贝叶斯分析

第一章：贝叶斯规则及应用

1、三种信息

统计学中有两个主要学派：频率学派与贝叶斯学派，他们之间既存在共同点，又存在不同点，主要与总体信息、样本信息和先验信息有关。基于总体信息和样本信息进行的统计推断被称为经典统计学，它的基本观点是统计推断是根据样本信息对总体分布或总体的特征数进行推断。贝叶斯学派观点除了上述两种信息以外，统计推断还应该使用第三种信息，即先验信息。下边来详细介绍这三种信息。

总体信息,即总体分布或总体所属分布族给我们的信息，譬如“总体是正态分布”这一句话就给我们带来很多信息:它的密度函数是一条钟形曲线;它的一切阶矩都存在;有关正态变量(服从正态分布的随机变量)的一些事件的概率可以计算;有正态分布可以导出 x^2 分布、t 分布和 F 分布等重要分布;还有许多成熟的点估计、区间估计和假设检验方法可供我们选用。总体信息是很重要的信息,为了获取此种信息往往耗资巨大。现实生活中美国军界为了获得某种新的电子元器件的寿命分布，常常购买成子上万个此种元器件,做大量寿命实验、获得大量数据后才能确认其寿命分布是什么。我国为确认国产轴承寿命分布服从两参数威布尔分布前后也花了五年时间,处理几千个数据后才定下的。又如保险费的确定与人的寿命分布密切相关,在保险业中,人的寿命分布被称为寿命表，中国人的寿命表不同于外国人的寿命表,男人的寿命表不同于女人的寿命表,北方人的寿命表不同于南方人的寿命表,当代人的寿命表与 50 年前人的寿命表也是不同的,而要确定这些寿命表是一项耗资费时的工作,至今我国还缺乏此类寿命表。确定我国各类人群的寿命表是我国统计工作者的重要任务。

样本信息,即从总体抽取的样本给我们提供的信息。这是最“新鲜”的信息,并且愈多愈好。人们希望通过对样本的加工和处理对总体的某些特征作出较为精确的统计推断。没有样本就没有统计学可言。这是大家都理解的事实。人们希望通过对样本的加工和处理对总体的某些特征做出较为精确的统计推断。例如有了样本观察值，我们可根据它大概知道总体的一些特征数(均值、方差等)在一个什么范围内。

先验信息,即在抽样之前有关统计问题的一些信息,一般说来,先验信息主要来源于经验和历史资料。先验信息在日常生活和工作中也经常可见，不少人在自觉地或不自觉地使用它。看下面两个例子：英国统计学家 Savagc(1961)曾考察如下两个统计实验:一位常饮牛奶加茶的妇女声称,她能辨别先倒进杯子里的是茶还是牛奶，对此做了十次试验,她都正确地说出了；

一位音乐家声称,他能从一页乐谱辨别出是海顿还是莫扎特的作品，在十次这样的试验中,他都能正确辨别。这些被实验者都是猜测吗？我们假设被实验者都在猜测，每次成功概率 0.5，那十次都猜中概率 2^{-10}=0.0009766，这几乎不可能发生，“每次成功概率为 0.5”的假设应被拒绝。结果中实验者每次成功概率比 0.5 大的多，可见经验在推断中不可忽视。

2、两大学派差异

贝叶斯统计学派与经典统计学派在很多问题上都有分歧，其中最根本的分歧是:

第一，是否利用先验信息。由于产品设计、生产都有一定继承性，故存在许多相关产品的信息以及先验信息利用，Byaes 统计学派认为利用这些先验信息不仅可以减少样本容量，而且在很多情况还可以提高统计精度;而经典统计学派忽略了这些信息。

第二，是否将参数θ看成随机变量。Byaes 统计学派的最基本的观点是:任一未知量θ都可以看成随机变量，可以用一个概率分布去描述，这个分布就是先验分布。因为任一未知量都具有不确定性，而在表述不确定性时，概率与概率分布是最好的语言;相反，经典统计学派却把未知量θ就简单看成一个未知参数，来对它进行统计推断。

3、贝叶斯公式的密度函数形式

现使用随机变量的密度函数叙述贝叶斯公式：

1）依赖于参数θ的密度函数在经典统计中记为 $p(x;\theta)$或 $p_\theta(x)$，它表示在参数空间 $\Theta=\{\theta\}$ 中不同的θ对应不同的分布。可在贝叶斯统计中记为 $p(x\mid\theta)$,它表示在随机变量θ给定某个值时,总体的条件概率函数。

2）根据参数θ的先验信息确定先验分布 $\pi(\theta)$。这是贝叶斯学派在最近几 1 年里重点研究的问题，已获得一大批富有成效的方法。

3）从贝叶斯观点看,样本 $x=(x_1,\cdots,x_n)$的产生要分二步进行。首先设想从先验分布$\pi(\theta)$产生一个样本θ'，这步人们是无法看到的。第二步是从总体分布 $p(x_i|\theta')$产生一个样本 $x=(x_1,\cdots,x_n)$，这个样本是具体的，人们能看得到的，此样本 x 发生的概率是与联合密函数成正比：$p(x|\theta')=\prod_{i=1}^{n}p(x_i|\theta')$，这个联合密度函数综合了总体信息和样本信息，常称为似然函数，记为 $L(\theta')$。频率学派和贝叶斯学派都承认似然函数，两派认为：在有了样本观察值 $x=(x_1,\cdots,x_n)$后,总体和样本中所含θ的信息都被包含在似然函数 $L(\theta')$之中,可在使用似然函数作统计推断时，两派之间还是有差异的。

4）由于θ'是设想出来的,它仍然是未知的,它是按先验分布$\pi(\theta)$而产生的,要把先验信息进行综合,不能只考虑θ',而应对θ的一切可能加以考虑。故要用$\pi(\theta)$参与进一步综合。这样一来,样本 x 和参数θ的联合分布 $h(x,\theta)=p(x\mid\theta)\pi(\theta)$把三种可用的信息都综合进去。

