统计因果推断的学习分享

黄勤来

2021.7.16

基础概率知识

- 随机试验，样本空间，样本点，随机事件
- 样本空间S中每个元素e都会对应一个实数，这种映射关系可以定义为一个函数f(e)，那么这个函数就称为随机变量，用大写符号X、Y、Z表示
- 条件概率：任意两个随机变量AB，有条件概率公式：
- $P(A|B)=\frac{P(AB)}{P(B)}$，那么$P(AB)=P(A|B)P(B)$
- 如果AB为随机事件，可以如上表示，如果AB为随机变量$P(A|B)$是$P\left(A=\mathrm{a_i}\middle|B=b_j\right)$的简写
- 通过条件概率我们可以定义
  (两随机变量相互独立的公式： if $P(A$jiao$B)$=$P(A)P(B)$ ⇒ $P(A|B,c)=P(A|c)$以及$P(B|A)=P(B)$
- $P(A|B,C)=P(A|C)$
- $but\ if\ P(AB)\neq P(A)P(B),\mathrm{and}\ P(A|B,C)=P(A|C)$，我们称AB为条件独立

如何理解条件独立

- 两随机变量相关：$\mathrm{Cov(A,B)} \neq 0$，线性相关系数
- 独立一定不相关（ $\mathrm{Cov(A,B)} = 0$ ），不相关不一定独立，相关一定不独立
- 相关不一定存在因果关系，不独立不一定存在因果关系
- 两随机变量独立一定不存在因果关系，两随机变量存在因果关系一定不独立
- 例子：
- 假如有100个家庭，在各自孩子出生的时候，都在自己的庭院里面种上一棵树。我们发现孩童的身高X与树的高度Y存在强的线性相关性。但是X 与 Y并不存在因果关系（混杂因素年龄Z）
- 那么$P(\mathrm{XY}) \neq P(X)P(Y)$，然而满足$P(Y|X,Z) = P(Y|Z)$以及$P(\mathrm{X}|\mathrm{Y},Z = z) = P(\mathrm{X}|Z)$
- $X = Z + U_x, Y = Z + U_y,$
  $while\ U_x\ and\ U_y\ are\ independent\ and\ characterized\ as\ exogenous\ variables$

贝叶斯公式与多维联合分布的条件展开

- 全概率公式：$B_1, B_2, \cdots, B_n$ 是随机变量B的一个划分，那么
- $P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n)$
- 贝叶斯公式为
- $P(B|A) = \frac{P(AB)}{P(A)} = \frac{P(AB)}{P(A|B_1)P(B_1)+P(A|B_2)P(B_2)+\cdots+P(A|B_n)P(B_n)}$
- 有随机变量$X_1, X_2, X_3, \cdots, X_n$,对于其联合概率（分布）可有：
- $P(X_1, X_2, \cdots, X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1X_2)P(X_4|X_1X_2X_3)\cdots P(X_n|X_1X_2 \dots X_{n-1})$
- 若n=4，那么：
- $P(X_1X_2X_3X_4) = P(X_1)\frac{P(X_1X_2)}{P(X_1)}\frac{P(X_1X_2X_3)}{P(X_1X_2)}\frac{P(X_1X_2X_3X_4)}{P(X_1X_2X_3)}$

图graph与因果模型 causal model

- 有向无环图DAG，directed acyclic graph
- Structural Causal Model：这里的Structural指的是结构性的，表明因果模型中变量的因果关系是概率性的，而非统计资料体现出来的。
- 将外源性的变量归入集合U表示；将依赖于其他变量的内源性变量归入集合V（即除了U之外的变量）；SCM中的箭头表示明确的函数赋值关系，这种函数赋值关系被认为是因果关系
- 如果SCM中变量Y是X的子节点，那么X为Y的直接因果关系，若X为Y的祖先节点，当因果出现不传递的情况，X与Y没有因果关系（独立）

