

统计因果推理：入门

1. 引言：统计和因果模型

1.1 为什么研究因果关系

“为什么要研究因果关系”这一问题的答案与“为什么要研究统计学”几乎一样直接。我们研究因果关系，是因为我们需要理解数据进而指导行动与决策，并从成功与失败中总结出经验与教训。例如，通过研究预估吸烟对肺癌发病率、受教育程度对薪资水平、碳排放对气候产生的影响等。在研究因果关系的基础上，还要进一步挖掘因果关系产生的原因以及对结果产生的影响，这一点是很有价值的。例如，知道疟疾是通过蚊子传播还是通过“异常空气”传播后，我们就知道下次湿地旅行时应该携带蚊帐还是呼吸面罩了，而疟疾通过异常空气传播曾被许多人相信。

为什么要将因果关系区别于传统的统计学习课程作为独立的话题来研究呢？这一问题的答案并不明显。就起本身而言，“因果关系”这一概念说明了这个世界的一些信息，而实证的统计学方法却不能。

已有大量的事实证明了这一点。严格来说，因果关系不只是统计学的一个方面，当它与传统的统计学结合后，将会揭示世界的运行机制，这是仅靠统计方法不能实现的。例如，前面提到的任何问题，都不能用标准的统计学语言来描述，这可能令许多人感到吃惊。为了理解因果关系在统计学中的特殊作用，首先来看一个有趣的统计学悖论。他可以形象地说明为什么传统统计学必须补充新内容，才能处理诸如上面提到的那些因果关系。

1.2 辛普森悖论

辛普森悖论（Simpson's paradox）以第一个论及该问题的统计学家Edward Simpson（生于1922年）命名，该悖论指出：存在这样的数据，总体上的统计结果与其每一个子部分的统计结果相反。下面通过一个实例说明。根据有关统计数据，平均来说，吸烟人群比不吸烟人群的收入更高；但当考虑吸烟人群的年龄因素时就可能发现，在每个年龄组，吸烟人群的收入低于不吸烟人群；如果再同时纳入年龄和学历这两个因素，可能又会发现相同年龄和学历的吸烟者比不吸烟者收入高。可见，随着考虑的因素增多，统计结果会不断发生逆转。在类似这样的问题中，想要确定吸烟是否会影响收入以及影响有多大，仅从数据表面似乎无法获得准确的答案。

在辛普森使用的经典例子（1951年）中，一组患者可以选择是否尝试一种新药。根据总体统计，服用该药的患者其痊愈率却低于未服药的患者。然而，当对患者按性别划分时，发现服药的男性患者比不服药男性患者痊愈率高，服药的女性患者也比不服药的女性患者痊愈率高！换句话说，这种药似乎分别有益于男性患者和女性患者，但从男性患者和女性患者所构成的全体受试者来看却是无益的，这似乎是矛盾的，也是不可思议的，这就是为什么该例子被认为是一个悖论的原因。有些人很难相信这一悖论，下面详细说明。

例1.2.1 记录选择服药与否的700例患者的痊愈率。其中，350例患者服药，350例患者不服药。研究结果如表1.1所示。

表1.1	用药	不用药
男性	81/87(93%)	234/270(87%)
女性	192/263(73%)	55/80(69%)
共计	273/350(78%)	289/350(83%)

如表1.1所示，第一行是男性患者服药与不服药的对比，第二行是女性患者服药与不服药的对比，第三行表示所有患者服药与不服药的对比。男性患者中，服药患者痊愈率（93%）比未服药患者痊愈率（87%）高。这一结果同样出现在女性患者中（分别是73%、69%）。然而，对全体受试者而言，服药患者痊愈率（83%）比未服药患者痊愈率（78%）高。

表1.1中数据似乎说明，如果知道患者的性别（男性或女性），那么就可以开出药物，但如果性别不明，则不能开药！然而，这个结论是荒谬的。如果药物有益于男性患者和女性患者，那么它必然对任何患者都有效，忽略患者的性别，并不会使药物变得无效。

鉴于这一研究结果，医生是否可为一名女性或男性或未知性别的患者开药呢？或者考虑一个正在评估药物对总人群有效性的决策者，他是否应该采信总人群的痊愈率数据？或者，他是否应该考虑按性别划分得出的亚群痊愈率数据？

这个问题无法简单地从统计学中找到答案。为了考察药物对患者的作用，首先要了解数据背后的原因，即产生结果的因果机制。例如，**假设已知另外一个事实**：雌激素对患者痊愈有负面效应，那么不管是否服药，女性患者总比男性患者难以痊愈。另外从表1.1所示数据可以看出，女性患者比男性患者更倾向于被选中服用药物。所以如果随机选择一名服药者，这个人更有可能是女性，因此与不服药的受试者相比，服药者更加倾向于未痊愈，这就给人感觉药物对于全体受试者是无效的。也就是说，女性是与用药和未痊愈都相关的共同因素。因此为了评估有效性，我们需要比较同一性别的受试者，从而确保服药与否的痊愈率差异并不归因于雌激素。这意味着我们应该研究分类数据，这些数据明确告诉我们药物是有益的，这也符合我们的直觉，即分类的数据比未分类的数据“更详细”，因此具有更大的信息量。

注：上述中仍然有一个假设，即先验知识。并且是因为这个先验知识后，才得知需要根据性别进行分组查看数据，进而得出结论的。因此，我目前认为先验知识(假设)尤其重要，它似乎很大程度上影响着结论。

对上面的例子做少许变化，进一步分析在连续取值的例子中是怎样发生类似的逆转的。以调研不同年龄段的每周锻炼量和胆固醇为例。如图1.1所示，当用横轴表示锻炼量，纵轴表示胆固醇，并按年龄划分时，可以看到每组呈总体下降趋势：即年轻人锻炼得越多，胆固醇越低，这同样适用于中年人和老年人。如果采用不按年龄划分的散点图，如图1.2所示，除未显示不同年龄组边界外，数据点与图1.1相同，而图1.2呈现出的是一个总体向上的趋势：人们越锻炼，胆固醇越高。为了探究这个问题的根源，我们再次挖掘数据背后的原因。如果已知**对于愿意锻炼的老年人（见图1.1）**，**无论锻炼与否，他们都更可能具有高的胆固醇**，那么就很容易解释这个逆转了。年龄是与锻炼和胆固醇都相关的共同因素。因此，应该考虑将数据按年龄分组，在同年龄人之间做比较，也就不会得出锻炼量大的人胆固醇反而高这种由年龄而非锻炼导致的错误结论了。

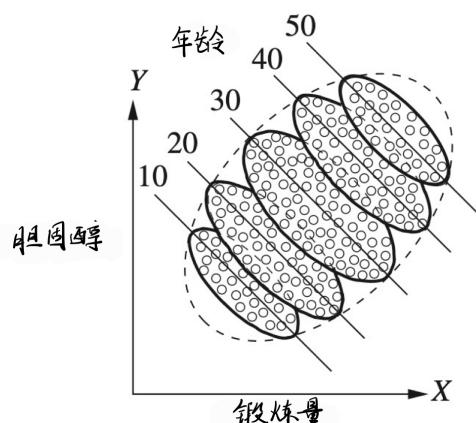


Figure 1.1 Results of the exercise–cholesterol study, segregated by age

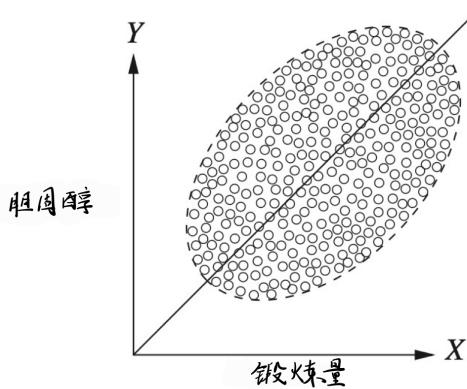


Figure 1.2 Results of the exercise–cholesterol study, unsegregated. The data points are identical to those of Figure 1.1, except the boundaries between the various age groups are not shown

注：在这个连续型变量的例子当中，同样也是需要有先验知识(年龄会影响胆固醇)，从而才能得知可能“正确”的结果。因此，我仍然认为若在陌生应用场景中，若不能发现这种先验知识，那还是要很大程度上依赖事实数据。

然而，分类的数据并不总能给出正确的答案，这可能让一些读者感到惊讶。假设使用与例1.2.1相同的用药和痊愈数据，但在试验结束后还要记录每位患者的血压，而非性别。试验数据如表1.2所示，对比表1.1可知，区别仅在于行列标签有差异。在这一案例中，我们知道药物会通过降低服药患者的血压来影响痊愈率，但不幸的是，药物也会产生副作用。

表1.2	用药	不用药
低血压	81/87(93%)	234/270(87%)
高血压	192/263(73%)	55/80(69%)
共计	273/350(78%)	289/350(83%)

现在，你会给患者推荐这种药物吗？

同样，答案因数据分类方式的不同而不同。对全体受试者而言，药物可能因其对血压的影响而提高痊愈率。但将受试者分类后，在治疗后血压偏高和治疗后血压偏低的亚群中，我们无法观察到这样的结果，而只能看出因药物副作用而降低痊愈率。

与例 1.2.1 一样，本试验的目的是评估药物对痊愈率的总体影响。但在这个例子中，由于降低血压是药物影响痊愈率的结果之一，所以基于血压的分类就变得没有意义了（如果在治疗前记录患者的血压，并且假定仅有血压对治疗有影响，那么情况就不同了）。我们再次统计、分析全体受试者的试验数据，发现药物治疗增加了痊愈的可能性，于是我们确定应该推荐药物治疗。值得注意的是，虽然表 1.1 和表 1.2 中的数值相同，但表 1.1 的正确结论体现在分类后的数据，而表 1.2 的正确结论体现在总体数据。

由前面的例子可知，数据并没有为治疗决策提供足够的信息，例如，无法知道何时测定药物的作用，无法知道药物如何影响血压，也无法知道血压如何影响痊愈率。事实上，正如统计学教科书所常提及（且正确地）指出的，相关性并不等于因果关系。利用统计方法无法仅根据数据确定因果关系，因此统计方法无法为决策提供支持。

然而，统计学一直是以某种因果假设来解释数据的。事实上，在辛普森悖论中，按性别对数据进行分类后之所以会得出矛盾的结论，其根本原因在于我们确信治疗不能影响性别。如果治疗可以影响性别，那么悖论就不存在了，因为我们可以轻松假设数据背后的因果关系具有与按血压分类那个例子相同的结构。尽管“治疗不能影响性别”这一命题看似平凡，但它无法通过数据验证，也无法依据

标准统计学写出数学表达式。事实上，列联表（如表 1.1 和表 1.2）无法表达任何因果信息，而统计推理通常是以列联表为基础的。

可喜的是，最新发展的统计学方法能够用于表达和解释因果假设。这些方法及其内涵是本书讨论的重点内容。通过运用这些方法，读者可以用数学语言描述任何复杂的因果场景，解决类似辛普森悖论引发的决策问题，就像在代数中求解未知量一样“有法可依”。利用这些方法，可以轻松识别上述例子中的问题，并用适当的统计分析方法加以解释。通过简单的逻辑操作组合而成的因果演算，将验证我们已有的直觉，包括不存在一种药物会仅对男性患者或女性患者有效，但对总体人群无效，以及对（服药后）具有相同血压患者的无用性。这些方法还能处理更复杂的问题，而在这些问题中无需再依赖直觉来分析。也就是说，简单的数学工具有助于解决政策评估中的实际问题，以及诸如事件如何发生和为何发生等科学问题。

但是，我们还没有做好完成这些任务的准备。为了严格表述数据中蕴含的因果关系，还需要完成以下工作：

第一，给出“因果关系”的可操作性定义；

第二，给出表达因果假设的形式化方法，即建立因果模型；

第三，给出因果模型结构与数据特征相联系的方法；

第四，给出从数据与模型的因果假设中得出结论的方法。

本书的前两章主要介绍因果假设建模及其与数据相关联的方法，接着在第3章我们使用这些假设和数据来解决因果问题。在开始所有讨论之前，首先必须给出因果关系的定义。因果关系看似简单直观，但几个世纪以来，仍没有一个得到统计学家和哲学家共同认可的、完整的因果关系定义。本书将因果关系的定义简化为：如果变量 Y 的值以某种形式依赖于变量 X 的值，那么变量 X 就是变量 Y 的原因。之后我们会对这个定义稍加扩展，但现在仅将因果关系看作是一种“听从”，即如果 Y 听从 X ，需依据 X 的取值来决定自身的值，那么 X 就是 Y 的原因。

为了理解上述的因果方法，读者还必须了解一些概率学、统计学及图论的基本概念。因此，下面两节将给出这些必要的基本概念和示例。读者若对概率、统计和图论有基本了解，可直接转到第1.5节，这并不会影响对全书内容的理解。

思考题

1.2.1

下面的命题有什么错误？

(a) 数据表明，收入与婚姻具有高度的正相关。因此，如果你结婚了，你的收入会增加。

(b) 数据表明，随着火灾数量的增加，消防员的数量也会增加。因此，为了减少火灾，应该减少消防员的数量。

(c) 数据表明，匆忙赶赴会议的人，通常更容易迟到。因此不要着急，否则你会迟到。

1.2.2

球手蒂姆的击球率比他的队友弗兰克高。然而，有人注意到，弗兰克分别跟右手投手和左手投手比赛时，击球率都比蒂姆高。为什么会这样？（用表格说明）

1.2.3

对于下面的每一个因果问题，判断应该使用分类数据还是总体数据以获得正确的结论。

(a) 肾结石治疗有两种方法：方案 A 和方案 B。医生对大的结石（因此病情更严重）更倾向于采用方案 A，对小的结石倾向于采用方案 B。如果一个患者不知道其体内结石的大小，为确定哪种治疗方案更有效，应该检索总体人群的数据还是结石大小不同亚群的数据？

(b) 在一个小镇上有两位医生，两人在其职业生涯中都曾做过 100 例手术。手术分两类：一类是非常复杂的，一类是非常简单的。相比而言，第一位医生经常做简单的手术，第二位医生经常做复杂的手术。假如你需要做手术，但不知道自己的情况是属于简单情况还是复杂情况。为尽可能提高手术成功概率，你应该查阅每位医生的总体成功率，还是分别查阅他们各自简单和复杂手术的成功率呢？

1.2.4

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，护士这一轮查房正巧经过诊疗病房）。奇怪的是，试验数据显示了一个辛普森悖论：尽管总体来看，药物对受试者是有益的，但在得到棒棒糖的亚群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) 药物对总体受试者是有益的还是有害的？

(b) 你的答案能反驳前面的按性别分类的例子吗？在那个例子中，依据性别对数据进行分类更合适。

(c) 画一个简略图解释这一内容（可参考 1.4 节相关内容）。

(d) 你怎么解释这个问题中出现的辛普森悖论？

(e) 如果棒棒糖在试验后的第二天分发（按同样的准则），答案会不同吗？

[提示：接受棒棒糖的患者更可能被安排药物治疗，同时这些患者更有抑郁表现，而抑郁是降低痊愈可能性的风险因素。]

1.3 概率和统计

由于统计本身通常关注的是可能性而非绝对性，因此对统计来说，概率表述极其重要。概率对于因果关系的研究同样重要，因为大多数因果命题都具有不确定性。例如，“粗心驾驶导致事故”这句话是对的，但并不意味着一位粗心的司机一定会引发事故。概率是表达不确定性的工具。本书将使用概率的语言及定理来描述我们对现实世界的信念和不确定性。为帮助读者更好地理解本书后续内容，下面介绍一些需要了解的重要术语和概念。

1.3.1 变量

变量是能取多个值的任意属性或符号。例如，在一项比较吸烟者和不吸烟者健康状况的研究中，变量可以是参与者的年龄、性别、是否有家族癌症史、吸烟多少年等。可以把变量看作一个问题，而变量的值就是这个问题的答案。例如，“这个参与者多大年龄了？”“38岁”，这里“年龄”是变量，“38”是变量的值。变量 X 所取值 x 的概率记作 $P(X = x)$ ，没有歧义时通常简写为 $P(x)$ 。也可以讨论多个变量同时取值的概率，例如， $X = x$ 与 $Y = y$ 的概率记作 $P(X = x, Y = y)$ 或 $P(x, y)$ 。因此 $P(X = 38)$ 表示人群中随机选择一人，其年龄为38岁的概率。

