+ P(Y = y \mid do(X = 1), Z = 4)P(Z = 4) \\ &+ P(Y = y \mid do(X = 1), Z = 5)P(Z = 5) \end{aligned}$$

之后可以根据（b）中公式继续化简，消去 do-算子。