5）要对未知数θ作出统计推断，在没有样本信息时,只能根据先验分布对θ作出推断。在有样本观察值 $x=(x_1,\cdots,x_n)$之后,应该依据 h(x,θ)对θ作出推断。为此我们把 h(x,θ)分解成 h(x,θ)=π(θ｜x)m(x)，其中 m(x)是 x 的边缘密度函数。$m(x)=\int_\Theta h(x,\theta)d\theta=\int_\Theta p(x|\theta)\pi(\theta)$，它与θ无关，或者 m(x)中不含θ的任何信息。因此能用来对θ作出推断的仅是条件分布 $\pi(\theta|x)$,它的计算公式是$\pi(\theta|x)=\frac{h(x,\theta)}{m(x)}=\frac{p(x|\theta)\pi(\theta)}{\int_\theta p(x|\theta)\pi(\theta)d\theta}$，这就是贝叶斯公式的密度函数形式。这个在样本 x 给定下，θ的条件分布被称为θ的后验分布。它是集中了总体、样本和先验等三种信息中有关θ的一切信息,而父是排除一切与θ无关的信息之后所得到的结果。故基于后验分布π(θ|x)对θ进行统计推断是更为有效,也是最合理的。

6）在θ是离散随机变量时，先验分布可用先验分布列$\pi(\theta_i)$,i=1,2,…表示。这时后验分布也是离散形式，即$\pi(\theta_i|x)=\frac{p(x|\theta_i)\pi(\theta_i)}{\sum_j p(x|\theta_j)\pi(\theta_j)},i=1,2,\cdots$

4、后验分布是三种信息的综合

一般说来,先验分布π(θ)是反映人们在抽样前对θ的认识后验分布π(θ|x)是反映人们在抽样后对θ的认识。之间的差异是由于样本 x 出现后人们对θ认识的一种调整。所以后验分布π(θ|x)可以看作是人们用总体信息和样本信息(综合称为抽样信息)对先验分布π(θ)作调整的结果。

设事件 A 的概率为θ，即 π(A)=θ。为了估计θ而作 n 此独立观察，其中事件 A 出现的次数为 X，X 服从二项分布 b(n,x)，即 $P(X=x|\theta)=\binom{n}{x}\theta^x(1-\theta)^{n-x},x=0,1,\cdots,n$，这就是似然函数。假若我们试验前对事件 A 没有什么了解，从而对其发生的概率θ也没有任何信息。在这种场合，贝叶斯本人建议采用“同等无知”的原则使用区间(0,1)上的均匀分布 U(0,1)作为θ的先验分布，因为它取(0,1)上的每一点的机会均等。贝叶斯的这个建议被后人称为贝叶斯假设。这时θ的先验分布为$\pi(\theta)=\begin{cases}1,0<\theta<1\\0,\text{其他}\end{cases}$，由此即可利用贝叶斯公式求出θ的后验分布，为了综合抽样信息和先验信息，可利用贝叶斯公式，为此先计算出样本 X 与参数θ的联合分布 $h(x,\theta)=\binom{n}{x}\theta^x(1-\theta)^{n-x},x=0,1,\cdots,n,0<\theta<1$，此式在定义域上与二项分布有差别。再计算 X 的边缘分布 $m(x)=\int_0^1 h(x,\theta)d\theta=\binom{n}{x}\frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)}$，最后得到θ的后验分布 $\pi(\theta|x)=\frac{h(x,\theta)}{m(x)}=\frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)}\theta^{(x+1)-1}(1-\theta)^{(n-x+1)-1},0<\theta<1$，这个分布就是参数为 x+1 的 n-x+1 的贝塔分布，这个分布记为 B(x+1，n-x+1)。

举一个例子，为了提高某产品质量，公司经理考虑增加投资来改进生产设备，预计需投资 90 万元，但从投资效果看，下属部门有两种意见：θ_1：改进生产设备后,高质量产品可占 90%。θ_2：改进生产设备后,高质量产品可占 70%。经理当然希望θ_1发生,公司效益可得很大提高,投资改进设备也是合算的，但根据下属两个部门过去建议被采纳的情况，经理认为，θ_1的可信程度只有 40%，θ_2的的可信程度是 60%。即 π(θ_1)=0.4，π(θ_2)=0.6。这两个都是经理的主观概率。经理想慎重一些,通过小规模试验后观其结果再定。为此做了一项试验，试验结果(记为 A)如下:A:试制五个产品,全是高质量的产品。经理希望用此试验结果修改原先对θ_1和θ_2的看法。根据题意，即计算后验概率π(θ_1｜A)与π(θ_2｜A)。这可以使用贝叶斯公式的离散形式来计算。已有先验概率π(θ_1)与π(θ_2)，计算两个条件概率即使用二项分布计算 P(A｜θ_1)与 P(A｜θ_2)，P(A｜θ_1)=0.9^5≈0.590，P(A｜θ_2)=0.7^5≈0.168，由全概率公式得到 P(A)=P(A｜θ_1)π(θ_1)+P(A｜θ_2)π(θ_2)=0.337，最后使用贝叶斯公式的离散形式计算 π(θ_1 ｜A)=P(A｜ θ_1)π(θ_1)/P(A)=0.236/0.337=0.700 ， π(θ_2 ｜A)=P(A｜ θ_2)π(θ_2)/P(A)=0.101/0.337=0.300，这表明经理根据实验 A 的信息将θ_1和θ_2 可信程度调整为 0.7 和 0.3，这就是贝叶斯的应用。