变量可分为离散变量和连续变量。离散变量（有时又称类别变量）可以从有限集或可数无限集中取值，这些集合中的值可以是任意范围的。例如，描述一个标准开关状态的变量是离散变量，因为它有两个值：“开”和“关”。连续变量可以从无限集中任取一个值，这些集合中的值是连续区间上的值（即，对于任何两个值，存在位于它们之间的第三个值）。例如，描述体重的变量是连续变量，因为重量的测量值是实数。

1.3.2 事件

为一个变量或者一个变量集合指定一个值（或一组值）称为一个事件，例如，“ $X = 1$ ”“ $X = 1$ 或 $X = 2$ ”“ $X = 1$ 与 $Y = 3$ ”“ $X = 1$ 或 $Y = 3$ ”是事件。“掷硬币正面朝上”“调查对象年龄大于40岁”“患者痊愈”等也是事件。“掷硬币的结果”是变量，“正面朝上”是它的值；“调查对象的年龄”是变量，“40岁以上”描述了变量可以取的值；“患者的状态”是变量，“痊愈”是值。这里“事件”的定义与人们日常概念并不完全一致，日常概念中的事件指发生某些变化。例如，在日常谈话中，我们不会将某人处于某个年龄作为一个事件，但我们会将他长了一岁作为一个事件。另一种从概率角度考虑事件的方法是：任何陈述性说明（说明可能是真或假）都是一个事件。

思考题

1.3.1

对于思考题 1.2.4 中棒棒糖的故事，可以识别出以下变量和事件。

1.3.3 条件概率

假设已知事件 B 已经发生，那么此时事件 A 发生的概率称为在 B 条件下 A 的条件概率。已知 $Y = y$ 条件下， $X = x$ 的条件概率记作 $P(X = x | Y = y)$ ，通常简记为 $P(x | y)$ ，表示事件“ $X = x$ ”的概率取决于给定的条件“ $Y = y$ ”。例如，你现在得流感的概率很低，但是如果你量体温发现是 39°C ，那你患流感的概率就大大提高了。

当数据乐使用由频次表示的概率时，一种考虑条件的方法是依据一个或多个变量的值过滤数据。例如，假设查看2012年美国总统选举中美国选民的年龄，根据人口普查品的统计，得到如表1.3所示的数据集。

在表1.3中，总共有132,949,000张选票，因此可以估算一位选民年龄小于45岁的概率。

$$P(\text{选民的年龄} < 45) = \frac{20,539,000 + 30,756,000}{132,949,000} = \frac{51,295,000}{132,949,000} = 0.38$$

假定已知一位选民的年龄大于29岁，想要估算其年龄小于45岁的概率，则只需用年龄大于29岁的条件简单过滤数据，形成一个新的数据集（如表1.4所示）。

这个新的数据中，总共有112,410,000张选票，因此可以估算

$$P(\text{选民的年龄} < 45 | \text{选民的年龄} > 29) = \frac{30,756,000}{112,410,000} = 0.27$$

表1.3 2012年美国总统选举选民/人数 (按年龄分组)

年龄组	选民人数
18~29	20,539,000
30~44	30,756,000
45~64	52,013,000
65以上	29,641,000
合计	132,949,000

表1.4 2012年美国总统选举选民数 (按大于29岁的年龄分组)

年龄组	选民人数
30~44	30,756,000
45~64	52,013,000
65以上	29,641,000
合计	112,410,000

在因果问题的研究中，你这样的条件概率起着重要的作用，因为经常安比软在不同的过滤或是不条件下，结果的概率（或风险）变化情况。例如，与不吸烟的人相比吸烟者患肺癌的概率。

思考题

1.3.2

如表1.5所示为某年美国成年人与性别与受教育水平的关系。

性别	最高学历	人数(单位：十万)
男性	高中以下	112
男性	高中	231
男性	大学	595
男性	研究生	242
女性	高中以下	189
女性	高中	189
女性	大学	763
女性	研究生	172

- (a) 估算 $P(\text{高中})$ 。
- (b) 估算 $P(\text{高中} + \text{女性})$ 。
- (c) 估算 $P(\text{高中} | \text{女性})$ 。

(d) 估算 $P(\text{女性} \mid \text{高中})$ 。

1.3.4 独立性

存在这样一种情况：一个事件的概率不随另一个事件的发生而改变。例如，当观察到事件“你体温升高会增加你患流感的率”时，对于事件“你的朋友乔38岁的概率没有任何影响”。在这种情况下，我们说这两个事件是独立的。一般地，如果

$$P(A \mid B) = P(A) \quad (1.1)$$

则称事件A与B相互独立。也就是说，事件B发生对A发生的概率不能提供任何额外信息。如果这一等式不成立，则说A与B相关。相关与独立具有对称关系：如果A与B相关，那么B也与A相关；如果A独立于B，那么B也独立于A。形式化表述为：如果 $P(A|B) = P(A)$ ，那么 $P(B|A) = P(B)$ 一定成立。从直观上说，如果“烟”能够提供一些关于“火”的信息，那么“火”一定会能提供一些关于“烟”的信息。

给定第三个事件C，如果

$$P(A \mid B, C) = P(A \mid C) \quad (1.2)$$

并且 $P(B \mid A, C) = P(B \mid C)$ 。则称两个事件A和B条件独立。例如，事件“烟雾检测器报警”与事件“附近有火”相关。但在第三个事件“附近有烟雾”的条件下，这两个事件可能变为独立的，烟雾检测器只对烟雾的存在做出响应，而不是对产生烟雾的原因做出响应。在处理数据集或列联表时，如果事件A和B在对C过滤所产生的新数据集中是独立的，则称在给定C的条件下，A和B是条件独立的。如果A和B在原始的、未经过滤的数据集中过滤后是独立的，则称它们为边独立的。

变量与事件类似，相互之间也存在相关或独立关系。对于X和Y的每一个取值x和y，如果有

$$P(X = x \mid Y = y) = P(X = x) \quad (1.3)$$

则称两个变量X和Y是独立的。与事件的独立性一样，变量的独立性也具有对称关系，因此式(1.3)意味着 $P(Y = y \mid X = x) = P(Y = y)$ 成立。若对X和Y的某一对取值，式(1.3)不成立，则称X和Y相关。从这个意义上说，变量的独立性可以理解为一组事件的独立性。例如，“高度”和“音乐才能”是两个独立的变量，对每一个高度h和音乐水平m，一个人身高h英尺的概率不会因为发现他有音乐水平m而改变。

1.3.5 概率分布

变量X的稳定率分布是X的每一个可能取值的概率的集合。例如，如果X能取三个值：1、2和3，则X的一种可能的概率分布是 $P(X = 1) = 0.5, P(X = 2) = 0.25, P(X = 3) = 0.25$ 。在概率分布中，概率的值必定为 $0 \sim 1$ ，并且所有可能取值的概率和必为1。概率为0的事件是不可发生事件，概率为1的事件是必然事件。

连续变量也有概率分布，连续变量 X 的概率分布用密度函数 f 来表示。当将密度函数 f 绘制到坐标平面上时，变量 X 的值在 a 和 b 之间的概率是曲线下 a 与 b 之间的面积，用积分表示该面积为 $\int_a^b f(x)dx$ 。整个曲线下的面积，即 $\int_{-\infty}^{+\infty} f(x)dx$ 必定为1。

变量集也有概率分布，称为联合分布。一组变量 V 的联合分布是 V 中变量值的每一种可能组合的概率的集合。例如，如果 V 中有两个变量 X 和 Y ，每个变量均可取两个值：1和2，那么， V 的一种可能的联合分布是： $P(X = 1, Y = 1) = 0.2$, $P(X = 1, Y = 2) = 0.1$, $P(X = 2, Y = 1) = 0.5$, $P(X = 2, Y = 2) = 0.2$ 。与单变量概率分布一样，联合分布的概率之和也一定是1。

1.3.6 全概率公式

下面介绍几个非常有用的概率论定理。首先，对于任何两个互斥事件A和B（即A和B不能同时发生），有

$$P(A + B) = P(A) + P(B) \quad (1.4)$$

由此，对于任何两个事件A和B，有

$$P(A) = P(A, B) + P(A, \bar{B}) \quad (1.5)$$

因为事件“ A 与 B ”与“ A 与 \bar{B} ”是互斥的——如果 A 为真，那么“ A 与 B ”或“ A 与 \bar{B} ”必定有一个为真。例如，“Dana是位高个子男性”与“Dana是位高个子女性”是互斥的，如果Dana是高个子，那么他必定要么是高个子男性，要么是高个子女性，因此，有

$$P(\text{Dana是个高个子}) = P(\text{Dana是个高个子男性}) + P(\text{Dana是个高个子女性})。$$

更一般地，如果任意一组事件 B_1, B_2, \dots, B_n ，其中恰好有一个事件必定为真（穷举的互斥集称为一个划分），那么

$$P(A) = P(A, B_1) + P(A, B_2) + \dots + P(A, B_n) \quad (1.6)$$

式 (1.6) 称为全概率公式，将其应用于日常生活中的实际例子，便可很容易地验证式 (1.6) 成立：例如，从整副牌中随机抽取一张，那么它正好是 J 的概率等于它是红桃 J 的概率加上梅花 J 的概率，再加上方块 J 和黑桃 J 的概率。事件 A 与每个 B_i 的联合概率称为 A 在 B_i 上的边缘概率，而对 A 在所有 B_i 上的边缘概率求和得到 $P(A)$ ，称为 A 的边缘概率。

如果已知 B 的概率以及在 B 条件下 A 的概率，可以通过简单的乘积推导求出 A 与 B 同时发生的概率：

$$P(A, B) = P(A | B)P(B) \quad (1.7)$$

例如，乔既风趣又聪明的概率等于一位聪明人风趣的概率，乘以乔聪明的概率。

除法法则

$$P(A | B) = \frac{P(A, B)}{P(B)}$$

被看作条件概率的形式化定义，正如表 1.3 和表 1.4 所示的案例，把条件视为过滤操作即可验证该法则的合理性。当以 B 为条件时，从原始联合表中去掉所有与 B 冲突的事件，得到的子表也表示一个概率分布，并且与所有概率分布一样，它的和必定为 1。由于子表中的每行在原表分布中的概率之和为 $P(B)$ （根据定义），那么可以通过每行乘以 $\frac{1}{P(B)}$ 来确定它们在新表中的分布概率。

式 (1.7) 蕴含了独立性的定义，到目前为止，对这一概念的使用是非正式的，意味着“没有给出额外的信息”。在概率分布中，它有一种数值形式的表示，即若事件 A 和 B 独立，则有：

$$P(A, B) = P(A)P(B)$$

例如，要检查两枚银币的结果是否真正独立，应该计算它们同时呈现背面的频次，并确定它等于每枚银币呈现背面的频次的乘积。

由式(1.7)及对称性 $P(A, B) = P(B, A)$ ，可以直接得到概率论中最重要的定理之一——贝叶斯法则：

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)} \quad (1.8)$$

借助式(1.7)中的乘法法则，可以把全概率公式表示为条件概率的加权和：

$$P(A) = P(A | B_1)P(B_1) + P(A | B_2)P(B_2) + \cdots + P(A | B_n)P(B_n) \quad (1.9)$$

式(1.9)非常有用，因为我们常常发现无法直接计算 $P(A)$ ，但通常情况下更容易计算条件概率，比如 $P(A | B_k)$ ，它与特定的背景条件有关，而不是与背景条件无关的 $P(A)$ 。例如，假设有来自两个产地的一批小零件，其中 30% 是工厂 A 生产的，次品率是 $\frac{1}{5,000}$ ；70% 是工厂 B 生产的，次品率是 $\frac{1}{10,000}$ ，求一个随机选择的小零件是次品的概率。这是一个伤脑经的问题，但是按照式(1.9)分解就变得容易了，有

$$\begin{aligned} P(\text{次品}) &= P(\text{次品} | A)P(A) + P(\text{次品} | B)P(B) \\ &= \frac{0.30}{5000} + \frac{0.70}{10000} = \frac{1.30}{10000} = 0.00013 \end{aligned}$$

或者举一个稍难的例子，假设掷两个骰子，想知道掷第二个骰子比第一个骰子点数大的概率，即求 $P(A) = P(\text{掷}2 > \text{掷}1)$ 。没有一个直观的方法可以立即算出这个概率。如果以第一个骰子的点数为条件，将其分解到条件 B_1, B_2, \dots, B_n ，就容易解决了，即有

$$\begin{aligned} P(\text{掷}2 > \text{掷}1) &= P(\text{掷}2 > \text{掷}1 \mid \text{掷}1 = 1)P(\text{掷}1 = 1) + P(\text{掷}2 > \text{掷}1 \mid \text{掷}1 = 2)P(\text{掷}1 = 2) + \dots \\ &= \left(\frac{5}{6} \times \frac{1}{6}\right) + \left(\frac{4}{6} \times \frac{1}{6}\right) + \left(\frac{3}{6} \times \frac{1}{6}\right) + \left(\frac{2}{6} \times \frac{1}{6}\right) + \left(\frac{1}{6} \times \frac{1}{6}\right) + \left(\frac{0}{6} \times \frac{1}{6}\right) \\ &= \frac{5}{12} \end{aligned}$$

式(1.9)所给出的分解法有时也称为“替代率”或“扩展法”，在本书中，我们将其称为条件化B。

1.3.7 使用贝叶斯法则

在使用贝叶斯法则时，有时把事件 A 称为假设，事件 B 称为证据。这种命名反映了贝叶斯定理的核心目的。在许多情况下，已知或可以轻易确定（假设 A 正确的情况下，证据 B 发生的概率），但很难计算（在得到证据 B 的情况下，假设 A 发生的概率）。然而，在现实世界中，后者是人们经常需要回答的问题。一般来说，在证据 B 发生后，希望将某些假设的信任度由 $P(A)$ 更新为 $P(A | B)$ 。在这种情况下，为了精确地使用贝叶斯法则，必须将某一个假设当作一个事件，并对所有假设赋予一个事先的概率分布，称为先验分布。

例如，假设在赌场中听到庄家喊出“11”。你恰好知道，该事件只能由骰子赌局和轮盘赌局这两种游戏引发，并且在任意时刻，这两种游戏正在进行的赌局数量完全相等。那么，在听到庄家报“11”的情况下，他正在进行骰子赌局的概率是多少呢？

在这种情况下，“骰子赌”是假设，而“11”是证据。要立即计算出这个概率是有困难的，但反过来，一轮骰子赌的结果是“11”的概率却很容易计算。骰子赌是一种游戏，掷两个骰子后计算点数的和。显然，两个骰子点数和为11的可能性是 $\frac{2}{36} = \frac{1}{18}$ ， $P(11 | \text{骰子赌}) = \frac{1}{18}$ 。在轮盘赌中，有等概率的38种结果，因此 $P(11 | \text{轮盘赌}) = \frac{1}{38}$ 。在此情况下，有两种可能的假设：“骰子赌”和“轮盘赌”，因为他们的赌局数相等， $P(\text{骰子赌}) = P(\text{轮盘赌}) = \frac{1}{2}$ ，这些是对两种赌局的先验知识。有全概率公式，

$$\begin{aligned} P(11) &= P(11 | \text{骰子赌})P(\text{骰子赌}) + P(11 | \text{轮盘赌})P(\text{轮盘赌}) \\ &= \frac{1}{2} \times \frac{1}{18} + \frac{1}{2} \times \frac{1}{38} = \frac{7}{171} \end{aligned}$$