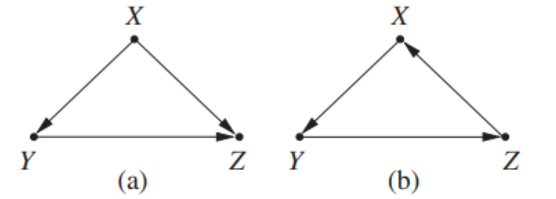
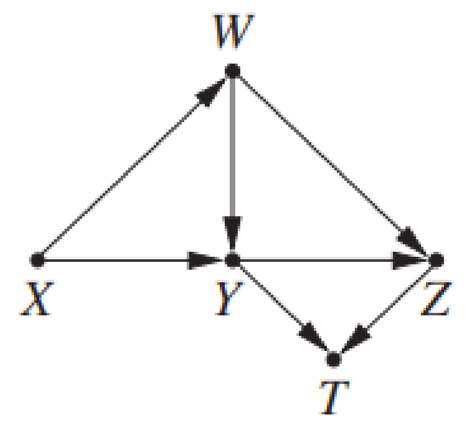
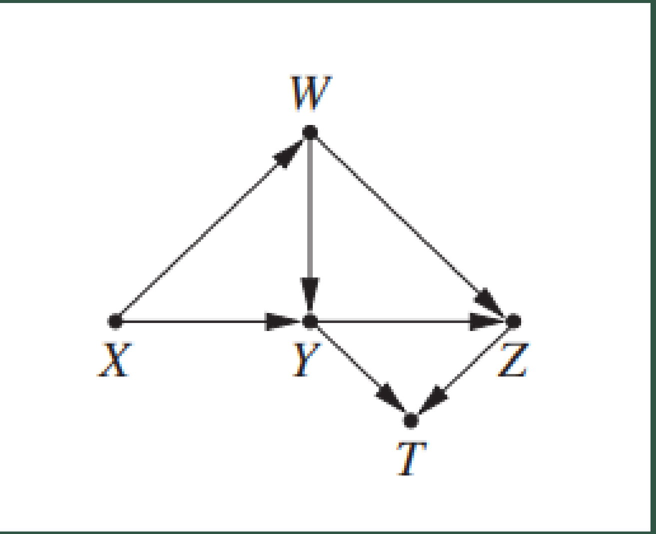