经过实验 A 后,经理对增加投资改进质量的兴趣增大。但因投资额大,还想再做一次小规模试验,观其结果再作决策。为此又做了一批试验,试验结果(记为 B)如下:B:试制 10 个产品,有 9 个是高质量产品。经理对此试验结果更为高兴。希望用此试验结果对 θ 1与 θ 2再作一次调整。为此把上次后验概率看作这次的先验概率,即 π(θ_1)=0.7，π(θ_2)=0.3，用二项分布可算得 P(B｜θ_1)=$10(0.9)^9$ (0.1)≈0.387， P(B｜θ_2)=$10(0.7)^9$ (0.3)≈0.121，由此可算得 P(B)=0.307 和后验概率π(θ_1｜B)=0.883，π(θ_2｜B)=0.117。经理看到，经过两次试验，θ_1（高质量产品可占 90%）的概率已上升到 0.883，能以 88.3%的把握保证此项投资能取得较大经济效益。

第二章：贝叶斯估计

1、贝叶斯估计推导

在贝叶斯估计的框架下，假设参数θ是一个随机变量，服从某种先验分布。在给定观测数据 D 的情况下，通过贝叶斯定理计算后验分布 p(θ ｜D)。贝叶斯估计的本质是通过贝叶斯决策得到参数 θ 的最优估计，使得总期望风险最小。即 p(θ｜D)在真实的θ值处有一个概率尖峰。我们对贝叶斯估计进行简单的推导。设 p(θ)是待估计参数 θ 的先验概率， θ 取值与样本集 D 有关，这里 D 取值为 x1,x2 到 xn。那么先验分布即为 p(θ)，样本信息为在给定参数 θ 的条件下，数据 D 发生的概率。将这两个式子代入到贝叶斯定理中，就可以得到后

验信息即在给定数据 D 的条件下，数据 θ 发生的概率，由此我们即基于后验信息进行统计推断。不同于极大似然估计，不再把参数θ看成一个未知的确定变量，而是看成未知的随机变量，通过对第 i 类样本 Di 的观察，使概率分布 $p(D|\theta)$转化为后验概率 $p(\theta|D)$，再求贝叶斯估计。

在统计推断中，我们常常需要寻找一个最优的估计量来对未知参数进行估计。这个最优估计量的选择通常是为了使得总期望风险最小化。假设我们有一个样本取值空间为 d 维欧式空间，参数取值空间为Φ。我们希望找到一个估计量$\hat{\theta}$，$\lambda(\hat{\theta},\theta)$是$\hat{\theta}$作为θ的估计量时的损失函数。条件风险表示在给定样本 x 的条件下，估计量$\hat{\theta}$对应的条件风险，等于损失函数$\lambda(\hat{\theta},\theta)$在后验概率密度函数 $p(\theta|x)$下的期望值。则总期望风险 R 为样本 x 下的条件风险对样本的概率密度函数 p(x) 的期望值。接下来，我们讨论在有限样本集合 D 取值 x1 到 xn 的情况下，如何求得使总期望风险最小的估计量θ∗。根据前面的公式，可以看出让总期望风险最小，那么就要让 D 的条件下参数θ的条件风险最小，所以我们的任务就转化为求给定样本集合 D 的情况下，寻找使得条件风险 $R(\hat{\theta}|D)$ 最小的估计量$\hat{\theta}$的表达式。 前面我们已经给出了最优估计量的计算方法。当损失函数是均方误差损失函数时，即损失函数 λ 等于参数 θ 和估计量$\hat{\theta}$之差的平方，此时，在样本 x 条件下的θ的贝叶斯估计量θ^*是在给定 x 下的条件期望：也就是参数 θ 在后验概率密度函数 $p(\theta|x)$下的条件期望。而对于一般的样本集 D，θ的贝叶斯估计量就是给定 D 下的条件期望。

2、贝叶斯估计基本步骤

至此，我们得出贝叶斯估计的基本步骤。首先需要确定参数θ的先验分布 $p(\theta)$。然后我们通过样本集合 D 取值 x_1, x_2 到 x_N，求出样本的联合分布 $p(D|\theta)$。之后利用贝叶斯公式，求得θ的后验分布，最终通过计算θ的后验分布的期望值，得到贝叶斯估计 θ^*。贝叶斯估计是参数θ在观察到样本数据后参数θ的条件期望。

下面我们讨论贝叶斯估计方法的优缺点：首先，我们来看贝叶斯估计方法优点：第一，贝叶斯估计方法中所有参数都被看作是随机变量，这意味着我们可以更容易地构建复杂的概率模型。相比之下，经典方法通常只考虑参数的点估计，可能会受限于简单模型的假设。第二，贝叶斯估计方法能够充分利用先验信息，结合似然原则，从而得到更精准的统计推断。先验信息是我们对参数的经验信念或知识，可以来自领域专家经验或历史数据。通过将先验信息与样本数据相结合，贝叶斯方法能够提供更准确的参数估计结果。

接着，我们来看贝叶斯估计方法的缺点：第一，贝叶斯估计方法具有一定主观性。由于先验信息的选择可能涉及主观判断，因此可能与科学的客观性相矛盾。不同的先验分布可能

导致不同的后验结果，这需要谨慎处理。第二，随着数据量的增加，先验的影响力会逐渐减小。在贝叶斯估计中，随着样本数据量的增加，后验概率会趋近于对应的似然函数，先验信息的影响逐渐减弱。这可能导致在大数据场景下，贝叶斯估计与经典方法趋于一致。值得一提的是，如果我们选择先验为均匀分布，那么贝叶斯方法等价于经典方法。这是因为均匀分布的先验不会对参数产生任何偏好，从而在一定程度上消除了主观性。