现在已经很容易地获取了确定 $P(\text{骰子赌} | 11)$ 所需要的全部信息：

$$P(\text{骰子赌} | 11) = \frac{P(11 | \text{骰子赌}) \times P(\text{骰子赌})}{P(11)} = \frac{1/18 \times 1/2}{7/171} = 0.679$$

贝叶斯法则的另一个应用例子是蒙提霍尔问题（即三门问题），这是统计学中一个经典的脑筋急转弯问题。在这个问题中，你是一位由蒙提霍尔主持的游戏节目中的参赛者。蒙提向你展示三扇门 A、B 和 C，其中只有一扇门后面有一辆新车（另外两扇门后面是山羊）。如果你选对了车的门，车就归你；否则，你得到一只山羊。你随便猜测了一扇门，比如 A，然后在没有任何暗示的情况下，蒙提打开了 C 门，C 门后面是山羊。他告诉你，现在你可以换到 B，或者坚持选择 A。无论你选哪一扇门，你都将得到它后面的东西。

你的最好策略是坚持打开A门，还是换到B门？

当第一次遇到这个问题时，很多人认为，由于车的位置与你第一次选择门之间是独立的，换门既没有好处也没有坏处；车在A门后面的概率等于在B门后面的概率。

但是多年以后，统计学的学生惊愕地发现，正确的答案为，与坚持A门相比，如果换成B门，赢得汽车的可能性会变为原来的两倍。对这个有些违背直觉的结果通常的解释是，当最初选择一扇门时，你有 $\frac{1}{3}$ 的概率选到有车的那扇门。因为不管最初你是否选中了有车的那扇门，蒙蒂总是打开有山羊的那扇门，之后你没有接收到任何新信息。因此，你选择的那扇门有车的概率仍然是 $\frac{1}{3}$ ，其余 $\frac{2}{3}$ 的概率必定属于剩下的另一扇关闭的门。

可以用贝叶斯法则来证明这一令人惊讶的事实：这里有三个变量： X ，参赛者选择的门； Y ，后面藏有汽车的门； Z ，蒙蒂打开的门。 X 、 Y 和 Z 均可取值 A、B 或 C。现在要证明 $P(Y = B | X = A, Z = C) > P(Y = A | X = A, Z = C)$ 。先验假设是车在 A 门后面，证据是蒙蒂打开了 C 门。我们把证明留给读者（见思考题 1.3.5）。为进一步加强直觉认识，可以把游戏推广到有 100 扇门（1 扇门后有汽车，99 扇门后有山羊）的情况。参赛者仍然选择一扇门，但这次蒙蒂打开了 98 扇门——所有刻意打开的门后面都是山羊——在最后两扇门之间，给参赛者提供是否更换选择的机会。现在，选择更换应该是显而易见的。

为什么蒙蒂打开 C 门构成了改变车位置的证据呢？毕竟它没有为你最初的选择是否正确提供任何证据。当他准备打开某扇门的时候，不管是 B 还是 C，你当然知道它的后面没有车。答案是，在你选择 A 门之后，蒙蒂不可以再打开它，但他本来可以打开其他门。但他没有这样做，这意味着他打开 C 门更有可能是因为被迫这样做：这提供了汽车在 B 门后面的证据。这是贝叶斯分析的一般性原理：任何经得起反驳的假设都会变得更有可能。门容易被反驳（即蒙蒂本来可以打开它），但 A 门不是。因此，B 门变成了更可能的位置。

读者可能会发现，上面的解释充满了反事实术语，例如，“他本来可以打开”“因为他可能被迫”“他本来打算打开”。事实上，**蒙蒂霍尔问题** 在概率游戏中如此特别的原因就在于其 对数据生成过程的高度依赖。它表明我们的概念不应该仅依赖于观察到的事实，还应该依赖于产生这些事实的过程。特别地，“汽车不在 C 门后”的信息本身，不足以说明这个问题。为了计算所涉及的概率，我们也必须知道，在打开 C 门之前主持人有哪些选项。本书第 4 章将形式化地陈述反事实理论，以便描述这样的过程和可能的选择，形成关于选择的正确概念。

有一些质疑贝叶斯法则的争议。当试图在给定一些证据条件下确定假设的概率时，通常无法根据情境的一部分或发生的频次来计算先验概率 $P(A)$ 。思考一下，如果不知道赌场里轮盘赌桌与骰子赌桌的比例，如何确定先验概率 $P(\text{骰子赌})$ 呢？也许会尝试设定 $P(A) = \frac{1}{2}$ ，用以表达一种无所知的情境。但如果获知了一些信息，比如轮盘赌桌在这个赌场里并不常见，或者报数人的声调让我们联想起昨天听到的骰子赌的庄家，这会如何影响结果？在这些情况下，为了使用贝叶斯法则，可以将 $P(A)$ 设置为对假设的主观确信度，这比其他概率值更为可靠。然而，对于贝叶斯法则的争论源于这种确信度的主观性质。例如，如何确定指定的 $P(A)$ 准确地概括了关于这个假设的已知信息？是否应该坚持将对某个问题的所有正反观点归结为一个数字？即使这样做了，为什么又要根据客观事件的发生频次来更新假设的主观确信度呢？一些行为实验表明，人们不会依照贝叶斯法则更新他们的信念，尽管很多人相信他们应该这样做。如果不是推理上有问题，那么这种偏离

规则的行为可能代表了一种妥协，并导致次优决策。关于如何恰当使用贝叶斯定理的争论一直持续到今天。然而，尽管存在这些争议，贝叶斯法则仍然是统计学中一个强大的工具。本书中将利用它来实现我们的目标。

思考题

1.3.3

思考第 1.3.6 节描述的赌场问题。

(a) 假设赌场中轮盘赌桌是骰子赌桌的两倍，计算 $P(\text{骰子赌} | 11)$ 。

(b) 假设赌场中骰子赌桌是轮盘赌桌的两倍，计算 $P(\text{骰子赌} | 10)$ 。

1.3.4

假设有三张卡片，卡片1正反面各一个黑脸，卡片2正反面各一个白脸，卡片3正反面分别为一个白脸和一个黑脸。你随机选择张卡片并放在桌子上，发现面朝上的是黑脸。那么，这张卡片明下的一面也是黑脸的概率是多少？

(a) 根据你的直觉说明，这张卡片朝下的一面也是黑脸的概率是 $\frac{1}{2}$ 。为什么它也许大于 $\frac{1}{2}$ ？

(b) 根据下列的变量，给出你认为容易估计的概率和条件概率，例如， $P(C_D = \text{黑脸})$ 。

$$I = \text{选择卡的标识}(\text{卡1}, \text{卡2}, \text{卡3})$$

$$C_D = \text{朝下一面的颜色}(\text{黑脸}, \text{白脸})$$

$$C_U = \text{朝上一面的颜色}(\text{黑脸}, \text{白脸})$$

使用你上面的估计，找出选择卡片朝下一面是黑脸的概率。

(c) 对随机选择的卡片，如果你看到它的正面是黑脸，使用贝叶斯法则找出其背面是黑脸的正确概率。

1.3.5

使用贝叶斯法则证明，在蒙蒂霍尔问题中，换门提高你赢得汽车的概率。

1.3.8 期望值

在统计学中，常常遇到要处理的数据量和概率分布规模太大，以至于无法有效检测变量取值的所有可能组合的情况。作为替代，常以损失部分信息为代价，用统计量来表示分布的、有意义的特征。这其中最常用的是期望值，也称为均值，它可以用于变量取值为数值的情况。变量 X 的期望值表示为 $E(X)$ ，等于变量的每一个可能的取值与变量取这个值的概率的乘积，再对来积求和，即

$$E(X) = \sum_x x P(X = x) \tag{1.10}$$

例如，用变量 X 表示掷一次六面骰子的结果，概率分布如下：

$$P(1) = \frac{1}{6}, P(2) = \frac{1}{6}, P(3) = \frac{1}{6}, P(4) = \frac{1}{6}, P(5) = \frac{1}{6}, P(6) = \frac{1}{6}$$

则 X 的期望值为

$$E(X) = (1 \times \frac{1}{6}) + (2 \times \frac{1}{6}) + (3 \times \frac{1}{6}) + (4 \times \frac{1}{6}) + (5 \times \frac{1}{6}) + (6 \times \frac{1}{6}) = 3.5$$

同样地， X 的任意函数 $g(X)$ 的期望值，可通过在 X 的所有取值上对 $g(x)P(X = x)$ 求和得到，即：

$$E(g(x)) = \sum_x g(x)P(x) \quad (1.11)$$

例如，若掷一个骰子后能得到点数平方的现金奖励，即有： $g(X) = X^2$ ，则奖励的期望值为：

$$E(g(X)) = (1^2 \times \frac{1}{6}) + (2^2 \times \frac{1}{6}) + (3^2 \times \frac{1}{6}) + (4^2 \times \frac{1}{6}) + (5^2 \times \frac{1}{6}) + (6^2 \times \frac{1}{6}) = 1$$

还可以计算 Y 关于条件 X 的期望值 $E(Y | X = x)$ ，方法是对 Y 的每一个可能取值 y 乘以 $P(Y = y | X = x)$ 并求和，即

$$E(Y | X = x) = \sum_y yP(Y = y | X = x) \quad (1.13)$$

$E(X)$ 是估计 X 可能取值的一种方法。特别地，对于 g 的值的所有猜测中，选择 $g = E(X)$ 可使得方差期望 $E(g - X)^2$ 最小。类似地，假定观察到 $X = x$ ，则 $E(Y | X = x)$ 表示对 Y 的最佳估计。如果 $g = E(Y | X = x)$ ，那么使得方差期望 $E((g - Y)^2 | X = x)$ 最小。

注：可得到常用的概论论结论 $g = E(X) \quad s.t. \quad \min E(g - X)^2$ ，若要证明，不妨计算 $h(g) = E(g - X)^2$ ，后计算 $\frac{d}{dg}h(g) = 0$ ，即可得证。

如表 1.3 所示，2012 年美国总统选举选民的期望年龄为

$$E(\text{选民的年龄}) = 23.5 \times 0.16 + 37 \times 0.23 + 54.5 \times 0.39 + 70 \times 0.22 = 48.9$$

对于该计算，假定每个类别中的各年龄的可能性相同，例如，选民是18岁的可能性与25岁相同，30岁的可能性与44岁相同。还假设投票者的最大年龄是75岁。这意味着，随机猜测一个选民的年龄，如果与真实值相差 e 年，将会失去 e^2 美元，那么猜测48.9，则平均损失会最小。同样，如果要求去猜测一个小于45岁的随机选民的年龄，则最好的赌注是

$$P(\text{选民的年龄} \mid \text{选民的年龄} < 45) = 23.5 \times 0.40 + 37 \times 0.60 = 31.6 \quad (1.14)$$

用期望作为预测或“最佳猜测”的依据，在很大程度上取决于 X 或 $(Y \mid X = x)$ 分布的假设，即它们的分布是近似对称的。然而，如果感兴趣的分布是高度偏态的，那么其他预测方法可能更好。例如，在这种情况下，可以使用 X 分布的中值作为“最佳估计”，这个估计使期望的绝对误差 $E(|g - X|)$ 最小，这里不再进一步探讨这种方法。

注：可以得到的结论为若对一个分布估计为 **中值**，则满足 $\min E(|g - X|)$ 。

1.3.9 方差和协方差

变量 X 的方差记为 $Var(X)$ 或 σ_X^2 ，用于粗略地衡量 X 的值在数据集或群体中“偏离”平均值的情况。如果 X 的值均处于一个值附近，则方差相对较小；如果 X 的值覆盖了较大的范围，方差则相对较大。在数学上，将变量的方差定义为变量与均值的平均平方差。可先找出均值 μ ，然后依据如下公式计算得到：

$$\sigma_X^2 = E((X - \mu)^2) \quad (1.15)$$

随机变量 X 的标准差 σ_X 是其方差的平方根。不同于方差， σ_X 与 X 有相同的单位表示。例如，根据表 1.3，小于 45 岁选民的年龄分布的方差，可以很容易地根据式 (1.15) 算出：

$$\begin{aligned} \sigma_X^2 &= ((23.5 - 31.6)^2 \times 0.40) + ((37 - 31.6)^2 \times 0.60) \\ &= (65.61 \times 0.40) + (29.16 \times 0.60) \\ &= 26.24 + 17.5 = 43.74 \end{aligned}$$

而标准差是

$$\sigma_X = \sqrt{43.74} = 6.61$$

这意味着，随机选择一个选民，他的年龄与平均年龄 31.6 岁相比，偏离值在 6.61 岁以内的概率很高。这种解释可以量化。例如，对于正态分布的随机变量 X ，其中约 2/3 的值落在期望值或均值的一个标准差内，并且，大约 95% 的值落在偏离均值的两个标准差内。

乘积 $(X - E(X))(Y - E(Y))$ 的期望具有特别的重要性，它称为 X 和 Y 协方差，记为 $Cov(X, Y)$ 或 σ_{XY} ：

$$\sigma_{XY} \triangleq E((X - E(X))(Y - E(Y))) \quad (1.16)$$

它用于衡量 X 和 Y 共变的程度，即这两个变量一起变化的程度，或称为“相关”。这种对相关性的度量实际上反映了 X 和 Y 共变的特定方式，它体现了 X 和 Y 线性共变的程度。读者可以把这看作是绘制点 (X, Y) ，然后通过一条直线来刻画 Y 随 X 的变化而变化的程度。

协方差 σ_{XY} 通常被归一化为相关系数

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad (1.17)$$

这是一个标量值，范围为 -1 到 1，表示通过各自的标准差归一化 X 和 Y 后，最佳拟合直线的斜率。相关系数 ρ_{XY} 等于 1 当且仅当一个变量能以线性方式预测另一个变量时；如果没有这样的线性预测，相关系数 ρ_{XY} 为 0。无论什么情况下，这样的预测方法都不如随机估计效果好。相关系数 σ_{XY} 和 ρ_{XY} 的含义将在后面讨论。这里需要注意的是，共变的程度可以利用式 (1.16) 和式 (1.17) 根据联合分布 $P(X, Y)$ 计算得出。此外，当 X 和 Y 相互独立时， σ_{XY} 和 ρ_{XY} 都为 0。还需特别注意， Y 和 X 的非线性关系无法通过简单的数值来刻画，需要借助完全明确化的条件概率 $P(Y = y | X = x)$ 。

思考题

1.3.6

- (a) 证明：当 X 和 Y 相互独立时， σ_{XY} 和 ρ_{XY} 都为 0。[提示：使用式(1.16)和式(1.17)。]
- (b) 给两个变量高度依赖，但它们的协相关系数仍然为 0 的例子。

1.3.7

同时抛掷两枚硬币来确定乡镇俱乐部两名玩家的收益。当且仅当至少一枚硬币正面朝上时，1号玩家获得1美元。当且仅当两枚硬币落地面相同时，2号玩家得到1美元。现 X 表示1号玩家的收益， Y 表示2号玩家的收益。

- (a) 指出并描述概率分布：

$$P(x), P(y), P(x, y), P(y | x) \text{ 和 } P(x | y)$$

- (b) 使用(a)中的描述，计算下面的度量：

$$E(X), E(Y), E(Y | X = x), E(X | Y = y), \sigma_X^2, \sigma_Y^2, \sigma_{XY}, \rho_{XY}$$

(c) 如果 2号玩家获得 1元, 则 1号玩家收益的最佳估计是多少?

(d) 如果 1号玩家获得 1元, 则 2号玩家收益的最佳估计是多少?

(e) 存在相互独立的两个事件 $X = x$ 和 $Y = y$ 吗?

1.3.8

一个掷双骰子游戏 (一次掷两个独立的骰子), 其中, X 表示骰子1投掷结果, Z 表示骰子2投掷结果, Y 表示骰子1和骰子2投掷结果和。计算下面理论估计的结果。

表1.6描述了12轮掷双骰子游戏的结果。

(a) $E(X)$, $E(Y)$, $E(Y | X = x)$, $E(X | Y = y)$ (对 x 和 y 的每一个值), 以及 σ_X^2 , σ_Y^2 , σ_{XY} , ρ_{XY} , σ_{XZ} 。

(b) 根据表1.6中的数据, 找出(a)中各个参数的样本估值。[提示: 使用软件包。]

(c) 假设测得 $X = 3$, 利用 (a) 中的结果, 确定 Y 的最佳估计。

(d) 假设测得 $Y = 4$, X 的最佳估计是什么?