Figure 1.7 (a) Showing acyclic graph and (b) cyclic graph

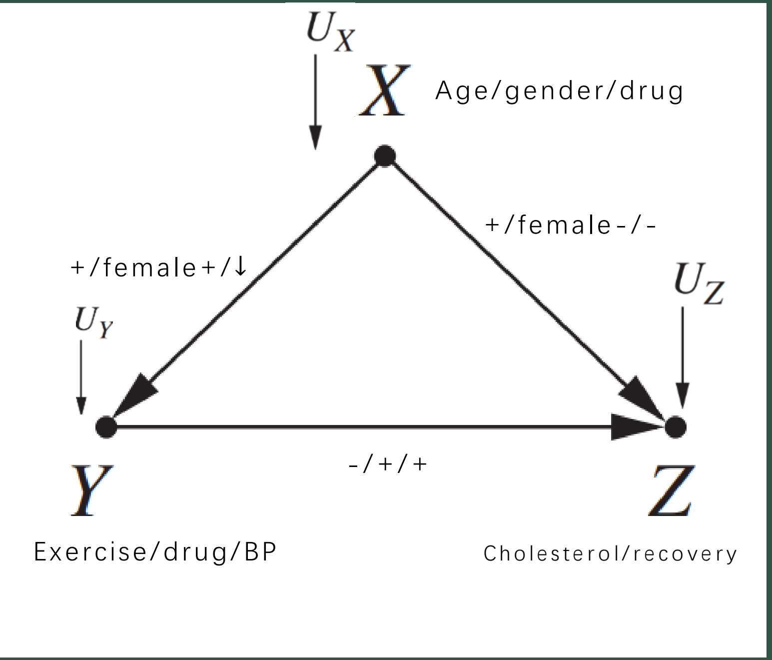


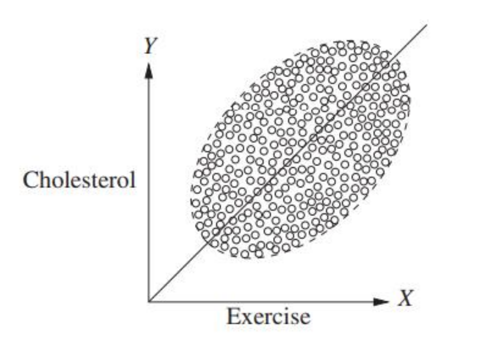
SCM and Product Decomposition

- $P(x_1, x_2, \cdots, x_n) = \prod_i P(x_i|pa_i)$
- $\text{In } it, pa_i\ stdans\ for\ values\ of\ the\ parents\ of\ variable\ X_i$
- 有随机变量 $\text{XYZT}W$,且因果图模型如图所示，对于其联合概率可有：
- $P(\text{XYZT}W) = P(X)P(Y|X)P(Z|XY)P(T|XYZ)P(W|XYZT)$, or
- $P(\text{XYZT}W) = P(X)P(W|X)P(Y|XW)P(Z|XYW)P(T|XYWZ)$
- 已知其因果图，
- $P(\text{XYZT}W) = P(X)P(W|X)P(Y|XW)P(Z|WY)P(T|YZ)$
- Eg2. $X \rightarrow Y \rightarrow Z$
- $Z = y + U_Z$
- $$\begin{aligned} P(X = x, Y = y, Z = z) &= P(X = x)P(Y = y|X = x)P(Z = z|X = x, Y = y) \\ &= P(X = x)P(Y = y|X = x)P(Z = z|Y = y) \end{aligned}$$



Simpson’s Paradox
辛普森悖论





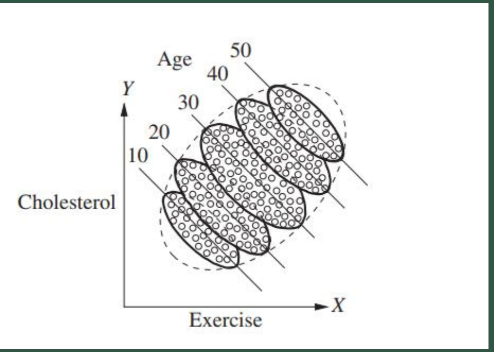


Table 1.1 Results of a study into a new drug, with gender being taken into account

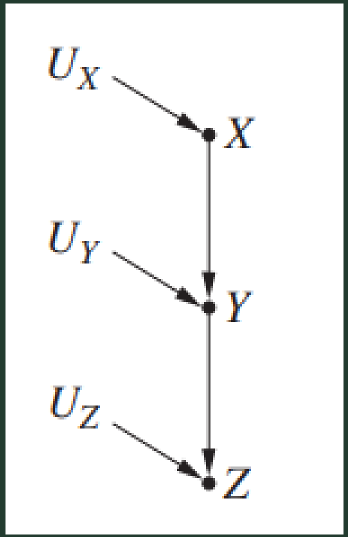
	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

Table 1.2 Results of a study into a new drug, with posttreatment blood pressure taken into account

	No drug	Drug
Low BP	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

因果模式
链式chains

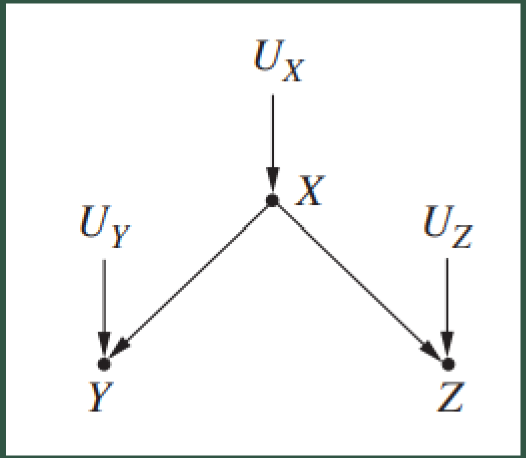
- $\mathrm{V} = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$
- $f_X: X = U_X$
- $f_Y: Y = \frac{x}{3} + U_Y$
- $f_Z: Z = \frac{y}{16} + U_Z$
- 其中 X 表示学校资金，Y表示该校学生平均成绩，Z表示该校大学录取率