下面我们来看一道例题。假设总体 X 服从参数为θ的伯努利分布，其中θ的取值范围是 $\Theta=(0,1)$。我们有一个样本集合 $x_1, x_2, ..., x_n$，它们来自总体 X。在这个问题中，我们使用均方误差损失函数 $\lambda(\theta, a)=(\theta-a)^2$，其中 a 是我们的估计量。参数θ服从均匀分布，先验分布为 U[0,1]。现在，让我们来求参数 θ 的贝叶斯估计值。首先求样本的概率密度函数 f(x)，也就是在给定样本集合 $x_1, x_2, ..., x_n$ 的情况下，对参数 θ 进行积分得到的结果。这个密度最终可以表示成 Beta 函数的形式。对于密度函数 f(x)，其中的 Beta 函数是一类有广泛应用的函数，是这样一个从 0 到 1 的区间内对变量 u 的积分表达式。Beta 函数与 Γ 函数之间的关系是 $B(s,t)=\frac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)}$。这个关系是一个重要的数学等式，它将 Beta 函数与 Γ 函数相联系，使得我们可以通过 Γ 函数的性质来求解 Beta 函数的值。由此可得到后验分布的密度函数 p(θ | x)的表达式。最后再由 Beta 函数与 Γ 函数之间的关系，求出有 θ 的 Bayes 估计为函数 T(x)，也就是在给定样本集合 $x_1, x_2, ..., x_n$ 的情况下，对参数 θ 进行条件期望计算，得到最终 θ 的 Bayes 估计为$\frac{1}{n+2}(1+\sum_{i=1}^{n}x_i)$

我们再来看一道例题，针对正态分布的情况，参数 θ 仅有均值 μ 未知，而方差已知。给定样本 D，$D\sim N(\mu, \sigma^2)$，均值变量的先验分布 $\mu\sim N(\mu_0, \sigma_0^2)$，求 μ 的贝叶斯估计。为了求得这一目标，首先我们需要计算 μ 的后验概率密度函数 $p(\mu|D)$，根据贝叶斯公式，后验概率密度函数可以表示为先验概率 p(μ) 与似然函数 p(D | μ)的乘积，再除以边缘概率 p(D)。简化表达式后，可以得到后验概率密度函数的表达式，我们使用 α 吸收了所有与 μ 无关的项。为了计算后验概率密度函数，我们对样本数据进行观察。根据题意知道样本 D 服从正态分布 $D\sim N(\mu, \sigma^2)$，均值变量的先验分布 $\mu\sim N(\mu_0, \sigma_0^2)$，将对应的表达式代入到后验概率密度函数的计算中，得到样本中每个数据的似然概率与先验概率的乘积，观察式子我们可以发现 $p(\mu|D)$是应该服从正态分布，不妨假设其均值为 μ_N，方差为 σ_N^2。由于 μ 的后验概率服从正态分布，因此由两式指数项中对应的系数相等，我们可以解出 μ_N与 σ_N^2的值如下。于是，就完全确定了 μ 的后验概率密度函数，我们即可求得 μ 的贝叶斯估计值，即 $\widehat{\mu}=$

对参数 μ 的后验概率密度函数求条件期望，利用正态分布期望的性质，从而 μ 的贝叶斯估计值为 μ_N。注意，当 σ_0^2=0 时，μ 的贝叶斯估计为 μ_0，此时先验知识过于强大(确定了 μ 的值)，样本不起作用；当 σ_0^2远大于 σ^2时，μ 的贝叶斯估计趋近于 $\widehat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} x_i$，此时先验信息作用基本为 0，贝叶斯估计完全依靠样本信息。

第三章：贝叶斯网络

1、不确定性推理与联合概率分布

不确定性推理是人工智能研究的重要课题之一.从 20 世纪六七十年代以来人们提出了多种方法，如概率方法、非单调逻辑、确定因子、Dempster-Shafer 证据理论、模糊逻辑等，在这些方法中，概率方法是最自然也是最早被尝试的方法之一，因为概率论本身是关于随机现象和不确定性的数学理论.使用概率方法进行不确定性推理就是:①把问题用一组随机变量 $X = \{X_1, X_2 \cdots, X_n\}$来刻画;②把关于问题的知识表示为一个联合概率分布 P(X);③按照概率论原则进行推理计算，下面来看一个例子。

Pearl 教授家住洛杉矶，那里地震和盗窃时有发生。教授家里装有警铃，地震和盗窃都可能触发警铃，听到警铃后，两个邻居 Mary 和 John 可能会打电话给他。一天，Pearl 教授接到 Mary 的电话，说听到他家警铃响，Pearl 教授想知道他家遭盗窃的概率是多大?这个问题包含 5 个随机变量:盗窃(B)、地震(E)、警铃响(A)、接到 John 的电话(J)和接到 Mary 的电话(M);所有变量的取值均是“y”或“n”，这里各变量间的关系存在不确定性：盗窃和地震以一定概率随机发生;它们发生后，并不一定触发警铃;而警铃响后，Mary 和 John 可能会因为某些原因，如在听摇滚乐或听力问题，而没有听到警铃;有时候，两人也会将其它声音误听为警铃声。假设 Pearl 教授对这 5 个变量的联合概率分布 P(B，E，A，J，M)的评估如表 1 所示。要计算的是接到 Mary 的电话(M=y)后，Pearl 教授对家里遭盗(B=y)的信度，即 P(B=y｜M=y)。

表 1：Alarm 问题的联合概率分布 P(B,E,A,J,M)