(e) 假设测得 $Y = 4$ 和 $Z = 1$, X 的最佳估计是什么? 说明为什么它与(d)的结果不一样。

轮次	X	Z	Y
第1轮	6	3	9
第2轮	3	4	7
第3轮	4	6	10
第4轮	6	2	8
第5轮	6	4	10
第6轮	5	3	8
第7轮	1	5	6
第8轮	3	5	8
第9轮	6	5	11
第10轮	3	5	8
第11轮	5	3	8
第12轮	4	5	9

1.3.10 回归

在统计学中, 通常希望根据一个变量 X 的值来预测另一个变量 Y 的值。例如, 根据年龄来预测一位学生的身高。通过前面的介绍可以看出, Y 基于 X 的最佳预测由条件期望 $E(Y | X = x)$ 给出, 但这需要假定已知条件期望, 或者能从联合分布 $P(y, x)$ 中计算出条件期望。而使用回归方法, 则可以直接从数据中做出预测。我们试图找到一个函数, 通常是线性函数, 这个函数以 X 的观测值作为输入, 以 Y 的值作为输出, 使得 Y 的预测值 (输出) 与实际值的平方误差量最小。

下面从一个散点图开始，将数据集的每种情况都画在同一个坐标平面上，其中，预测变量或输入变量位于横轴，预测结果变量位于纵轴上，如图1.3所示。

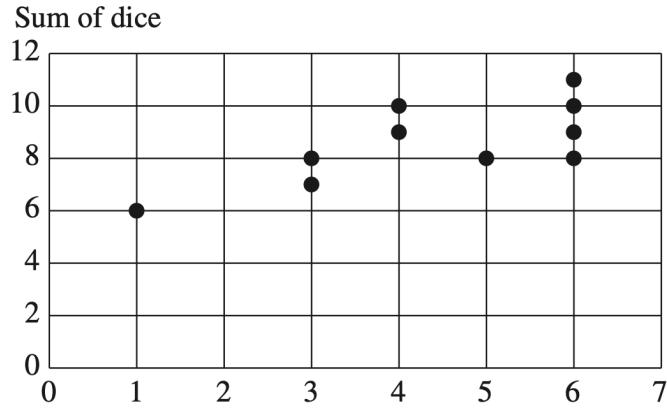


Figure 1.3 Scatter plot of the results in Table 1.6, with the value of Die 1 on the x -axis and the sum of the two dice rolls on the y -axis

最小二乘回归线是使得散点图上的点到这条线的垂直距离的均方差量最小的直线，也就是说，如果散点图上有 n 个数据点 (x, y) ，对于任何一个数据点 (x_i, y_i) ，直线 $y = \alpha + \beta x$ 在 x_i 的值用 y'_i 表示，那么最小二乘回归线就是使式 (1.18) 的值最小的直线：

$$\sum_i (y_i - y'_i)^2 = \sum_i (y_i - \alpha - \beta x_i)^2 \quad (1.18)$$

下面分析斜率 β 与概率分布 $P(x, y)$ 是如何关联的。假设连续投掷12轮骰子，得到如表1.6所示结果。根据该数据，根据骰子 1(X) 的值来预测两个骰子的点数和 Y ，图1.3为该数据的散点图。对于该例子，最小二乘回归线如图1.4所示，虚线表示样本数据的最佳拟合直线，实线表示总体数据的最佳拟合直线。注意，我们使用的样本数据的回归线不必与总体数据的回归线相同。当允许样本量增加到无限大时，将得到总体数据的回归线。图1.4中的实线表示理论上的最小二乘线，由式 (1.19) 给出：

$$y = 3.5 + 1.0x \quad (1.19)$$

虚线表示样本的最小二乘线，由于抽样的原因，在斜率和截距两方面均与理论直线存在差异。

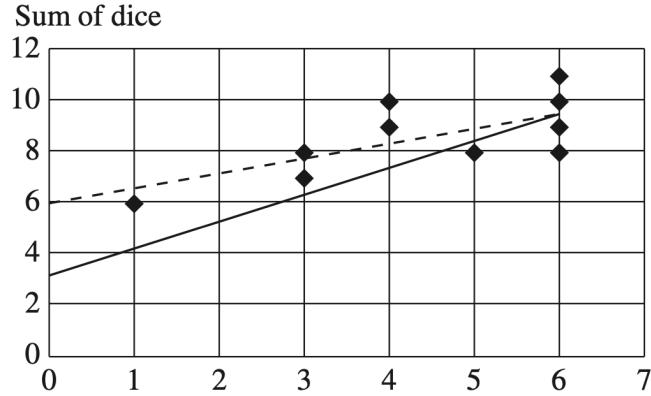


Figure 1.4 Scatter plot of the results in Table 1.6, with the value of Die 1 on the x -axis and the sum of the two dice rolls on the y -axis. The dotted line represents the line of best fit based on the data. The solid line represents the line of best fit we would expect in the population

因为已知给定第一个骰子值为 x 的条件下两个骰子和的期望值，因此在图1.4中，能够得到总体数据的回归线方程。计算比较简单：

$$E(Y | X = x) = E(\text{骰子}2 + X | X = x) = E(\text{骰子}2) + x = 3.5 + 1.0x$$

这个结果并不奇怪，因为 Y （两个骰子之和）可以写成

$$Y = X + Z$$

其中 Z 是骰子2的投掷结果。容易看出，如果 X 增加一个单位，比如从 $X = 3$ 增加到 $X = 4$ ，那么 $E(Y)$ 也将同样增加一个单位。然而，读者可能有点奇怪，这个结果反过来则不成立， X 关于 Y 的回归线的斜率不为1.0。为了解释原因，有

$$E(X | Y = y) = E(Y - Z | Y = y) = y - E(Z | Y = y) \quad (1.20)$$

需要注意的是项 $E(Z | Y = y)$ ，因为它（线性）依赖于 y ，因此 $y - E(Z | Y = y) < y$ ，从而使斜率小于1.0。为了计算 $E(X | Y = y)$ 的准确值，根据 X 和 Z 的对称性，有

$$E(X | Y = y) = E(Z | Y = y)$$

带入公式 (1.20) 后，得到

$$E(X | Y = y) = 0.5y$$

这里斜率减小的原因是，当给 Y 增加1个单位时，平均来说，每一个 X 和 Z 对此的贡献相同，这是与直觉相符的。当观察到两个极值的和是 $Y = 10$ 时，对每个数值的最佳估计是 $X = 5$ 和 $Z = 5$ 。

一般来说，如果写出 Y 关于 X 的回归方程为

$$y = a + bx \quad (1.21)$$

斜率 b 由 R_{YX} 表示，它可以用协方差 σ_{XY} 来计算如下：

$$b = R_{YX} = \frac{\sigma_{XY}}{\sigma_X^2} \quad (1.22)$$

从这个方程可以清楚地看出， Y 关于 X 的斜率可能与 X 关于 Y 的斜率不一样，也就是说，在大多数情况下， $R_{YX} \neq R_{XY}$ ($R_{YX} = R_{XY}$ 仅当 X 的方差与 Y 的方差相等时成立)。回归线的斜率可以为正、负或零。如果为正，称 X 和 Y 正相关，意思是随着 X 值的增大， Y 的值也增大；如果为负，称 X 和 Y 负相关，意思是随着 X 值的增大， Y 的值减小；如果为零（一条水平线），称 X 和 Y 线性无关，这意味着，至少从线性关系上来说，知道 X 的值并不能有助于预测 Y 的值。如果两个变量是相关的，无论是正相关还是负相关（或其他方式），它们都是相互依赖的。

1.3.11 多元回归

用多元线性回归也可以实现对多个变量进行回归分析。例如，如果想用变量 X 和 Z 的值来估计变量 Y 的值，则可以在 $\{X, Z\}$ 上进行 Y 的多元线性回归，并估计一个回归关系。

$$y = r_0 + r_1x + r_2z \quad (1.23)$$

它表示三维坐标系中的一个斜面。

可以创建一个三维散点图， Y 、 X 和 Z 的值分别表示在纵轴、横轴和竖轴上。然后，沿竖轴将散点图切片。每个切片将构成如图1.4所示的那种二维散点图。每个二维散点图将有一个斜率为 r_1 的回归线。沿横轴的切片给出的斜率是 r_2 。

当保持 Z 不变时， Y 关于 X 的斜率称为偏回归系数，表示为 $R_{YX.Z}$ 。注意当 R_{YX} 为正时， $R_{YX.Z}$ 可能为负，如图1.1所示。这从普森悖论中可以看出， Y 与 X 之间总是正相关的，但当增加第三个变量 Z 为条件时， Y 与 X 就变成负相关了。

利用一个定理来极大地简化偏回归系数的计算，例如式 (1.23) 中的 r_1 和 r_2 ，该定理是回归分析中最基础的定理之一。这个定理表述为，如果 Y 是变量 X_1 、 X_2 、 X_3 与噪声项 ϵ 的线性组合，

$$Y = r_0 + r_1 X_1 + r_2 X_2 + \cdots + r_k X_k + \epsilon \quad (1.24)$$

那么，不管 Y, X_1, X_2, \dots, X_k 的实际分布如何，当 ϵ 与每一个回归元 X_1, X_2, \dots, X_k 都不想关时，即

$$\text{Cov}(\epsilon, X_i) = 0 \quad \text{for } i = 1, 2, \dots, k$$

可以得到最佳的最小二乘系数。

这一条件称为正交性原理，下面解释如何利用它来简化计算。假设想要在已知两个骰子点数之和 $Y = \text{骰子1} + \text{骰子2}$ 的条件下，计算骰子1的最优估计，记

$$X = \alpha + \beta Y + \epsilon \quad (1.25)$$

我们的目标是按照可估计的统计度量找到 α 和 β 。假设不失一般性，令 $E(\epsilon) = 0$ ，在方程的两边取期望值，得到

$$E(X) = \alpha + \beta E(Y) \quad (1.26)$$

将式 (1.25) 两边乘以 Y ，并取期望值，得

$$E(XY) = \alpha E(Y) + \beta E(Y^2) + E(Y\epsilon) \quad (1.27)$$

根据正交性原理得到 $E(Y\epsilon) = 0$ ，由式 (1.26) 和式 (1.27) 得到关于两个未知数 α 和 β 的两个方程，求解方程得到：

$$\begin{aligned} \alpha &= E(X) - E(Y) \frac{\sigma_{XY}}{\sigma_Y^2} \\ \beta &= \frac{\sigma_{XY}}{\sigma_Y^2} \end{aligned}$$

推导完毕。斜率 β 其实已经可以从式 (1.22) 通过互换 X 与 Y 得到，但上面的推导展示了一个在二维或多维情况下计算斜率的一般性方法。

作为例子，考虑给定两个观测值 $X = x$ 和 $Y = y$ 时，找出 Z 的最优估计问题。如前所述，写出回归方程

$$Z = \alpha + \beta_Y Y + \beta_X X + \epsilon$$

但现在，为得到关于 β 、 β_Y 和 β_X 的三个方程，也需要两边乘以 Y 和 X ，并求期望。利用正交条件 $E(\epsilon Y) = E(\epsilon X) = 0$ ，解得到的方程，得

$$\beta_Y = R_{ZY \cdot X} = \frac{\sigma_X^2 \sigma_{ZY} - \sigma_{ZX} \sigma_{XY}}{\sigma_Y^2 \sigma_X^2 - \sigma_{YX}^2} \quad (1.28)$$

$$\beta_X = R_{ZX \cdot Y} = \frac{\sigma_Y^2 \sigma_{ZX} - \sigma_{ZY} \sigma_{YX}}{\sigma_Y^2 \sigma_X^2 - \sigma_{YX}^2} \quad (1.29)$$

式 (1.28) 和式 (1.29) 具有通用性；对于任何三个变量，这两个公式以方差和协方差的形式给出了线性回归系数 $R_{ZY \cdot X}$ 和 $R_{ZX \cdot Y}$ 。并且同样地，它们也反映了这些斜率受模型其他参数影响的敏感程度。然而，在实际求解中，回归斜率是通过有效的“最小二乘”算法从样本数据中估计出来的，一般很少通过数学方程直接进行演算。一个例外是在获得数据之前，需要估计这些斜率中是否有一个为零。当考虑为某种情况选择一组回归变量时，这样的预测很重要，正如后面将在第3.8节中提到的，这项任务可利用因果图进行非常有效的处理。

思考题

1.3.9

(a) 使用正交性原理证明式 (1.22) [提示：参照式 (1.27) 的处理方式。]

(b) 为思考题1.3.8中描述的骰子游戏，计算所有的偏回归系数。[提示：应用式 (1.28) 及思考题 1.3.8 (a) 中计算得到的方差和协方差。]

$$R_{YX \cdot Z}, R_{XY \cdot Z}, R_{YZ \cdot X}, R_{ZY \cdot X}, R_{XZ \cdot Y} \text{ 和 } R_{ZX \cdot Y}$$

1.4 图

从辛普森悖论可知，某些决策无法仅从数据本身获得有效信息，而要依赖于 **数据背后的原因**。本节介绍一种表述这些原因的数学工具——图论。在高中数学中一般并不讲授图论的内容，但作为一种有用的数学语言，它能像解决算术问题一样，用简单的运算解决因果问题。

虽然“图”这个词通常来说或多或少与“图表”这个词相当，用于指代一大类视觉对象，但在数学中，图是一个形式化定义的对象。数学中的图是由顶点（或称为节点）和边组成的集合，图中的节点由边相连接（或不连接）。图1.5给出了一个简单的图，X、Y 和 Z（圆点）是节点，A 和 B（直线）是边，节点 X 和 Y 相邻，节点 Y 和 Z 相邻，节点 X 和 Z 不相邻。

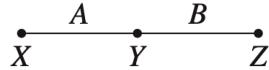


Figure 1.5 An undirected graph in which nodes X and Y are adjacent and nodes Y and Z are adjacent but not X and Z

如果两个节点间有边连接，则称这两个节点为相邻节点。在图1.5中， X 和 Y 是相邻节点， Y 和 Z 是相邻节点。若图中的每一对节点都有边连接，这种图称为完全图。

节点 X 和 Z 之间的路径是指从 X 开始，到 Z 结束的一系列由边首尾连接的节点。例如，在图1.5中，由于 X 和 Y 相连，而 Y 和 Z 相连，因此从 X 到 Z 有一条路径。

图中的边可以是有向的，也可以是无向的。图1.5中的两条边都是无向边，因为它们没有指定的输入端与输出端。有向边是指以一个节点指向另一个节点的边，并用箭头指示方向。所有边都有方向的图称为有向图。图1.6展示了一个有向图，其中 A 是从 X 到 Y 的有向边， B 是从 Y 到 Z 的有向边，节点 X 是 Y 的父节点，节点 Y 是 Z 的父节点。



Figure 1.6 A directed graph in which node X is a parent of Y and Y is a parent of Z

有向边的起始节点称为其终止节点的父节点；反之，有向边的指向节点称为其起始节点的子节点。在图1.6中， X 是 Y 的父节点， Y 是 Z 的父节点；相应地， Y 是 X 的子节点， Z 是 Y 的子节点。

如果两个节点间的路径能沿着箭头方向追踪，那么这条路径称为有向路径。也就是说，路径上不存在该路径上的某条边同时指向的节点，也不存在两条边共同起始的节点。如果两个节点由一条有向路径连接，那么，第一个节点称为该路径上所有节点(除自己外)的祖先节点，路径上的其他每个节点(除第一个节点)都是第一个节点的后代。类似于父节点与子节点的关系：父节点是其子节点的祖先，也是子节点的子节点的祖先，以此类推。例如，在图1.6中， X 是 Y 和 Z 的祖先， Y 和 Z 都是 X 的后代。