1. Z与Y，Y与X相互不独立
2. P(X|Y)≠ $P(X)$ P(Z|X)≠ $P(Z)$
3. Z与X很可能不独立
4. **如果对Y取条件，Z与X相互独立**
   if Y is a path between X and Z

$$f_Y : Y = \begin{cases} a & \text{IF } X = 1 \text{ AND } U_Y = 1 \\ b & \text{IF } X = 2 \text{ AND } U_Y = 1 \\ c & \text{IF } U_Y = 2 \end{cases}$$

因果模式 - 分叉forks



1. X与Y，Z与X相互不独立

   $for\ some\ x, y\ P(X = x|Y = y) \neq P(X=x)$

2. Z与Y很可能不独立

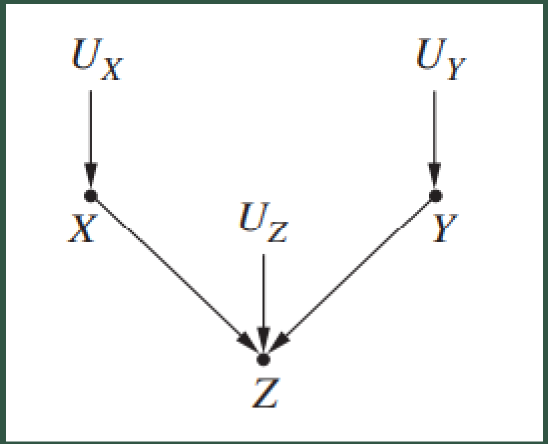
3. **如果对X取条件，Y与Z相互独立， X为Z与Y的共同的因素，且Y与Z之间不存在其他路**

- For $all\ x, y, z,$

  $P(Y = y|Z = z, X = x) = P(Y = y|X = x)$

- $V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$
- $f_X: X = U_X$
- $f_Y: Y = 4x + U_Y$
- $f_Z: Z = \frac{x}{10} + U_Z$
- 其中 X 表示温度，Y表示冰淇淋销售量，Z表示犯罪率

因果模式 - 汇聚/碰撞 collider



1. X与Z，Y与Z相互不独立

   $for\ some\ x, z, P(X = x|Z = z) \neq P(X{=}x)$
2. X与Y相互独立，因为U_X and U_Y相互独立
3. **如果对Z取条件，那么X与Y相互不独立，Z为X与Y的collider或collider的子孙节点**

- For $some\ x, y, z,$

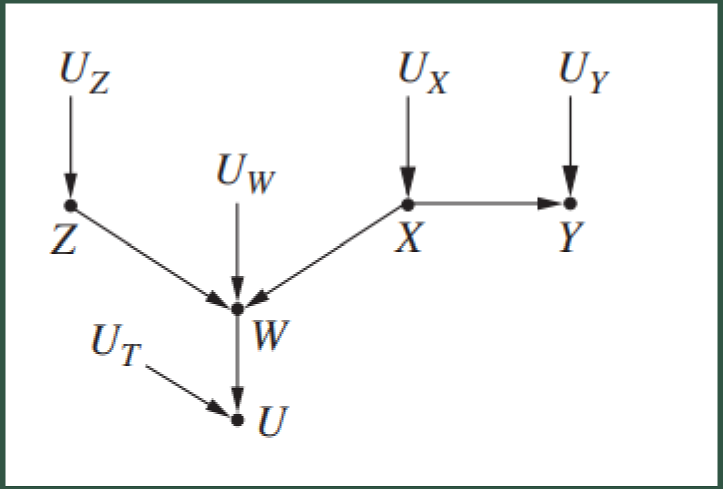
  $P(Y = y|X = x，Z = z) \neq P(Y = y|Z = z)$

- $V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$
- $f_X: X = U_X$
- $f_Y: Y = U_Y$
- $f_Z: Z = x + y + U_Z$
- 或者X表示一枚硬币的正反面{0,1}，Y表示另一枚硬币的正反面{0,1}，Z表示两枚硬币的情况 {0,1,2}
- $P(Y = 1|X = 1，Z = 2)$=1
- $P(Y = 1|Z = 2)$=1 ???
- $P(Y = 1|X = 1，Z = 1)$=0
- $P(Y = 1|Z = 1)$=0.5