B	E	A	M	J	概率	B	E	A	M	J	概率
y	y	y	y	y	1.2E-4	n	y	y	y	y	3.6E-3
y	y	y	y	n	5.1E-5	n	y	y	y	n	1.6E-3
y	y	y	n	y	1.3E-5	n	y	y	n	y	4.0E-4
y	y	y	n	n	5.7E-6	n	y	y	n	n	1.7E-4
y	y	n	y	y	5.0E-9	n	y	n	y	y	7.0E-6
y	y	n	y	n	4.9E-7	n	y	n	y	n	6.9E-4
y	y	n	n	y	9.5E-8	n	y	n	n	y	1.3E-4
y	y	n	n	n	9.4E-6	n	y	n	n	n	1.3E-2
y	n	y	y	y	5.8E-3	n	n	y	y	y	6.1E-4
y	n	y	y	n	2.5E-3	n	n	y	y	n	2.6E-4
y	n	y	n	y	6.5E-4	n	n	y	n	y	6.8E-5
y	n	y	n	n	2.8E-4	n	n	y	n	n	2.9E-5
y	n	n	y	y	2.9E-7	n	n	n	y	y	4.8E-4
y	n	n	y	n	2.9E-5	n	n	n	y	n	4.8E-2
y	n	n	n	y	5.6E-6	n	n	n	n	y	9.2E-3
y	n	n	n	n	5.5E-4	n	n	n	n	n	9.1E-1

从联合概率分布 $\mathrm{P(B,E,A,J,M)}$出发，先计算边缘分布 $\mathrm{P(B,M)}=\sum_{\mathrm{E,A,J}}\mathrm{P(B,E,A,J,M)}$，得到下表 2。

表 2 边缘分布

B	M	P (B, M)
y	y	0.000115
y	n	0.000075
n	y	0.00015
n	n	0.99966

再按照条件概率定义，得

P(B=y｜M=y)=P(B=y,M=y)/P(M=y)=P(B=y,M=y)/(P(B=y,M=y)+P(B=n,M=y)

=(0.000115)/(0.000115+0.000075)≈0.61

2、条件独立与联合分布的分解

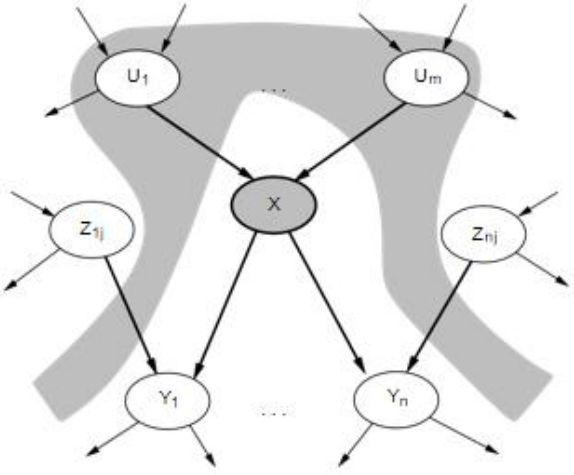
利用变量间的条件独立关系可以将联合分布分解成多个复杂度较低的概率分布，从而降低模型表达的复杂度，提高推理效率，使得人们可以应用概率方法来解决大型问题。

基于 Alarm 问题，运用链规则，将联合概率分布 $P(B, E, A, J, M)$分解成复杂度低的乘积形式。可以把联合概率分布 $P(B, E, A, J, M)$表示为

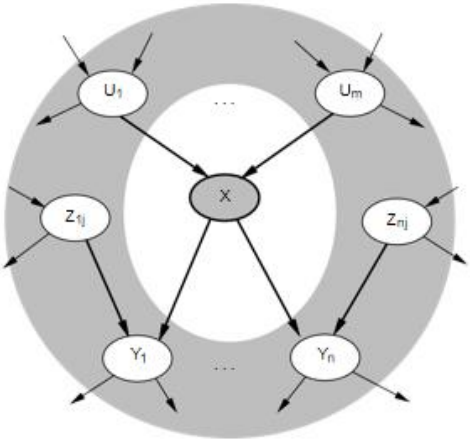
$P(B,E,A,J,M)=P(B)P(E|B)P(A|B,E)P(J|B,E,A)P(M|B,E,A,J)$，然而这一步并没有降低模型复杂度，但是它使得可以根据问题的背景知识做一些合理独立假设降低复杂度，例如，地震(E)应与盗窃(B)无关，于是假设 E 与 B 相互独立，即 $P(E|B)=P(E)$,这样就把 $P(E|B)$简化成了 $P(E)$。另外，John(J)和 Mary(M)是否打电话直接取决于他们是否听到警铃(A)，所以可以假设给定 A 时，J 与 B 和 E，以及 M 与 J、B 和 E 都相互独立。即 $P(J|B,E,A)=P(J|A)$和 $P(M|B,E,A,J)=P(M|A)$。将这些独立假设放在一起，得到复杂度低的概率分布乘积形式：$P(B,E,A,J,M)=P(B)P(E)P(A|B,E)P(J|A)P(M|A)$，一般地，考虑一个包含 n 个变量的联合分布 $P(X_1,\cdots,X_n)$，利用链规则，可以写为

$P(X_1,\cdots,X_n)=P(X_1)P(X_2|X_1)\cdots P(X_n|X_1,X_2,\cdots,X_{n-1})=\prod_{i=1}^{n}P(X_i|X_1,X_2,\cdots,X_{i-1})$，对于任意$X_i$，如果存在 $\pi(X_i)\subseteq\{X_1,\cdots,X_{i-1}\}$，使得给定 $\pi(X_i)$，X_1与$\{X_1,\cdots,X_{i-1}\}$中其他变量条件独立，即 $P(X_i|X_1,X_2,\cdots,X_{i-1})=P\left(X_i\middle|\pi(X_i)\right)$，那么有 $P(X_1,\cdots,X_n)=\prod_{i=1}^{n}P\left(X_i\middle|\pi(X_i)\right)$，这样就得到了联合分布的一个分解，其中当$\pi(X_i)=\emptyset$时，$P(X_i|\pi(X_i))$为边缘分布 $P(X_i)$。

下面简单介绍全局语义和局部语义。全局语义将整个联合分布定义为局部条件分布的乘积，而局部语义表示每个节点在给定其父节点的情况下，与其非后代节点条件独立。如图所示，节点和箭头展示了随机变量及其条件依赖关系，强调了在父节点已知的情况下，节点的概率分布与其他非后代节点无关。这些概念是构建和理解复杂概率模型的基础。