当一个节点存在返回自身的有向路径时，这个路径（或图）称为环（有环图），没有环的有向图称为无环图。例如，图1.7（a）是无环图，而图1.7（b）是有环图。注意，在图1.7（a）中不存在从任何节点回到该节点自身的有向路径，而图1.7（b）中存在从 X 回到 X 的有向路径。

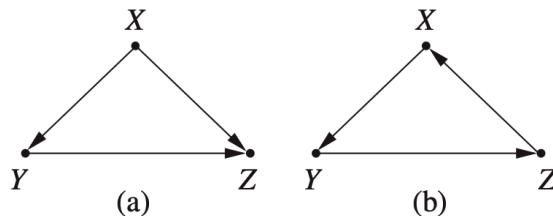


Figure 1.7 (a) Showing acyclic graph and (b) cyclic graph

思考题

1.4.1

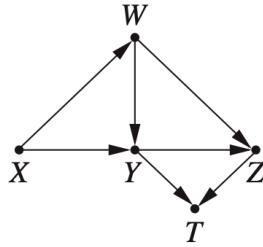


Figure 1.8 A directed graph used in Study question 1.4.1

如图1.8所示：

- (a) 找出Z的所有父节点。
- (b) 找出Z的所有祖先。
- (c) 找出W的所有子节点。
- (d) 找出X的所有后代。
- (e) 画出X到T的所有简单路径（即每个节点只出现一次）。
- (f) 画出X到T的所有有向路径。

1.5 结构因果模型

1.5.1 因果假设建模

为了能够严格地处理因果关系问题，需要寻找一种能够形式化表述数据背后因果假设的方法。为此，引入结构因果模型 (structural causal model, SCM) 的概念，用于描述现实世界关联特征及其相互作用。具体来说，结构因果模型描述了如何为感兴趣的变量赋值。

从形式上看，结构因果模型含有两个变量集 U 和 V ，以及一组函数

$$f = \{f_X : W_X \rightarrow X \mid X \in V\}$$

其中 $W_X \subseteq (U \cup V) - \{X\}$ ，即函数 f_X 根据模型中其他变量的值给变量 X 赋值。在这里，我们展开因果的定义：若 Y 存在于 f_X 后者的定义域中，则变量 Y 是变量 X 的直接原因。若 Y 是 X 的直接原因或原因的原因，则 Y 是 X 的原因。

U 中的变量称为外生变量，简单地说，它们属于模型的外部，不必解释它们变化的原因。 V 中的变量称为内生变量，模型中每一个内生变量都至少是一个外生变量的后代。外生变量没有祖先节点，因此不是任何其他变量的后代，特别地，外生变量不能是内生变量的后代，在图中表现为一个根节点。如果知道每个外生变量的值，那么利用函数 $f_X \in f$ ，就能完全确定每个内生变量的值。

例如，假设研究治疗方案 X 与哮喘患者的肺功能 Y 之间的因果关系。再假设 Y 也依赖于或归因于空气污染水平，由变量 Z 表示。在这种情况下，把 X 和 Y 看作内生变量， Z 看作外生变量。这是因为我们假设空气污染是一个外部因素，也就是说，它不能由个人特定的治疗方案或肺功能引发。

每一个结构因果模型 (SCM) 都与图形化的因果模型相关联，俗称“图模型”或简称“图”。图模型的节点表示 V 中的变量，节点之间的边表示图中的函数。设有SCM M 的图模型 G ， M 中的每个变量都表示为一个节点，对于变量 $X \in V$ ，假如 X 的定义域中包含变量 Y （即，如果 X 依赖于 Y 的值），那么在图 G 中，会有一条从 Y 到 X 的有向边。我们将主要讨论图模型是有向无环图 (Directed Acyclic Graph, DAG) 的结构因果模型，由于结构因果模型和图模型之间的这种关系，可以给出因果关系的图形化定义：在图模型中，如果变量 X 是另一个变量 Y 的子节点，那么 Y 是 X 的直接原因；如果 X 是 Y 的后代，那么 Y 是 X 的一个潜在原因（存在特殊的非传递的情况， Y 不是 X 的原因，我们将在第2章讨论）。

因果模型和图用这种方式将因果假设表示出来。例如，考虑下面简单的SCM：

SCM 1.5.1(学历、工龄和工资)

$$U = \{X, Y\}, \quad V = \{Z\}, \quad F = \{f_Z\}$$

$$f_Z : Z = 2X + 3Y$$

这个模型表示雇主付给一位学历为 X 、工龄为 Y 的员工的工资。 X 和 Y 都出现在 f_Z 中，因此， X 和 Y 都是 Z 的直接原因。如果 X 和 Y 有祖先，这些祖先将是 Z 的潜在原因。

图1.9展示了与SCM 1.5.1相关的图模型。

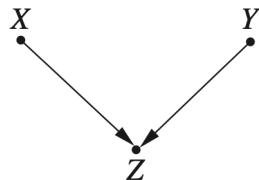


Figure 1.9 The graphical model of SCM 1.5.1, with X indicating years of schooling, Y indicating years of employment, and Z indicating salary

因为有从 X 和 Y 连接到 Z 的边，因而只需根据图模型就可以得知，模型中存在基于 X 和 Y 给 Z 赋值的函数，因此 X 和 Y 是 Z 的原因。然而，由于没有关于SCM更详细的描述，故不能从图中得到定义2的具体内容，或者换句话说，无法得知 X 和 Y 是如何引起 Z 的变化。

如果图模型比结构因果模型含有更少的信息，为什么还要使用它们呢？有以下几个原因。关于因果关系的知识通常不像SCM要求那样是定量的，而只是像图模型表示的那样是定性的。例如，容易知道性别是身高的原因，身高是打篮球成绩的原因，但要给它们之间的定量关系却是困难的。下面简单地创建一个部分细化的SCM来代替图模型。

SCM 1.5.2(身高、性别和篮球成绩)

$$V = \{\text{身高, 性别, 成绩}\}, \quad U = \{U_1, U_2, U_3\}, \quad F = \{f_1, f_2\}$$

$$\text{性别} = U_1$$

$$\text{身高} = f_1(\text{性别}, U_2)$$

$$\text{成绩} = f_2(\text{身高}, \text{性别}, U_3)$$

这里， $U = \{U_1, U_2, U_3\}$ 表示未知的外部因素，我们并不关心它们的名字，但它们影响 V 中的可测变量。 U 有时称为“误差项”或“省略因素”，表示观察变量的未知的和（或）随机的外生原因。

相比较部分细化的SCM，图模型提供了对因果关系更直观的理解。考虑上面介绍的SCM和与之相关的图模型，尽管SCM和它的图模型包含相同的信息，即 X 引发 Y 和 Y 引发 Z ，但通过查看图模型，可以更快、更容易地确定这些信息。

思考题

1.5.1

假设有如下的SCM，所有的外生变量独立并且期望值是0。

SCM 1.5.3

$$\begin{aligned} V &= \{X, Y, Z\}, \quad U = U_X, U_Y, U_Z, \quad F = f_X, f_Y, f_Z \\ f_X : X &= U_X \\ f_Y : Y &= \frac{X}{3} + U_Y \\ f_Z : Z &= \frac{Y}{16} + U_Z \end{aligned}$$

- (a) 画出这个模型的图。
- (b) 假定观察到 $Y = 3$ ，确定 Z 的最优估计(期望值)。
- (c) 假定观察到 $X = 3$ ，确定 Z 的最优估计。
- (d) 假定观察到 $X = 1$ 和 $Y = 3$ ，确定 Z 的最优估计。
- (e) 假设所有的外生变量服从均值为0、方差为1(即 $\sigma = 1$)的正态分布。
 - (i) 假定观察到 $Y = 2$ ，确定 X 的最优估计。
 - (ii) 假定观察到 $X = 1$ 和 $Z = 3$ ，确定 Y 的最优估计。

[提示：可以用多元回归技术，并可以利用如下事实：对三个正态分布的变量 X 、 Y 和 Z ，有 $E(Y | X = x, Z = z) = R_{YX \cdot Z}x + R_{YZ \cdot X}z$ 。]

1.5.2 乘法分解

图模型的另一个优点是，它能非常有效地表达联合分布。到目前为止，已经给出联合分布的两种表示方式。第一种是使用表格，对其中每个可能的组合值给出一个概率，这从直观上很容易理解，但在具有多个变量的模型中，它会占用大量的空间，10个二进制变量需要一个1024行的表！

第二种是结构因果模型，利用它可以更有效地表示 n 个变量的联合分布：只需要确定决定变量之间关系的 n 个函数，然后通过误差项的概率分析，就能发现支配联合分布的所有概率要素。但是，我们并不总是能够完全详尽地描述一个模型。我们可能知道一个变量是另一个变量的原因，而不知道与之相关的表达形式，或者可能不知道误差项的分布。即使获知了以上所有信息，但要将它们清晰地表达出来可能是很困难的，特别是当变量是离散的，且函数没有常见的代数表达式时。

幸运的是，可以利用以下法则及图模型克服以上障碍。

乘积分解法则

对任何无环的图模型，模型中变量的联合分布可通过对图中所有“家庭成员”计算条件分布概率 $P(\text{子节点} \mid \text{父节点})$ 的积给出。这个法则可形式化表述为：

$$P(x_1, x_2, \dots, x_n) = \prod_i P(x_i \mid pa_i) \quad (1.30)$$

其中， pa_i 表示变量 x_i 的所有父节点，积对 \prod_i 从 1 到 n 做乘积。式 (1.30) 应用了变量之间某些普遍成立的独立性，这些将在下一章中进行更详细的讨论。

例如，在一个简单的链 $X \rightarrow Y \rightarrow Z$ 中，可以直接写出：

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

这一法则使得我们在表达联合分布时，可以节省大量的空间。因为不再需要创建一个概率表，列出每个三元组 (x, y, z) 的值，只需为 $X, (Y \mid X)$ 和 $(Z \mid Y)$ 创建三个表就足够了，需时把值相乘即可。

为了估计由上述模型生成的数据集的联合分布，不必统计每个三元组的频次，而是统计每一个 $x, (y \mid x), (z \mid y)$ 及其乘积的频次，这为海量模型的处理节省了大量时间，实质上也提高了频次统计的准确性。因此，图的深层意蕴是将一个“高维”的分布估计问题变为一些“低维”的分布估计问题，所以图简化并提供了更精确的估计。如果我们不知道SCM的图形结构，那么对于具有众多变量的数据集合，哪怕只是中小规模，这种估计都将变得不可能，这就是所谓的“维数灾难”。

图模型让我们做到这一切，而不必总是需要知道那些与变量、参数或者误差项分布有关的函数。

如果要求不那么严格，这里给出一个实例来展示这种可节省时间和空间的策略。考虑链 $X \rightarrow Y \rightarrow Z \rightarrow W$ ，其中 X 代表有云/无云， Y 表示有雨/无雨， Z 代表潮湿路面/干燥路面， W 代表打滑路面/不打滑路面。

根据你的经验判断， $P(\text{有云, 无雨, 干燥路面, 打滑路面}) = 0.23$ 的可能性有多大？

这是一个很难直接回答的问题。但使用乘积法则，可以把它分解为

$$P(\text{有云})P(\text{无雨} \mid \text{有云})P(\text{干燥路面} \mid \text{无雨})P(\text{打滑路面} \mid \text{干燥路面})$$

根据常识， $P(\text{有云})$ 应该是相对比较高的，也许是0.5（当然，对于居住在季节不明显的洛杉矶居民，可能较低）。同样， $P(\text{无雨} \mid \text{有云})$ 是相当高的，比如0.75。 $P(\text{干燥路面} \mid \text{无雨})$ 仍然会很高，也许是0.9。但 $P(\text{打滑路面} \mid \text{干燥路面})$ 应该很低，应小于0.05。于是根据所有这些数据，可大概估算为 $0.5 \times 0.75 \times 0.9 \times 0.05 = 0.0169$ 。

在本书中，当需要用数值概率推理但又希望免给出大的概率表时，常常会使用这一乘积分解法则。

在处理估计问题时，乘积分解法则的重要性尤其值得重视。事实上，许多统计方法集中在有效的抽样设计和估算策略上，在研究适当的数据集时就可获得所需精度的概率估计。再看一下对于链 $X \rightarrow Y \rightarrow Z \rightarrow W$ ，估计概率 $P(X, Y, Z, W)$ 的问题。然而这次，我们尝试从数据中估计概率，而不是依据主观的判断。需要进行概率赋值的 (x, y, z, w) 的组合数是 $16 - 1 = 15$ 。假设随机观察45次，每次生成一个向量 (x, y, z, w) ，则平均每个 (x, y, z, w) 单元对应大约3个样本，有的对应一个或两个，有的干脆没有。即使每个单元都获得了足够数量的样本，也很难估计其在大规模群体（即样本量趋于无穷）中的比例。

然而，如果使用乘积分解法则，这45个样本会被分成更大的类别。为了确定 $P(x)$ ，每个 (x, y, z, w) 样本必落入两个单元 ($X = 1$) 和 ($X = 0$) 之一。显然，使其中之一为空集的概率很低，估计这两个单元在群体中频次的准确性要比估计前述15个单元的准确性高很多。要确定 $P(y | x)$ ，需要考虑4个单元：

$(Y = 1, X = 1), (Y = 0, X = 1), (Y = 1, X = 0)$ 和 $(Y = 0, X = 0)$ ；同理需要确定 $P(z | y)$ ，也涉及4个单元： $(Y = 1, Z = 1), (Y = 0, Z = 1), (Y = 1, Z = 0)$ 和 $(Y = 0, Z = 0)$ ；要确定 $P(w | z)$ ，涉及4个单元： $(W = 1, Z = 1), (W = 0, Z = 1), (W = 1, Z = 0)$ 和 $(W = 0, Z = 0)$ ，其中的分解同样成立。相比最初分为15个单元，每一个这样的分解将有助于实现更为精确的频次估计。通过SCM的图结构，可清楚地看到更简单的估计计算，并提高了频次估计的准确性。

上述实例不是为图模型提供定性知识的唯一用途。正如在下一节中将要看到的，相比直观理解，图模型能够揭示更多的信息。即使仅使用数据集中因果关系所构成的图模型，也能从该数据中了解并推理出很多信息。

思考题

1.5.2

假设一群患者中有 $r\%$ 的人患有某种致命的疾病，即表现为具有症状 Z ，并且他们可以服用延长寿命的药物 X ，效果为 Y （见图 1.10）。令 $Z = z_1$ 和 $Z = z_0$ 分别表示有和没有这种症状， $Y = y_1$ 和 $Y = y_0$ 分别表示死亡和存活， $X = x_1$ 和 $X = x_0$ 分别表示服用和未服用该药物。假设对于没有这种症状的患者 $Z = z_0$ ，他们服用该药物死亡的概率为 p_2 ，不服用该药物死亡的概率为 p_1 。另一方面，对于有症状的患者 $Z = z_1$ ，不服用该药物死亡的概率为 p_3 ，服用该药物死亡的概率为 p_4 。而且，有症状的患者更倾向于不服用药物，概率 $q_1 = P(x_1 | z_0)$ 和 $q_2 = P(x_1 | z_1)$ 。

- (a) 基于这个模型，根据参数 $(r, p_1, p_2, p_3, p_4, q_1, q_2)$ ，对于 x, y 和 z 的所有取值，计算联合分布 $P(x, y, z)$ 、 $P(x, y)$ 、 $P(x, z)$ 和 $P(y, z)$ 。[提示：参照 1.5.2 节的乘积分解法则。]
- (b) 对于有症状的患者、没有症状的患者及所有患者这三种人群，计算差值 $P(y_1 | x_1) - P(y_1 | x_0)$ 。
- (c) 根据(b)的结果，找出呈现辛普森悖论的参数组合。