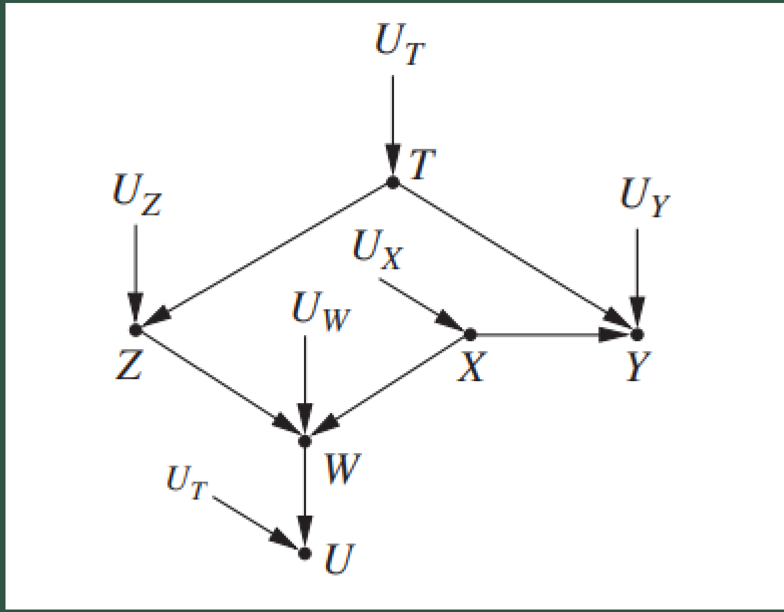
有向分割与有向连接

directional-separation and d-connection

- Z={}
- Z={W} 、 {U}
- Z={W，X}

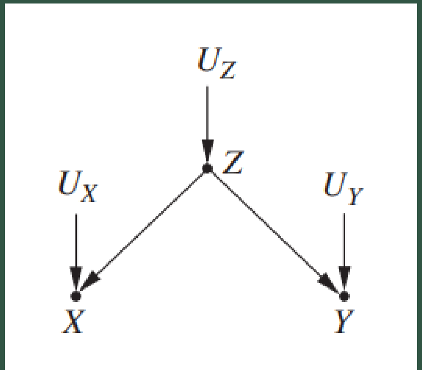


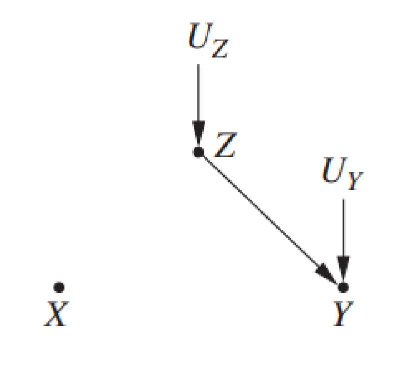
1. 应用之前的规则来处理复杂因果图
2. 如果两个节点/随机变量是可有向分割的，那么两随机变量一定是相互独立的
3. 如果两个节点是有向连接的，那么随机变量有极大可能是不独立的、有关联的
4. 依赖的传递：如果不通过对某节点取条件，那么只有collider会阻断依赖的传递；
5. 如果要对集合Z取条件进行阻断依赖的传递；那么需要满足1）collider以及其子孙没在Z中；2）链与分叉的中间节点在Z中



干预 Interventions

- 干预 VS 取条件概率
- 干预会改变整个系统，系统中其他随机变量的值也会随之改变
- 对某个变量取条件概率，则不对系统做任何改变。我们只是在范围更小的样本集合中做计算
- 当我们干预某个变量的时候，我们希望遏制该随机变量由于其父节点的变化而改变的趋势，那么相当于减掉指入该节点的箭头。