每个节点在给定其马尔可夫毯的情况下，与其他所有节点都是条件独立的。马尔可夫毯包括该节点的父节点、子节点以及子节点的父节点。图示中，节点 X 的马尔可夫毯被高亮显示，展示了其父节点 $U_1,\ldots,Um$、子节点 $Y_1,\ldots,Y_n$以及子节点的父节点 $Z_{1j},\ldots,Z_{nj}$。



3、贝叶斯网络概念

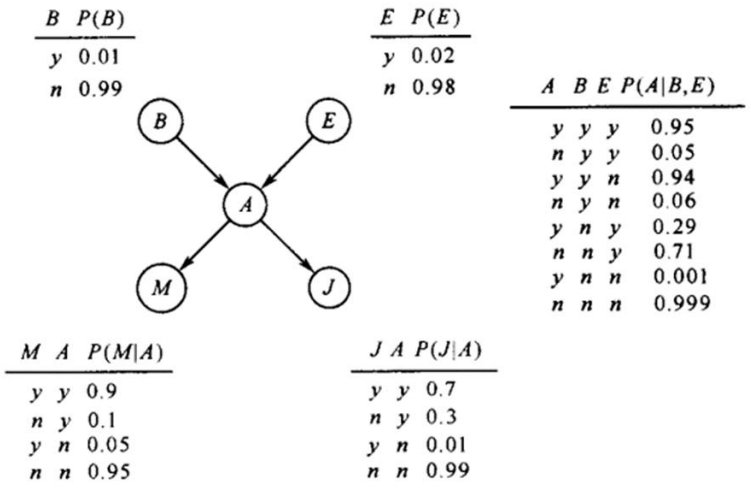
前面公式 $P(X_1,\cdots,X_n)=\prod_{i=1}^{n}P\left(X_i\middle|\pi(X_i)\right)$中，变量$X_i$的分布直接依赖于$\pi(X_i)$的取值，如果给定$\pi(X_i)$，则$X_i$与$\{X_1,\cdots,X_{i-1}\}$中的其他变量条件独立。Pearl 提出用如下方法构造一个有向图来表示这些依赖和独立关系:①把每个变量都表示为一个节点;②对于每个节点X_i，都从$\pi(X_i)$中每个节点画一条有向边到X_i。

在一个有向图中，如果从节点 X 到节点 Y 存在一条边，那么称 X 为 Y 的父节点，而 Y 为 X 的子节点。一个节点的所有父节点和子节点称为它的邻居节点。没有父节点的节点称为根节点，没有子节点的节点则称为叶节点。一个节点的祖先节点包括其父节点及父节点的祖先节点，根节点无祖先节点。一个节点的后代节点包括其子节点及子节点的后代节点，叶节点无后代节点。一个节点的非后代节点包括所有不是其后代节点的节点。节点 X 的父节点为 pa(X)或$\pi(X)$，子节点为 ch(X),邻居节点为 nb(X),祖先节点为 an(X),后代节点为 de(X),非后代节点为 nd(X)。在一有向图中，若某节点是它自己祖先节点，则该图包含一有向环。有向无环图是不含有向环的有向图。贝叶斯网是一个有向无环图，其中节点代表随机变量，节点间的边代表变量之间的直接依赖关系，每个节点都附有一个概率分布，根节点 X 所附的是它的边缘分布 P(X)，而非根节点 X 所附的是条件概率分布 $P(X|\pi(X))$。

贝叶斯网可以从定性和定量两个层面来理解，在定性层面，它用一个有向无环图描述了变量之间的依赖和独立关系；在定量层面，它则用条件概率分布刻画了变量对其父节点的依赖关系。在语义上，贝叶斯网是联合概率分布的分解的一种表示。更具体地，假设网络中的变量为$X_1,\cdots,X_n$，那么把各变量所附的概率分布相乘就得到联合分布，即 $P(X_1,\cdots,X_n)=\prod_{i=1}^{n}P\left(X_i\middle|\pi(X_i)\right)$，其中当$\pi(X_i)=\emptyset$时，$P(X_i|\pi(X_i))$为边缘分布 $P(X_i)$。联合概率分布的分解降低了概率模型的复杂度，贝叶斯网的引人虽然没有进步降低复杂度，但它为概率推理提供了很大的方便，这主要是因为贝叶斯网一方面是严格的数学语言，适合于计算机的处理;

另一方面，它又直观易，方便人们讨论交流和建立模型，另外，Pearl 还认为，贝叶斯网提供了人脑推理过程的一个模型，因为依赖和独立关系是人们日常推理的基本工具，而且人类知识的基本结构也可以用依赖图来表达，事实上，越来越多的研究领域开始采用贝叶斯网来展示问题的结构，从而使贝叶斯网的影响远远超出了不确定性推理和人工智能的范围。

基于 Alarm 问题，构建贝叶斯网。在前面的问题中可以得到 π (B)=π(E)=∅，π(A)={B,E},π(J)={A}，π(M)={A}。根据 Pearl 提出的方法，得到有向图。观察有向图：A 依赖于 B 和 E，M 和 J 都依赖于 A。根据条件概率分布 P(A | B,E)可以定量 A 是如何依赖于 B 和 E。当盗窃和地震都发生时，警铃响的概率 P(A=y | B=y,E=y)是 0.95;当只发生盗窃但没有发生地震时，警铃响的概率 P(A=y | B=y,E=n)是 0.29;当只发生地震但没有发生盗窃时，警铃响的概率 P(A=y | B=n,E=y)是 0.94;而当盗窃和地震都没发生时，警铃响的概率 P(A=y | B=n,E=n)是 0.001。类似地，P(M | A)和 P(J | A)分别定量刻画了 M 和 J 如何依赖于 A，变量 B 和 E 不依赖于其它变量，P(B)和 P(E)给出它们的边缘分布。将所示的有向图与这 5 个概率分布合在一起，就构成一个贝叶斯网，如图所示。