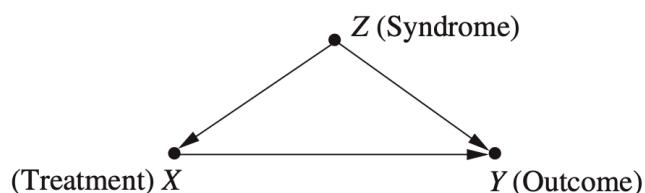


Figure 1.10 Model showing an unobserved syndrome, Z , affecting both treatment (X) and outcome (Y)

1.5.3

考虑二值随机变量的链 $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$ ，假设任意两个相邻变量之间的条件概率为

$$\begin{aligned}P(X_i = 1 | X_{i-1} = 1) &= p \\P(X_i = 1 | X_{i-1} = 0) &= q \\P(X_1 = 1) &= p_0\end{aligned}$$

计算下列概率：

$$\begin{aligned}P(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0) \\P(X_4 = 1 | X_1 = 1) \\P(X_1 = 1 | X_4 = 1) \\P(X_3 = 1 | X_1 = 0, X_4 = 1)\end{aligned}$$

1.5.4

给出对应于豪蒂霍尔问题的结构模型，并用它描述所有变量的联合分布。

2. 图模型及其应用

2.1 模型与数据的联系

在第1章中，分别介绍了概率、图和结构方程，但没有提及它们之间的关系。实际上，这三者是密切相关的。在概率语言中，独立的概念是用代数等式定义的，本章将展示如何用有向无环图（DAG）形象地表示概念。此外，图形化的表示还有利于刻画结构方程模型中隐含的概率信息。

熟悉结构方程模型的研究人员可以仅仅根据图模型的结构就能预测数据中的独立性质，而不需要依赖方程式或误差分布所携带的任何定量信息。反过来，这意味着观察数据中的独立性质对判断一个假设模型是否正确是有帮助的。在第3章还将介绍与数据结合的图结构，利用它能够定量预测干预的结果，而不必实际进行干预措施。

2.2 链结构和分叉结构

到目前为止，我们一直把因果模型看作是对数据背后的“因果故事”的表示。另一种看法是因果模型代表了数据产生的机制，可以把它看作世界相关部分的某种设计蓝图，我们可以用它来模拟世界中的数据。例如，如果给出高中三年级学生数学考试成绩的一个真实、完整的因果模型，并且给出一张完整的包含该模型各个外生变量值的表，那么理论上可以得出每个学生的考试成绩。当然，这需要指定所有可能影响学生考试成绩的因素，但这是不现实的。在大多数情况下，我们无法知道一个模型如此精确的知识。作为一种弥补，可以用一个概率分布来刻画外生变量，这样就可以生成整个学生群体和相关学生子群体的考试分数的近似分布了。

假设只有模型的一个图结构，但因果模型中变量的概率分布不明确，只知道哪些变量是由哪些其他变量引起的，但不知道关联的强度或性质，在此情况下，即使使用如此有限的信息，也可以获得由此模型生成的数据集的许多信息。从一个不明确的图形化因果模型中，即只知道该模型中哪些变量是哪些其他变量的函数，但不知道函数的具体形式，但可以获知数据集中哪些变量是相互独立的，以及哪些变量在哪些其他变量的条件下是相互独立的。这些独立性适用于由该图形化因果模型生成的各个数据集，而不论结构因果模型（SCM）附带的具体函数如何。

举个例子，以下三个假设的结构因果模型共享一个相同的图形化模型，第一个结构因果模型表示美国一所高中的学费（ X ，单位为美元）、SAT平均分（ Y ）和某年的大学录取率（ Z ）之间的因果关系。第二个结构因果模型表示一个灯泡开关状态（ X ）、一个与之相关的电路的状态（ Y ）和一个灯泡的状态（ Z ）之间的因果关系。第三个结构因果模型与竞赛参与者相关，它代表了参与者每周工作时间（ X ）、参与者每周参加训练的时间（ Y ）以及参与者完成比赛的时间（ Z ）之间的因果关系。在这三个模型中，外生变量（ U_X 、 U_Y 、 U_Z 等）代表了可能改变内生变量关系的未知或随机的影响因素。具体来说，在SCM2.2.1和SCM2.2.3中， U_Y 和 U_Z 是导致个体差异的附加因素。在SCM2.2.2中，如果有一些未被发现的异常，则 U_Y 和 U_Z 取值为1；反之，则 U_Y 和 U_Z 取值为0。

SCM 2.2.1（高中学费、SAT平均分和大学录取率）

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$

$$f_X : X = U_X$$

$$f_Y : Y = \frac{x}{3} + U_Y$$

$$f_Z : Z = \frac{Y}{16} + U_Z$$

SCM 2.2.2（开灯、电路和灯泡状态）

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$

$$f_X : X = U_X$$

$$f_Y : Y = \begin{cases} \text{关} & \text{当}(X = \text{"打开"}) \text{且} U_Y = 0 \text{ 或 } (X = \text{"合上"}) \text{且} U_Y = 1 \\ \text{开} & \text{否则} \end{cases}$$

$$f_Z : Z = \begin{cases} \text{亮} & \text{当}(Y = \text{"关"}) \text{且} U_Z = 0 \text{ 或 } (Y = \text{"开"}) \text{且} U_Z = 1 \\ \text{灭} & \text{否则} \end{cases}$$

SCM 2.2.3（工作时间、训练时间和比赛完成时间）

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$

$$f_X : X = U_X$$

$$f_Y : Y = 84 - x + U_Y$$

$$f_Z : Z = \frac{100}{y} + U_Z$$

SCM2.2.1-SCM2.2.3共享如图2.1所示的图模型。

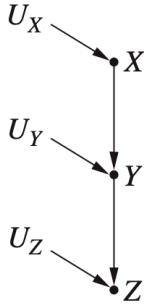


Figure 2.1 The graphical model of SCMs 2.2.1–2.2.3

SCM 2.2.1 和 SCM 2.2.3 处理的是连续变量，SCM 2.2.2 处理的是离散变量。在 SCM 2.2.1 中，自变量（父变量）的值越大，因变量（子变量）的值越大；SCM 2.2.3 中，自变量（父变量）的值越大，因变量（子变量）的值越小；SCM 2.2.2 中，变量之间的相关关系是逻辑的。虽然没有任何两个结构因果模型包含共同的函数，但是由于它们对应同一个图结构，因此所有由这三个结构因果模型生成的数据必然具有某些相同的独立性质，可以通过检测图 2.1 中的图模型来简单地预测这些独立性质。这三种结构因果模型生成的数据具有共同的独立性和可能的依赖关系如下。

1. Z 和 Y 可能是相互依赖的：对于某些 z, y ，有：

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

2. Y 和 X 可能是相互依赖的：对于某些 y, x ，有：

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

3. Z 和 X 可能是相互依赖的：对于某些 z, x ，有：

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

4. Z 和 X 在 Y 的条件下是独立的：对于所有的 x, y, z ，有：

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

为了理解这些独立性和依赖关系，我们来检测图模型。首先，验证由一条边连接的两个变量可能是相互依赖的。注意，从一个变量指向另一个变量的箭头表示第一个变量是第二个变量的原因，也就是说，第一个变量是确定第二个变量值的函数的一部分。因此，第二个变量的值依赖于第一个变量的值；在某些情况下，改变第一个变量的值会引起第二个变量值的改变。这意味着，当检测数据集

中的这些变量时，如果已知一个变量的值，则另一个变量取某个值的概率可能会改变。所以在一个典型的因果模型中，不管具体的函数是什么，由边连接的两个变量是依赖的。有这个推论可知，在SCM2.2.1~SCM2.2.3中， Z 和 Y 可能是相互依赖的， Y 和 X 也可能是相互依赖的。

根据这两个事实，可以得出结论：如果 Z 依赖于 Y 的值，而 Y 依赖于 X 的值，那么 Z 很可能依赖于 X 的值，即 Z 和 X 可能是依赖的。然后，在某些特殊的情况下不是这样的。例如，考虑下面的结构因果模型，他的图模型也与图2.1相同。

SCM 2.2.4 (非传递依赖的情况)

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$

$$f_X : X = U$$

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

$$f_Z : Z = \begin{cases} i, & \text{当 } Y = c \text{ 或 } U_Z = 1 \\ j, & \text{当 } U_Z = 2 \end{cases}$$

在这个例子中，无论 U_Y 和 U_Z 取什么值， X 都对 Z 值没有任何影响。 X 的变化会造成 Y 在 a 和 b 之间发生变化，但除非 Y 取值 c ，否则 Y 不会影响 Z 。因此， X 和 Z 在这个模型中独立地变化。此类情况称为非传递情况。

然而，非传递情况并不常见，在大多情况下， X 和 Z 的值会一起变化，就像 X 和 Y 、 Y 和 Z 一样，因此，在数据集中它们可能是相互依赖的。

现在再来讨论：在条件 Y 下， Z 和 X 是独立的。回顾一下，以 Y 为条件时，我们基于 Y 的值将数据过滤划分成不同的组，然后分别比较 $Y = a, Y = b$ 时的情况。首先来看 $Y = a$ 的情况。我们想知道在这些情况下， Z 的值是否独立于 X 的值。之前，我们认为 X 和 Z 可能是相互依赖的，因为当 X 的值发生变化时， Y 的值可能会发生改变，而当 Y 的值发生变化时， Z 的值可能会改变。但是现在仅检测 $Y = a$ 值的情况，即选择具有不同的 X 值时， U_Y 的值需随之变化而使得 Y 的值保持为 a 。但因为 Z 值仅取决于 Y 和 U_Z ，不依赖于 U_Y ，所以（无论 U_Y 怎么变） Z 的值都不会改变。这样一来，选择一个不同的 X 值并不会改变 Z 的值，因此，在 $Y = a$ 的情况下， X 与 Z 是独立的。无论 Y 的值怎样变化，这个结论显然都是正确的。所以在 Y 的条件下， X 与 Z 是独立的。

这种由三个节点和两条边组成，并且中间变量有一条边进入和一条边射出的结构称为链结构。以上的推理说明，在任何图模型中，对于给定的任意两个变量 X 和 Y ，如果 X 和 Y 之间的唯一路径完全由链组成，那么以该路径上的任何中间变量为条件， X 和 Y 都是独立的。无论连接变量的函数是什么，这种独立关系都成立。由此得到第一条规则：

规则 2.2.1 (链结构中的条件独立性) 如果变量 X 和 Y 之间只有一条单向路径， Z 是截断这条路径的任何一组变量，则在 Z 的条件下， X 和 Y 是独立的。

注：链式结构当中两头(即 X 和 Z)不一定关联，但是在 Y 的条件下，两头一定独立。

需要注意的是，只有当假设误差项 U_X, U_Y 和 U_Z 相互独立时，规则2.2.1才成立。例如，如果 U_X 是 U_Y 的原因，那么以 Z 为条件不一定会使 X 和 Y 相互独立，因为 X 的变化可能仍然通过它们的误差项与 Y 的变化关联。

现在来看如图2.2所示的图模型。举例来说，这个结构可能表示一个城市某天的温度（ X ），当天本地一家冰激凌店的销售量（ Y ），以及当天城市暴力犯罪的数量（ Z ）之间的因果机制。SCM 2.2.5给出了这些变量之间可能的函数关系。这种结构也可以代表SCM 2.2.6中关于开关（ X ）的状态（开或关），第一个灯泡（ Y ）的状态（亮或灭）以及第二个灯泡（ Z ）的状态（灭活亮）之间的因果关系，外生变量 U_X, U_Y 和 U_Z 代表影响这些变量相互之间关系的其他随机因素。

SCM 2.2.5 (温度、冰激凌销量和罪犯)

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$

$$f_X : X = U_X$$

$$f_Y : Y = 4x + U_Y$$

$$f_Z : Z = \frac{x}{10} + U_Z$$

SCM 2.2.6 (开关和两个灯泡的状态)

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$

$$f_X : X = U_X$$

$$f_Y : Y = \begin{cases} \text{亮, 当 } (X = \text{"开"}) \text{ 且 } U_X = 0 \text{ 或 } (X = \text{"关"}) \text{ 且 } U_Y = 1 \\ \text{灭, 否则} \end{cases}$$

$$f_Z : Z = \begin{cases} \text{亮, 当 } (X = \text{"开"}) \text{ 且 } U_Z = 0 \text{ 或 } (X = \text{"关"}) \text{ 且 } U_Z = 1 \\ \text{灭, 否则} \end{cases}$$

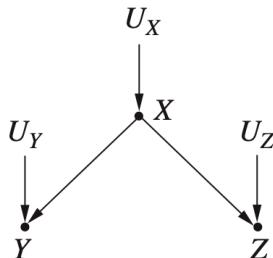


Figure 2.2 The graphical model of SCMs 2.2.5 and 2.2.6

如果假设误差项 U_X, U_Y 和 U_Z 是相互独立的，那么通过检测图2.2中的图模型，可以确定SCM 2.2.5和SCM 2.2.6共享如下的依赖性和独立性。

1. X 和 Y 可能是相互依赖的：对于某些 x, y ，有

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

2. X 和 Z 可能是相互依赖的：对于某些 y, z ，有

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

3. Z 和 Y 是独立的：对于所有的 z, y ，有

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

4. Y 和 Z 在 X 的条件下可能是独立的：对于某些 x, y, z ，有

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

对于第1点和第2点，由于 Y 和 Z 也都通过肩头与 X 直接相连，因此当 X 的值改变时， Y 的值和 Z 的值都可能发生变化。这使我们进一步想到：当 X 改变时， Y 会发生变化， Z 也会发生变化，那么当 X 改变时， Y 与 Z 可能（虽然不确定）会一起发生变化，反之亦然。由于可以从 Y 值的变化中得到 Z 值发生相应变化的信息，所以 Y 和 Z 可能是相互依赖的变量。

那么，为什么在 X 的条件下 Y 和 Z 独立呢？当以 X 为条件时发生了什么？以 X 的值来筛选数据，所以只比较 X 是一个固定值的情况。由于 X 的值是不变的，所以 Y 和 Z 的值不会随着 X 的变化而变化，它们只会随着 U_Y 和 U_Z 而变化，由于已经假设 U_Y 和 U_Z 是独立的，因此， Y 和 Z 的值是独立变化的。

这种具有三个节点，并且有两个箭头从中间变量射向的结构称为分叉结构。分叉结构中的中间变量是其他两个变量和它们任何后代的共同原因。如果两个变量共享一个共同原因，并且这个共同原因是它们之间唯一路径的一部分，那么上述推理说明，这些变量的依赖关系和条件独立性是成立的，因此，得到另一个规律：

规则2.2.2（分叉结构的条件独立性） 如果变量 X 是变量 Y 和 Z 的共同原因，并且 Y 和 Z 之间只有一条路径，则 Y 和 Z 在 X 的条件下独立。

2.3 对撞结构

到目前为止，已经研究了两种简单的边和节点的结构：链结构和分叉结构，这两种结构均允许边和节点出现在两个变量之间的路径上。还有第三种结构，因为这种结构有其独特的考虑和挑战，因此在这里专门介绍。这种结构包含一个对撞节点，它指的是一个节点接收来自两个节点射出的边。图 2.3 展示了一个最简单的，包含对撞结构的因果模型图，它代表两个原因 X 和 Y 的共同效应 Z 。

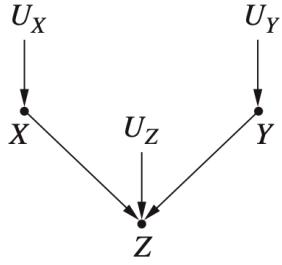


Figure 2.3 A simple collider

与每个因果模型图一样，所有具有像图2.3这样结构的因果模型都可以从图中得出相同的依赖性和独立性。对于如图2.3所示的模型，假设 U_X 、 U_Y 和 U_Z 是独立的，可以得出如下的独立性。