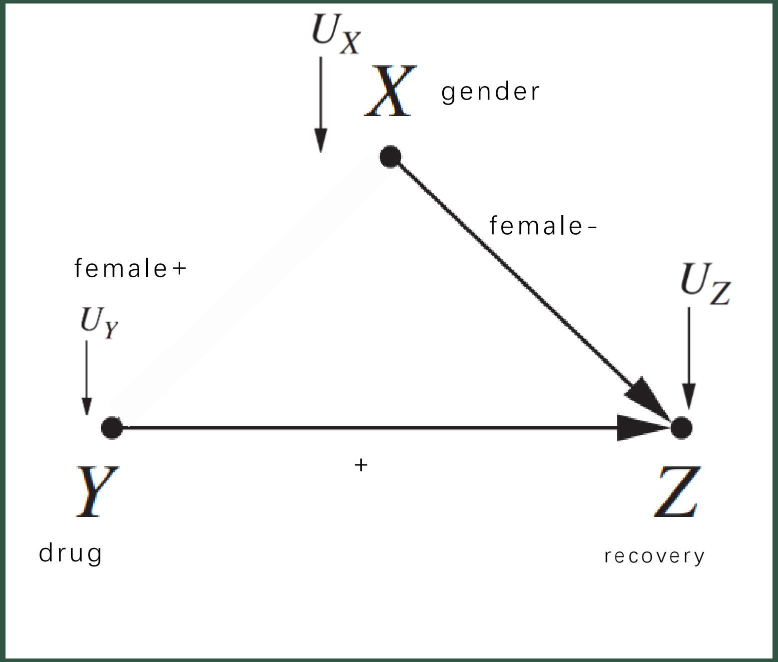




1. 干预一个变量：
2. $do(X=x)$来表示对X变量进行干预，使其取值为x
3. 条件概率：$P(Y=y|X=x)$与干预之后的概率$P(Y=y|do(X=x))$所代表的的含义不同。值取决于SCM，可能相同，可能不同
4. 从分布的术语来讲，$P(Y=y|X=x)$表示总体中X=x的部分的Y=y的概率；而$P(Y=y|do(X=x))$表示如果总体所有个体的X都等于x的时候，Y=y的概率。
5. $\mathrm{P}(Y=y|do(X=x),Z=z)$表示如果总体每个个体的X都等于x的时候，已知Z=z，Y=y的条件概率

通过干预从相关关系中析取因果

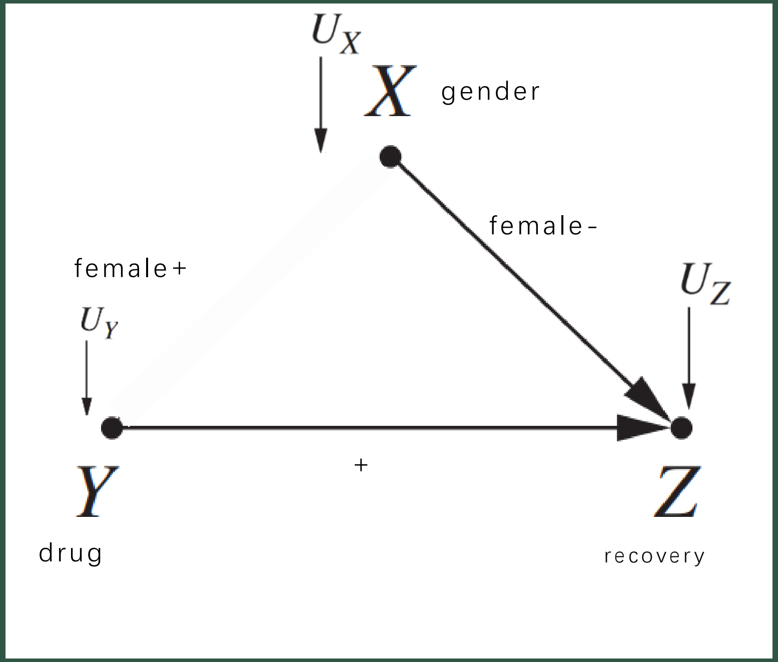
- 前提，已获得真实有效的SCM。同时干预措施不会在真实世界中造成“副作用”，否则应该将该副作用加入SCM中
- $P(Z=1|do(Y=1))-P(Z=1|do(Y=0))$
- 以上average causal effect （ACE）