4、贝叶斯网络的构造

贝叶斯网的构造方法有两种，一种是通过咨询专家手工构造，另一种是通过数据分析来获得。本节详细讨论如何手工构造贝叶斯网，这包括确定网络结构和评估条件概率两个子任务。确定网络结构分为以下几个步骤：

(1)选定一组刻画问题的随机变量{X1,X2,⋯,Xn }

(2)选择一个变量顺序α=⟨X1,X2,⋯,Xn ⟩

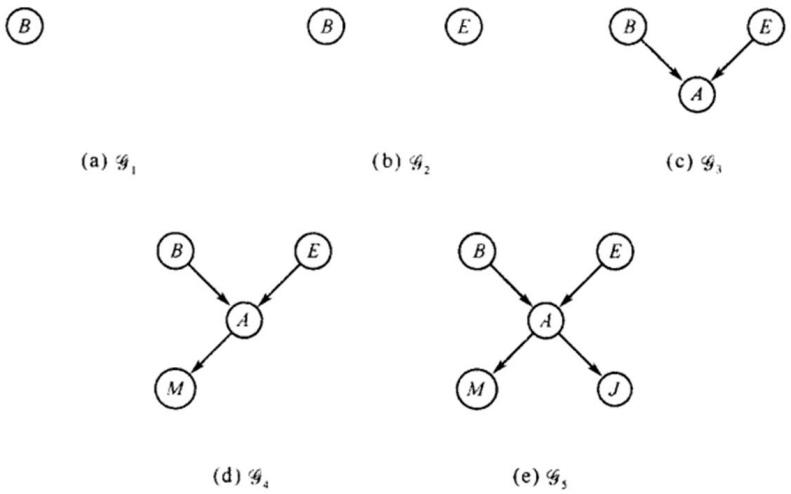
(3)从一个空图出发，按照顺序α逐个将变量加入ξ中

(4)在加入变量 X_i 时，ξ中的变量包括 X1,X2,⋯,X(i-1)

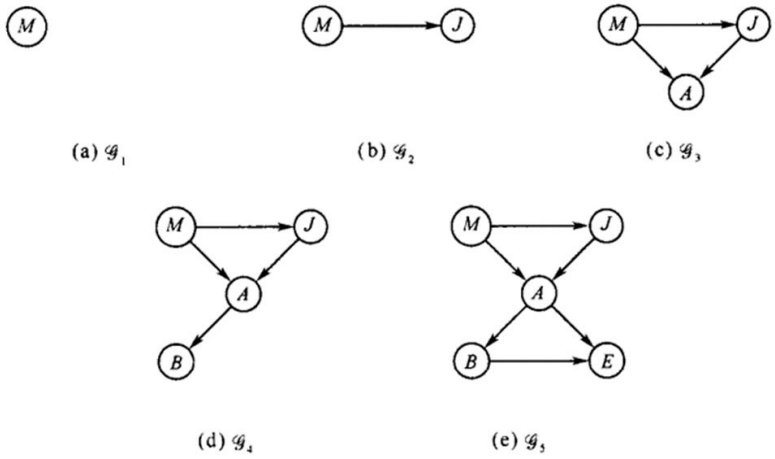
①利用问题的背景知识，在这些变量中选择一个尽可能小的子集π(Xi)，使得假设“给定π(Xi)，Xi 与ξ中的其它变量条件独立”合理;

②从 π(Xi)中的每一个节点添加一条指向 X_i 的有向边.

基于 Alarm 问题，此问题中涉及 5 个随机变量，假设选用序 α_1=<B,E,A,M,J>来构造一个贝叶斯网结构，具体流程为：①首先把 B 加入空图，得到ξ_1;②接着加入 E:假设 B 和 E 相互独立，π(E)=∅，因此无需加边，结果是ξ_2;③然后加入 A:我们认为 A 同时依赖 B 和 E，所以π(A)={B,E}，于是分别从 B 和 E 画一条到 A 的边，得到ξ_3;④之后加入 M:假设给定 A，M 与 B，E 条件独立，所以π(M)={A},于是从 A 画一条到 M 的边，得到ξ_4;⑤最后加入 J:假设给定 A，J 与 B，E 和 M 相互条件独立，所以 π(J)={A}，于是画一条 A 到 J 的边，得到ξ_5。 ξ_1~ξ_5 分别如图(a)~(e)所示



如果按照序 α 2=<M,J,A,B,E>来构造一个贝叶斯网结构，具体过程如图所示。



观察上面两个例题可知，不同的变量顺序导致不同的网络结构，不同的网络结构表示了联合分布的不同分解，而不同的分解则意味着不同的复杂度。一般有三种原则来选择变量顺

序：1）Smith(1989)认为应以模型的复杂度为标准。2）Howard 和 Matheson(1984)认为变量顺序的选取应以条件概率评估的难易程度为标准。3）Pearl(2000)提出应该用因果关系来决定变量顺序，原因在前，结果在后。

4、因果关系与贝叶斯网

在实际应用中，往往利用因果关系来确定贝叶斯网的结构。在利用因果关系建立起来的贝叶斯网中，变量间的边表示的是因果关系，而非简单的概率依赖关联。这样的贝叶斯网称为贝氏因果网，简称因果网，在贝氏因果网中，除了可以进行概率推理外，还可以进行干预的推理以及反事实推理。利用因果关系建立贝叶斯网网络结构有两点需要注意：