1. X 和 Z 可能是相互依赖的：对于某些 x, z ，有

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

2. X 和 Z 可能是相互依赖的：对于某些 y, z ，有

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

3. X 和 Y 是独立的：对于所有的 x, y ，有

$$P(X = x | Y = y) = P(X = x)$$

4. X 和 Y 在 Z 的条件下可能是相互依赖的：对于某些 x, y, z ，有

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

前两点已在2.2节中解释了。第3点是不言而喻的，因为 X 和 Y 都不是彼此的后代或祖先，也没有依赖于同一个变量的值，它们只会分别受 U_X 和 U_Y 的影响，而 U_X 和 U_Y 是独立的，所以没有使 X 值变化与 Y 值变化相关联的因果机制。这种独立性也反映了因果关系在时间上的机制，当前独立的事件不会因为将来产生共同的效应而变得相互依赖。

那么什么第4点成立呢？为什么以共同效应作为条件时，两个独立的变量实际上变得相互依赖了呢？为了回答这个问题，再改回到以条件变量的值来过滤的定义上。当以 Z 为条件时，我们将比较限定在 Z 取相同值的情况下。由于 Z 的值依赖于 X 和 Y ，因此，在对 Z 取相同值做比较时， X 值的任何变化必须通过 Y 值的变化来补偿，否则 Z 值也会改变。

对撞结构的这种特性，以对撞节点为条件会使该节点的父节点互相依赖，背后的原理一开始很令人费解，例如，一个最基本的例子： $Z = X + Y$ ，且 X 和 Y 是独立的变量，有以下推理：如果已知 $X = 3$ ，你不知道关于 Y 可能值的任何信息，因为这两个数字是独立的；另一方面，如果已知 $Z = 10$ ，那么再告诉 $X = 3$ ，相当于立刻确认 Y 必然是7。因此，当给定 $Z = 10$ 时， X 和 Y 是依赖的。

可以通过现实生活的例子进一步阐明这种现象。例如，假设某所大学为两类学生提供奖学金：一类是具有超常音乐天赋的学生，另一类则是拥有超常学业成绩的学生。通常，音乐天赋和学业成绩是独立的特质，所以在广泛人群中，对于一个有音乐天赋的人，无从获得关于其学业成绩的任何信息。但是，如果发现一个人获得了奖学金，并且知道这个缺乏音乐才能，那么会立即推断他可能有很高的学业成绩。因此，在固定两个独立变量的共同效应（第三变量为奖学金）的值时，这两个独立变量会变得相互依赖。

再来研究一个数值例子：同时（独立）投掷两枚质地均匀的硬币，至少有一枚硬币落地时正面向上时，铃就会响。令 X 和 Y 分别表示两枚硬币的投掷结果，令 Z 代表铃的状态， $Z = 1$ 表示铃响了， $Z = 0$ 表示铃没响。这种机制可以表示为如图2.3所示的对撞结构，图中两枚硬币的投掷结果是父节点，铃的状态是对撞节点。

如果已知硬币1落地时正面向上，由于两枚硬币是独立的，我们不能获得硬币2的结果。但是假设听到铃声，且知道硬币1落地时反面向上，则可确认硬币2落地时必然正面向上。同样，如果假设听到了铃声，且知道硬币2也是落地时正面向上，那么硬币1落地时正面向上的概率会改变，这种概率变化比前一种情况更微妙。

为了后面的计算，表2.1给出了两枚质地均匀的硬币同时投掷的结果的概率分布。其中， X 表示第一枚硬币， Y 表示第二枚硬币， Z 表示铃，如果任何一枚硬币落地时正面向上，则铃响。

表2.1 两枚质地均匀的硬币同时投掷的结果的概率分布

X	Y	Z	$P(X, Y, Z)$
正面	正面	1	0.25
正面	反面	1	0.25
反面	正面	1	0.25
反面	反面	0	0.25

由表2.1知

$$P(X = \text{正面} \mid Y = \text{正面}) = P(X = \text{反面} \mid Y = \text{反面}) = \frac{1}{2}$$

也就是说， X 和 Y 是相互独立的。现在，以 $Z = 1$ （铃声响）和 $Z = 0$ （铃声不响）为条件，所得数据子集如表2.2所示。

表2.2 基于表2.1分布的条件概率分布

X	Y	$P(X, Y Z = 1)$	$P(X, Y Z = 0)$
正面	正面	0.333	0
正面	反面	0.333	0
反面	正面	0.333	0
反面	反面	0	1

通过计算这些表中的概率，有

$$P(X = \text{正面} | Z = 1) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$$

如果进一步筛选出 $Z = 1$ 的子表，然后研究 $Y = \text{正面}$ 的情况，将得到

$$P(X = \text{正面} | Y = \text{正面}, Z = 1) = \frac{1}{2}$$

可以看出，在 $Z = 1$ 的情况下，在知道 $Y = \text{正面}$ 时， $X = \text{正面}$ 的概率从 $\frac{2}{3}$ 变到 $\frac{1}{2}$ 。显然，给定 $Z = 1$ 后， X 和 Y 是相互依赖的。当然，当铃声不响 ($Z = 0$) 时，更明显的相关依赖发生了，因为我们知道这两枚硬币落地时肯定都是反面向上的。

另一个对撞结构的例子是1.3节中遇到的蒙蒂霍尔问题。这个例子有助于进一步理解这种结构引起的问题。从本质上来说，蒙蒂霍尔问题反映了对撞结构的存在。你最初选择的门是一个父节点；后而有车的门是另一个父节点；蒙蒂打开的、后而有山羊的门是对撞节点，该节点受其他两个变量的共同影响。这里的因果关系是：如果你选择门 A ，如果门 A 后面有一只羊，那么蒙蒂将被迫打开剩下的、后面有羊的门。

最初关于门的选择和汽车的位置是独立的，这就是为什么虽初有 $\frac{1}{3}$ 的机会选择有车的门。然而，正如投掷两枚独立的硬币一样，在蒙蒂对门进行选择的条件下，最初的选择和车的位置是相互依赖的。在这个例子中，最初车在门 B 后面的概率为 $\frac{1}{3}$ ，但是在选择门 A 并且蒙蒂打开门 C 的情况下，车在门 B 后面的概率将为 $\frac{2}{3}$ 。

在对撞节点的条件下，可以使先前独立的变量变得相互依赖，以对撞点的任何后代为条件也是如此，再次回到投掷两枚独立的硬币的例子看看为什么是这样。假设没有直接听到铃声，而是依靠一个不太可靠的目击者。当铃不响时，目击者有 50% 的概率谎报铃响。令 W 表示目击者的报告，因果结构如图2.4所示，其中， X 代表第一枚硬币投掷的结果， Y 代表第二枚硬币投掷的结果， Z 表示当 X 或 Y 是正面向上就会响的铃， W 代表报告铃是否响的目击者， X, Y 和 W 的所有组合的概率如表2.3所示。

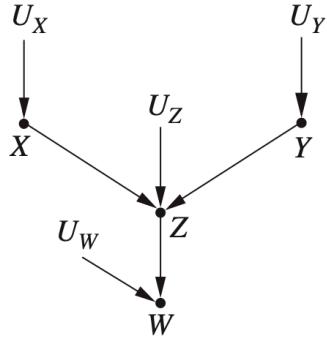


Figure 2.4 A simple collider, Z , with one child, W , representing the scenario from Table 2.3, with X representing one coin flip, Y representing the second coin flip, Z representing a bell that rings if either X or Y is heads, and W representing an unreliable witness who reports on whether or not the bell has rung

表2.3 两枚质地均匀的硬币同时投掷的结果的概率分布。

X	Y	W	$P(X, Y, W)$
正面	正面	1	0.25
正面	反面	1	0.25
反面	正面	1	0.25
反面	反面	1	0.125
反面	反面	0	0.125

基于表2.3，可以很容易地验证如下式子：

$$P(X = \text{正面} \mid Y = \text{正面}) = P(X = \text{正面}) = \frac{1}{2}$$

$$P(X = \text{正面} \mid W = 1) = \frac{0.25 + 0.25}{0.25 + 0.25 + 0.25 + 0.125} = \frac{0.5}{0.875}$$

$$P(X = \text{正面} \mid Y = \text{正面}, W = 1) = \frac{0.25}{0.25 + 0.25} = 0.5 < \frac{0.5}{0.875}$$

由上述结果可得出，在得到目击者报告之前， X 和 Y 是独立的，但之后就变得相互依赖了。进而得出2.2节两个规则之外的第三条规则：

规则 2.3.1 (对撞结构中的条件独立性) 如果变量 Z 是变量 X 和 Y 之间的对撞节点，并且 X 与 Y 之间只有一条路径，那么 X 与 Y 是无条件独立的，但是在 Z 或 Z 的任何子孙条件下是相互依赖的。

规则2.3.1对于研究因果关系是极其重要的，在接下来的章节中，该规则可用于测试一个数据集是否由某个因果模型所生成，用于从数据中发现模型，或者用于在混杂情况下确定应该检测哪个变量以及如何估计因果效应，从而彻底解决因果推断问题。

备注：好奇的读者可能想知道为什么在对撞节点条件下产生的依赖关系会令大多数人都感到惊讶，正如蒙蒂霍尔问题所示的那样。原因是人们倾向于将依赖关系和因果关系联系起来。因此，他们错误地认为两个变量之间的统计依赖之所以存在，只能是因为有产生这种依赖性的因果机制；也就是说，要么两个变量中的一个是另一个的原因，要么出现第三个变量是这两个变量的共同原因。在对撞结构中，他们惊奇地发现还有第三种方式可以产生这种依赖，违背了“没有因果关系就无关联”的假设。

思考题

2.3.1

$$X \rightarrow R \rightarrow S \rightarrow T \leftarrow U \leftarrow V \rightarrow Y$$

Figure 2.5 A directed graph for demonstrating conditional independence (error terms are not shown explicitly)

(a) 列出图2.5中在集合 $Z = \{R, V\}$ 条件下的每一对独立的变量。

(b) 列出图2.5中使每一对不相邻的变量独立的条件变量。

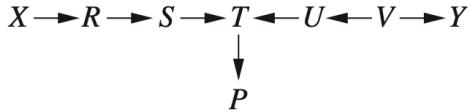


Figure 2.6 A directed graph in which P is a descendant of a collider

(c) 列出图2.6中在集合 $Z = \{R, P\}$ 条件下的每一对独立的变量。

(d) 列出图2.6中使每一对不相邻的变量独立的条件变量。

(e) 假设用图2.5中的模型产生一组数据，并用线性方程 $Y = a + bX + cZ$ 来拟合这组数据，为确保斜率 $b = 0$ ， Z 可以是模型中的那些变量？[提示：回忆一下，一个非零斜率意味着 X 与 Y 在 Z 条件下相互依赖。]

(f) 继续问题(e)，但现在参考图2.6，如果方程 $Y = a + bX + cR + dS + eT + fP$ 来拟合数据，哪些系数可能为0？

2.4 d-分离

因果模型通常不会像前面遇到的例子那样简单。具体来说，变量之间只有一条路径的图模型是很少的。在大多数的图模型中，变量可能有多条路径连接，且每个路径包含多个链、分叉和对撞结构。因此需要考虑这样的问题，即对于任意复杂的图因果模型，是否存在一个准则或方法，用来预测由该模型生成的数据所具有的相关性质呢？

实际上，根据前面章节介绍的规则，可以得到一个这样的方法：d-分离。**d-分离**（d表示“方向的”）使我们能够确定任何一对节点是否是d-连通的，即它们之间是存在一条连通路径；或者确定任何一对节点是否是d-分离的，即它们之间不存在连通的路径。当说一对节点是d-分离的，指的是这两个变量是绝对独立的；当说一对节点是d-连通的，指的是这两个变量可能或很有可能是相互依赖的。

如果两个节点 X 和 Y 之间存在的任何路径都被阻断，则它们是（关于这些阻断变量）d-分离的；如果 X 和 Y 之间存在一条路径没有被阻断，那么 X 和 Y 是连通的。可以把变量之间的路径看作管道、依赖性就像通过管道里的水：如果存在一根管道是未被阻断的，水就可以从一个地方流到另一个地方，如果有一条路径是通畅的，那么两端的变量就是相互依赖的。仅在一个地方阻断管道就可以阻止水流通过，同样地，只需要一个节点就可以阻断整个路径上的依赖性传递。

d-分离分为两类：以某些节点为条件和不以任何节点为条件。如果不以任何节点为条件，那么只有对撞节点可以阻断一条路径。原因是：不以任何变量为条件时，对撞结构会阻断依赖关系，正如在2.3节中所提到的那样。因此，如果两个节点 X 和 Y 之间的每条路径都有一个对撞节点，则 X 和 Y 不会有依赖关系，它们必须是边缘独立的。

然而，如果以一组节点 Z 为条件，那么以下类型的节点可以阻断一条路径：

- 自身不在 Z 中且其子孙节点也不在 Z 中的对撞节点
- 在 Z 中的链节点或分叉的中间节点

其背后的原因要追溯到在2.2节和2.3节介绍的内容。对撞节点不允许依赖性在其父节点之间传递，因此阻断了路径。而规则2.3.1告诉我们，当以对撞节点或其后代为条件时，父节点之间可能会变成互相依赖的。因此，不在条件集 Z 中的对撞节点会使依赖关系无法在路径上传递，而在条件集中的对撞节点或其后代则不会阻断依赖关系。相反，非对撞结构（包括链结构和分叉结构）不会阻断依赖关系，但规则2.2.1和规则2.2.2又告诉我们，当以非对撞结构的中间节点为条件时，这些路径的两端节点会变得独立（每次只考察一条路径）。因此，条件集中的任何非对撞节点将阻断依赖关系，而在条件集中的非对撞节点将允许依赖关系在路径上传递。

定义2.4.1 (d-分离) 一条路径会被一组节点 Z 阻断，当且仅当：

- 路径包含链结构 $A \rightarrow B \rightarrow C$ 或分叉结构 $A \leftarrow B \rightarrow C$ ，其中间节点 B 在 Z 中（即以 B 为条件）；或者
- 路径 p 包含一个对撞结构 $A \rightarrow B \leftarrow C$ ，且对撞节点 B 及其子孙节点都不在 Z 中。

如果 Z 阻断了 X 和 Y 间的每一条路径，则 X 和 Y 在 Z 的条件下是d-分离的，因此 X 和 Y 在以 Z 为条件时是独立的。

我们使用d-分离工具来探究一些更复杂的图模型，并确定其中的独立变量和依赖变量，包括不以其它变量为条件和以其它变量为条件这两种情况。以如图2.7所示的图模型为例，图2.7可能与许多因果模型相关联。变量可能是离散的、连续的或两者的混合，它们之间的关系可能是线性的、指数的或其它任意形式的。然而，无论模型如何，d-分离将总是描述该模型生成的数据集所具有的独立性质。

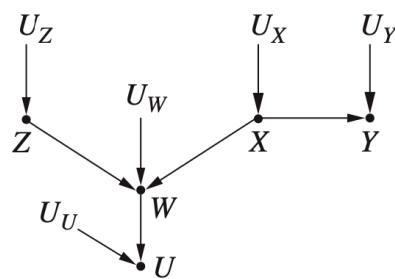


Figure 2.7 A graphical model containing a collider with child and a fork

特别地，我们来分析图2.7模型中 Z 和 Y 之间的关系。使用一个空的条件集时，它们是d-分离的，这说明 Z 和 Y 是无条件独立的。为什么呢？因为 Z 和 Y 之间只有一条路径，并且该路径被对撞结构 ($Z \rightarrow W \leftarrow X$) 阻断，所以它们之间就没有未被阻断的路径。

但如果以 W 为条件呢？d-分离告诉我们在 W 的条件下， Z 和 Y 是d-连通的。原因是：当条件集是 $\{W\}$ 时， Z 和 Y 之间的唯一路径包含一个分叉节点 (X)，该节点不在条件集内，并且路径中的唯一对撞节点 (W) 在条件集内，因此该路径没有被阻断（注意以对撞节点为条件时，会“解除阻断”）。如果以 U 为条件，由于 U 是 Z 和 Y 之间路径上对撞节点的后代，因此结果也是如此。