1. 干预一个变量：
2. $P(Z=z|do(Y=y))= P_m(Z=z|Y=y)$
3. 那么我们考虑P_m与原本的概率P分布的关系。首先X的概率分布不会受到$do(Y=y)$的影响。即样本中男女的比例不会改变；其次$P(Z=z|Y=y,X=x)$概率仍然不会发生改变，因为Z只由x，y决定，一旦x，y给定，z的概率只由U_Z决定
4. $\Rightarrow$ $P_m(Z=z|Y=y,X=x)=P(Z=z|Y=y,X=x)$
   $$P_m(X=x)=P(X=x)$$
5. 同时在施加干预之后，Y与X独立了，那么$P_m(X=x|Y=y)=P_m(X=x)=P(X=x)$
6. $P_m(Z=z|Y=y)=$
   $\sum_x P_m(Z=z|Y=y,X=x)P_m(X=x|Y=y)\Rightarrow$
   $\sum_x P(Z=z|Y=y,X=x)P(X=x)\Rightarrow$
   ???? $P(Z=z|Y=y)$

Control or not control

- $P(Z = z|do(Y = y)) = \sum_x P(Z = z|Y = y, X = x)P(X = x)$, adjust for X
- 注意，如果做的是随机试验，不需要控制混杂因素X，因为随机试验要求除了干预措施Y，其他变量都是随机的，从而X不会对Y产生影响，而X对Z的影响对ACE不影响。



1. 从而干预一个变量:
2. $P(Z = z|do(Y = y)) = P_m(Z = z|Y = y) = P(Z = z|Y = y)$
3. $P_m(Z = z|Y = y) = \sum_x P_m(Z = z|Y = y, X = x)P_m(X = x|Y = y) \Rightarrow \sum_x P(Z = z|Y = y, X = x)P(X = x)$
4. In practice, investigators use adjustments in randomized experiments as well, for the purpose of minimizing sampling variations

Table 1.1 Results of a study into a new drug, with gender being taken into account

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

$$P(Z = 1|do(Y = 1))$$

$$= \sum_{0,1} P(Z = 1|Y = 1, X = 0|1)P(X = 0|1) = 0.832$$

$$P(Z = 1|do(Y = 0))$$

$$= \sum_{0,1} P(Z = 1|Y = 0, X = 0|1)P(X = 0|1) = 0.7818$$

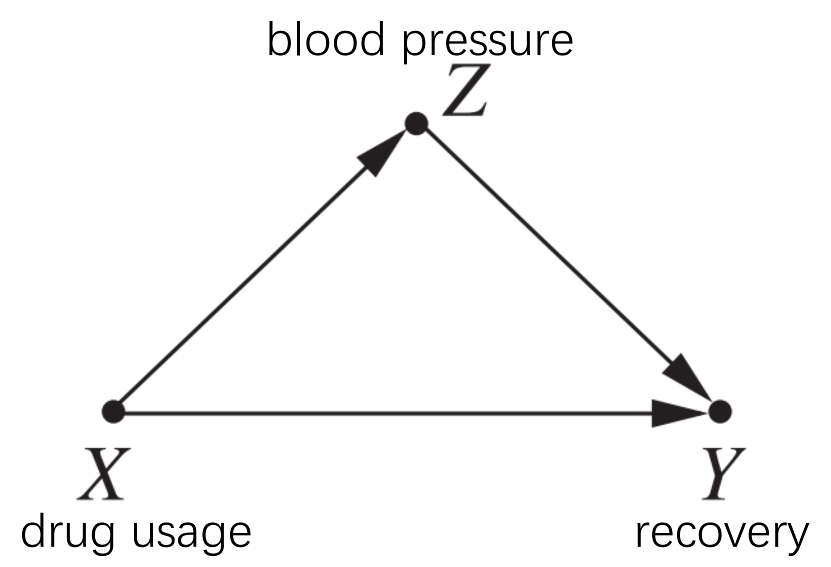
$$ACE = 0.832 - 0.7818 = 0.0502$$

指导我们按照X分层看好坏

Control or not control

Table 1.2 Results of a study into a new drug, with posttreatment blood pressure taken into account

	No drug	Drug
Low BP	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)



$$P(\mathrm{Y}=1|do(X=1)) = P(\mathrm{Y}=1|X=1)$$
$$= \sum_{z=0,1} P(\mathrm{Y}=1|X=1, Z=z)\, P(Z=z|X=1)$$

指导我们看总的表格

Adjust or control the parents of X (the intervention) rather than the parents of result variables in the original graph

$$P(Y=y|do(X=x))$$
$$= \sum_{z} P(Y=y|X=x, PA_X=z) P(PA_X=z)$$
$$\Rightarrow mul\ and\ divide\ factor\ P(X=x|PA_X=z)$$
$$\Rightarrow P(y|do(x)) = \sum_{z} \frac{P(X=x, Y=y, PA_X=z)}{P(X=x|PA_X=z)}$$

propensity score 倾向性分数

多变量干预以及截断乘积法则

multiple intervention and the truncated product rule

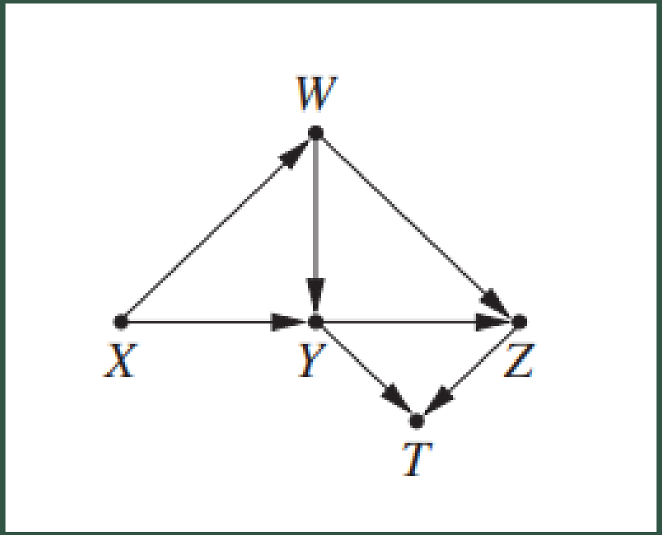
假如我们要对不止一个变量做干预，应该如何调整

$$P(XTW|do(Y=y, Z=z)) = P(X)P(W|X)P(T|yz)$$

$$P(T|do(Y=t, Z=z)) = \sum_{x,w} P(X)P(W|X)P(T|yz) = P(T|yz)$$

$$P(x_1, x_2, \cdots, x_n|do(\vec{x})) = \prod_i P(x_i|pa_i)\,, for\ all\ x_i\ not\ in\ \vec{x}$$

Adjust or control the parents of X (the intervention) rather than the parents of result variables in the original graph

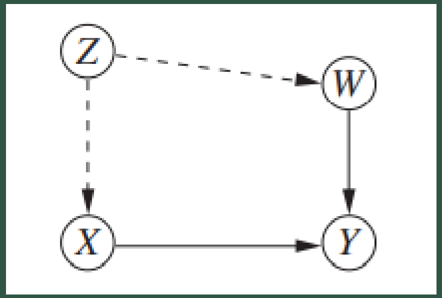


$$P(XYZTW) \\ = P(X)P(W|X)P(Y|XW)P(Z|WY)P(T|YZ)$$

$$P(XTW|do(Y=t, Z=z)) \\ = P_m(X)P_m(W|X)P_m(Y|XW)P_m(Z|WY)P_m(T|YZ$$

后门准则
Backdoor criterion

控制 施加干预措施的变量 的父节点，但存在有些父节点不可观测的情况，这个时候需要使用后门准则来阻断不可观测的父节点的影响



backdoor criterion：给定一组有序变量（X，Y）存在于因果图中，如果存在一组变量Z能够阻断（X，Y）之间的所有路，并且Z中不存在X的后代节点，那么，调整后公式为

$$P(Y=y|do(X=x)) = \sum_z P(Y=y|X=x, Z=z)P(Z=z)$$

Adjust or control the parents of X (the intervention) rather than the parents of result variables in the original graph

1. We block all spurious paths between X and Y.
2. We leave all directed paths from X to Y unperturbed.
3. We create no new spurious paths.

$$P(Y=y|do(X=x)) = \sum_z P(Y=y|X=x, PA_X=z)P(PA_X=z)$$

CAUSAL INFERENCE IN STATISTICS A PRIMER
Judea Pearl