另一方面，如果以集合 $\{W, X\}$ 为条件，则 Z 和 Y 仍然是独立的。此时，基于规则2.2.1，路径上有一个在条件集合中的非对撞节点 (X)，因此 Z 和 Y 之间的路径被阻断。虽然通过以 W 为条件解除了阻断，但有一个阻断节点就足以阻断整条路径。由于 Z 和 Y 之间的唯一路径被该条件所阻断，所以 Z 和 Y 在 $\{W, X\}$ 的条件下是d-分离的。

现在考虑，当在 Z 和 Y 之间添加另一个路径时，如图2.8所示，会发生什么情况。 Z 和 Y 现在是无条件依赖的。为什么？因为它们之间有一条路径 ($Z \leftarrow T \rightarrow Y$)，且该路径不包含对撞节点。然而，如果以 T 为条件，则该路径被阻断， Z 和 Y 再变得独立。另一方面，以 $\{T, W\}$ 为条件时，它们再次变成d-连通（以 T 为条件时，阻断了路径 $Z \leftarrow T \rightarrow Y$ ，但以 W 为条件则解除了路径 $Z \rightarrow W \leftarrow X \rightarrow Y$ 的阻断）。如果将 X 添加到条件集中，条件集变成 $W, U, \{W, U\}, \{W, T\}, \{U, T\}, \{W, U, T\}, \{W, X\}, \{U, X\}$ ，或 $\{W, U, X\}$ 时， Z 和 Y 是d-连通的（因此也可能是相互依赖的）。而当条件集为 $T, \{X, T\}, \{W, X, T\}, \{U, X, T\}, \{W, U, X, T\}$ 时，它们是d-分离的（因此是独立的）。注意： T 在使 Z 和 Y 是d-分离的每个条件集合中；这是因为 T 是使 Z 和 Y 无条件d-连通的路径上的唯一节点，所以除非以它为条件，否则 Z 和 Y 将始终是d-连通的。

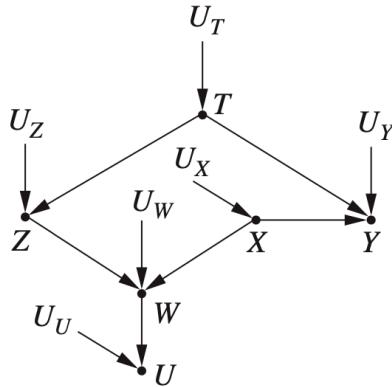


Figure 2.8 The model from Figure 2.7 with an additional forked path between Z and Y

思考题

2.4.1

图2.9表示一个误差项已被移除的因果图，其中，假设所有 U （未显示）都是独立的。假设所有这些误差项都是相互独立的。

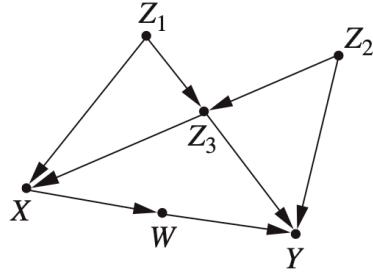


Figure 2.9 A causal graph used in study question 2.4.1, all U terms (not shown) are assumed independent

- (a) 对于图2.9中每一对不相邻的节点，找到一组使它们d-分离的节点。在数据独立性方面，这个结果说明了什么？
- (b) 假设只有集合 $\{Z_3, W, X, Z_1\}$ 中的节点可以测量，重新回答问题(a)。
- (c) 对于图2.9中每一对不相邻的节点，在以其他所有节点为条件的情况下，确定它们是否是独立的。
- (d) 对于图中每一个节点 V ，找到一个最小的节点集合，使得 V 与图中其他所有节点独立。
- (e) 假设想要用模型中其他所有变量的值来估计 Y 的值。我找到一个最小的变量集合，能够获得与所有变量同样准确的对于 Y 的估计。
- (f) 假设希望估计 Z_2 ，重新回答问题(e)。
- (g) 假设希望用 Z_3 的值来预测 Z_2 的值，如果增加 W 的值，精度会提升吗？解释原因。

2.5 模型检验与因果搜索

前面的章节表明因果模型在其生成的数据集上有可被检验的推论式。例如，如果认为图 G 可能生成了数据集 S ，d-分离将告诉我们图 G 中的哪些变量在哪些其他变量的条件下一定是独立的，而我们可以用数据集来检验条件独立性。假设列出图 G 中的d-分离条件，并注意到变量 A 和 B 在 C 的条件下一定是独立的，然后，基于数据集 S 来估计概率，发现 A 和 B 在 C 的条件下不是独立的，我们就拒绝将图 G 作为数据集 S 的因果模型。

现在以图2.9的因果模型来说明。由模型列出的条件独立性可知， W 和 Z_1 在给定 X 的条件下是独立的，因为 X “d-分离” W 与 Z_1 。现在在 X 和 Z_1 上回归 W ，也就是说，要寻找一条最符合我们的数据的直线：

$$w = r_X x + r_1 z_1$$

如果 $r_1 \neq 0$ ，表明在给定 X 的条件下， W 的值依赖于 Z_1 ，因此，模型是错误的（回想一下，条件相关就是条件依赖）。我们不仅知道模型是错误的，而且也知道错误在哪儿。在正确的模型中，在 W 和 Z_1 间一定有一条不被 X “d-分离”的路径，这是适用于所有具有误差独立性特质的（有向）无环模型的一个理论结果（Verma et al., 1990）。而且如果模型中的每个d-分离条件均与数据中的条件独立性一致，则任何进一步的检验也不能否定该模型。这意味着，对于任何（与模型独立性相符的）数据集，人们总能为这个模型找到一组函数，并指定误差项的摄率，从而确地生成该数据。

还可以利用其他方法来检验模型（对于数据集）的适合度。评估适合度的标准方法涉及对模型的统计假设检验，也就是说，评估所观察的样本不是凭运气，而是根据由假设模型产生的可能性有多大。然而，由于模型没有完全确定，在评估这种可能性之前需要先估计模型参数。可以通过假设线性高斯模型（即，模型中的所有函数都是线性的，所有误差项都是正态分布的）来近似估计参数，因为在这样的假设下，联合分布（也是高斯分布）可以用模型参数来简洁地表达，然后可以评估观察样本由完全参数化的模型生成的可能性（Bollen, 1989）。

然而，这个过程也存在很多问题。首先，如果有某些参数不能被估计，那么就得不到联合分布，则模型不能被检验。正如将在3.8.3节中看到的，当某些误差项相关或者某些变量不可观测时，会出现这种问题。其次，这是一个全局性模型检验过程，如果发现这个模型不能较好地适合数据，我们无法去寻找其原因，也没有办法确定应该在模型中删除或添加哪些边来提高这种适合性。第三，当对一个模型实施局部性检验时，涉及的变量可能很多，如果每个变量均存在测量噪声和/或采样差异，那么检验将不可靠。

相比于这种局部性检验方法，d-分离有以下优点。首先，它是非参数的，这意味着它不依赖于任何具体的变量间的函数，而是仅使用问题中的图模型。其次，它仅能依赖局部性检验模型，而不是全局性检验，这使我们能够识别假设模型中有缺陷的特定区域，并修复该区域，从而得到一个全新的模型。这也意味着，如果无法确定模型中某个区域的参数，无论是什么原因导致的，我们仍然可以获得模型剩余部分的一些不完整的信息（与全局性检验方法相反，在全局性检验方法中，如果无法估计一个参数，就不能检验模型的任何部分）。

如果有一台计算机，就可以用d-分离方法来检验和排除很多可能的模型，最终将得到一些经过检验，与数据结果中依赖关系不矛盾的模型。最终结果是几个模型，而不是一个模型，这是因为有些图有不可区分的蕴涵式。具有不可区分蕴涵式的一组图被称为等价类。如果两个图 G_1 和 G_2 有相同的骨架（即有相同的边，而不管边的方向），并且它们具有同样的v-结构（即父节点不相邻的对撞结构），则这两个图在同一个等价类内。满足这条规则的任何两个图都具有相同的d-分离条件集，因此具有相同的可检验蕴涵式集（Verma et al., 1990）。

这个结果的重要性在于它使我们能够为数据集寻找可能产生它的因果模型。因此，不仅可以从一个因果模型出发生成一个数据集，也可以从一个数据集出发反推出因果模型。由于大多数数据分析研究的目标正是要找到一个解释数据的模型，因此，这个结果是非常有用的。

还有其他的因果模型搜索方法，包括在本节开始时提到的依赖于全局校验的一些方法，但是对它们进行全面研究超出了本书的范围，对此感兴趣、想要了解更多的读者可以参考Pearl (2000)、Pearl祭 (1991)、Rebane等 (1987)、Spirtes等 (1991) 以及Spirtes等 (1993)。

思考题

2.5.1

- (a) 图2.9中的哪些箭头反向后可不被任何统计检验识别出来？[提示：使用等价类准则。]
- (b) 应用等价类准则列出所有等价于如图2.9所示的因果图。
- (c) 列出图2.9中可以由非实验数据确定方向的箭头。
- (d) 给出一个 Y 的回归方程，使得当方程中某个系数非零时，如图2.9所示的模型是错误的。
- (e) 给出一个 Z_3 的回归方程，使得当方程中某个系数非零时，如图2.9所示的模型是错误的。
- (f) 假设 X 不可测量，重新回答问题(e)。

(g) 对(d)和(c)中的这类回归方程一共需要多少个，才能确保模型能够被完全检验？也就是说，如果图通过了所有的这些检验，它将不能被其他这类检验所否定。[提示：确保你检验了每一个由式(1.30)乘积分解所隐含的偏回归系数为0的情况。]

3 干预的效果

3.1 干预

许多统计研究的最终目标是预测干预措施的效果。例如，我们收集西部火灾相关因素的数据，实际上是要寻找可以用于预测的因素，以减少火灾的发生；当对一种新的癌症药物进行研究时，通过让患者服药以实施干预，观察患者用药后的反应；而当研究暴力电视节目与儿童的攻击性行为之间的相关性时，是想尝试确认少数儿童接触暴力电视节目的干预措施能否降低儿童的攻击性。

在统计学课程中常会提到“相关关系不是因果关系”。两个变量之间的关系并不仅仅只有一个变量引起另一个变量的变化（关于这个性质有一个著名例子：冰激凌销量的增加与暴力犯罪数目的增加是有关系的，不是因为冰激凌导致犯罪，而是因为冰激凌销量和暴力犯罪都在炎热天气中更常见）。因此，随机对照试验被认为是统计学中的黄金准则。在一个正确的随机对照试验中，除了输入变量，所有影响输出变量的因素要么是不变的，要么是随机变化的，因此输出变量的任何改变必然由这一个输入变量引起。

不幸的是，很多问题不适合用随机对照试验来解决。我们不能控制天气，所以无法将引起火灾的变量随机化；研究暴力电视节目的时候，虽可以随机选取参与者，但很难有效地控制每个孩子电视的行为，而且几乎不可能知道我们对孩子的控制是否有效；甚至在随机药物试验中，也会出现很多问题，参与者退出了、没有吃药或者弄虚作假吃药。

在随机对照试验不可行的情况下，研究人员实施观察性研究，他们仅仅记录数据，而不是擦除数据。这种研究方法的问题在于很难将因果关系从相关关系中提取出来。常识告诉我们，对冰激凌的销量进行干预，不会影响犯罪的数目，但事实不都是这么清晰。例如，温尼伯大学最近的一项研究表明，青少年过度发短信与（知识）“肤浅”相关。有媒体证实说，发短信使青少年更加肤浅（从干预角度说，对青少年进行干预，使他们减少发短信的数量，从而不让他们那么“肤浅”）。但是，这个试验没有证明任何事情，可能是肤浅使青少年发短信更多；也可能肤浅和短信过度是由一个共同因素引起的，例如基因，如果可能的话，对该基因因素进行干预，可以避免这两个方面的问题。

对一个变量进行干预与以该变量为条件的区别是很明显的。当干预模型中的一个变量时，我们固定这个变量的值，这意味着改变了系统，其他变量的值通常会因此发生变化。当以一个变量为条件时，我们不做任何改变；仅仅关注问题的子集，在这个子集中，变量的值都是我们感兴趣的，这里改变的是我们对世界的看法，而不是世界本身。

例如，图3.1展示了冰激凌销量例子的图模型， X 表示冰激凌销量， Y 表示犯罪率， Z 表示温度。当采取干预措施、固定变量的值时，意味着削弱了该变量为响应其他变量而变化的自然趋势。这相当于在图模型上进行一种处理，即删除指向该变量的所有边。如果采取的干预措施是降低冰激凌销量（比如，关闭所有的冰激凌店），将得到如图3.2所示的图模型。检验图3.2中的相关性可以发现，犯罪率与冰激凌的销量完全独立（即不相关），这是因为后者不再与温度（2）相关。换句话说，即使改变了固定值的水平，这个变化也不会传递到变量 Y （犯罪率）。也就是说，干预一个变量会产生一种与以变量为条件完全不同的依赖模式。此外，以变量为条件可以用第1章中描述的方法直接从数据集中获得，而干预的变化依赖于因果图的结构。对于任何给定的干预，可以根据图模型来确定应该删除哪些边。

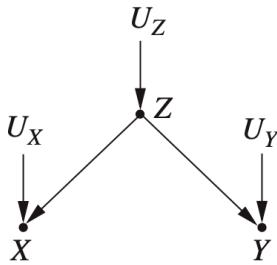


Figure 3.1 A graphical model representing the relationship between temperature (Z), ice cream sales (X), and crime rates (Y)

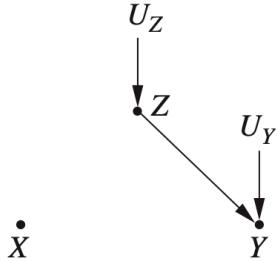


Figure 3.2 A graphical model representing an intervention on the model in Figure 3.1 that lowers ice cream sales

我们在符号上区分变量 X 自然地取值 x 的情况和固定 X 取值 x 的情况，后者用 $do(X = x)$ 来表示。因此， $P(Y = y | X = x)$ 表示在 $X = x$ 的条件下 $Y = y$ 的概率：

$P(Y = y | do(X = x))$ 表示通过干预使 $X = x$ 时 $Y = y$ 的概率。以分布的术语来说，

$P(Y = y | X = x)$ 反映了在 X 的值都是 x 的个体上 Y 的总体分布；另一方面，

$P(Y = y | do(X = x))$ 反映了如果群体中的每个个体均将 X 值固定为 x 时， Y 的总体分布。

类似地，用 $P(Y = y | do(X = x), Z = z)$ 表示对于给定的 $Z = z$ ，干预 $do(X = x)$ 得到的分布中 $Y = y$ 的条件概率。

利用 do -表达式和图模型，可以将因果关系从相关关系中分解出来。在这一章剩下的部分，我们将学习一种通过单纯地观察数据就能神奇地分解出因果关系的方法，当然，首先要假设图是实际问题的有效表述。箭要注意的是，默认假设的干预不会造成其他影响，也就是说，当对一个个体给变量 X 分配值 x 时，不会直接改变其他变量的值。比如，给一个患者分配一种药物和违背了他的宗教信仰强迫其服用药物，在恢复上可能有不同的效果，当改变了其他变量的值时，这些改变必须在模型中明确地表示出来。

3.2 校正公式

冰激淋的例子代表了一种极端的情况，在这个例子中， X 和 Y 之间的相关性从因果角度完全是假设的，因为从 X 到 Y 没有因果路径，但现实生活中大多数情况并不那么明确。为了探讨一个更现实的情况，我们来分析图3.3，其中 X 代表使用药物， Y 代表痊愈， Z 代表性别， Z 和 X 都对 Y 有影响，这个模型实际上反映的就是辛音森悖论。为了确定药物在人群中的有效性，设想一种假设性的干预措施，即对整个人人统一服用这种药物，并与补充干预下的痊愈率进行比较，补充干预指阻止每个服用药物。用 $do(X = 1)$ 表示第一种干预，用 $do(X = 0)$ 表示第二种干预，现在要估计它们的差异。

$$P(Y = 1 | do(X = 1)) - P(Y = 1 | do(X = 0)) \quad (3.1)$$