1.因果关系没有一个能被广泛接受的严格定义，对它到底是客观世界本身属性，还是人的意识为了理解世界而创造出来主观概念，未能达成共识。例如多数医生认为“抽烟(S)导致肺(L)”，但烟草行业却辩解说“存在一个既诱发抽烟(S)又能导致肺癌(L)的基因(G)”，这两派观点对应两个不同的因果模型。在实际应用中，往往采用如下方式来判断因果关系:假设一个万能的上帝可以介入改变任何变量的状态。对变量 X 和 Y,如果知道 X 的状态被上帝改变了会影响你对 Y 的信度,而反过来知道 Y 的状态被上帝改变并不影响你对 X 的信度,那么就说 X 是 Y 的原因。

2.因果关系与条件独立之间的关系。贝叶斯网蕴含了许多条件独立关系，所以当利用因果关系建立贝叶斯网时，实际上是在基于因果关系进行条件独立的假设，所做的假设可以归纳为因果马尔可夫假设，这个假设是因果关系和条件独立之间的桥梁。

在统计学和机器学习中,理解因果效应是关键的任务之一。因果效应揭示了一个变量(通常是一个干预措施）对另一个变量（通常是结果）的影响。在贝叶斯分析中，我们利用概率模型来量化和推断这些因果关系。因果效应可以通过比较不同干预措施下的潜在结果来衡量。考虑一个简单的例子：我们希望了解服用某种药物（干预措施）对健康状况（结果）的影响。为了说明这一点，我们使用以下符号：

T：观察到的干预措施，例如是否服用药物。

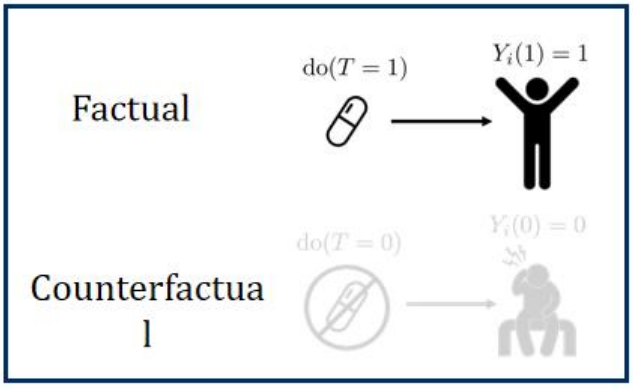
Y：观察到的结果，例如健康状况。

i：特定的个体或单位。

潜在结果用 $Y_i(1)$ 和 $Y_i(0)$ 表示，分别代表在干预和无干预情况下的潜在结果。潜在因果模型的目的：推断干预/政策的效果。

在因果推断中，我们面临的一个核心挑战是如何评估一个干预措施对结果的真实影响。为了更好地理解这个问题，我们需要区分事实（Factual）和反事实（Counterfactual）。假设

我们想了解服用某种药物对健康状况的影响。对于某个个体 i，我们有两个潜在的结果：当个体 i 服用药物时（do(T=1)），其健康状况 $Y_i(1)=1$，表示健康改善；当个体 i 不服用药物时（do(T=0)），其健康状况 $Y_i(0)=0$，表示健康状况未改善。下图清晰地展示了事实和反事实的区别：事实是指实际观察到的结果，而反事实则是指在另一种干预条件下可能发生的结果。



下表列出了多个个体在不同干预措施下的结果。可以看出由于现实条件的限制，我们无法同时观察到同一个体在服药和不服药两种情况下的结果。这意味着我们无法直接得到反事实数据。例如，对于个体 1，我们只能观察到其在不服药情况下的结果 Y=0，但无法知道如果他服药后的结果 $Y_i(1)$。这种数据缺失给因果推断带来了巨大的挑战。

i	T	Y	$Y(1)$	$Y(0)$	$Y(1)-Y(0)$
1	0	0	?	0	?
2	1	1	1	?	?
3	1	0	0	?	?
4	0	0	?	0	?
5	0	1	?	1	?
6	1	1	1	?	?

在因果推断中，我们常常需要评估某种干预措施对结果的影响。随机对照试验（Randomized Controlled Trial, RCT）是一种有效的方法，用来解决因果推断中的基本问题。通过随机分组和控制变量，我们可以公正地估计干预措施的实际效果。随机对照试验通过以下步骤来实现因果推断：1.随机分组：将实验个体随机分为实验组和对照组。实验组接受干预（Treatment），而对照组则不接受干预。这样可以确保两组在其他变量上的差异仅来自干预措施，从而保证比较的公平性。2.实验设计：实验组接受干预，对照组不接受干预；除干预（Treatment）外，其他特征保持相同，以减少混杂因素的影响。通过以下公式，我们可以估计平均处理效应:$\mathrm{E[Y(1)-Y(0)]=E[Y|T=1]-E[Y|T=0]}$。

随机对照试验虽然是解决因果推断问题的重要方法，但也存在诸多问题。首先，成本高昂，通常需要大量资金和资源。其次，实验中很难做到完全的随机化，这可能会影响结果的准确性。例如，在新药测试中，无法确保受试者之间除了用药之外所有其他条件都一致。理想情况下，干预（T）能够直接影响结果（Y），但实际操作中，中介变量（C）也会影响结果。此外，由于这些问题，实际中有时需要依赖观测试验。

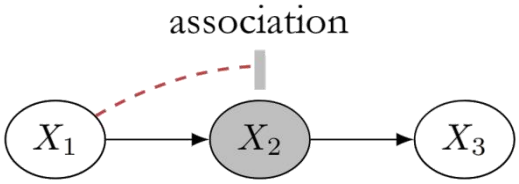
观测试验虽然可以弥补随机对照试验的不足，但也有其自身的局限性。首先，处理措施（Treatment）的随机性无法得到保障，这对因果推断的准确性提出了挑战。伦理原因也是一个重要因素，例如，不能随机让人们吸烟以衡量吸烟对肺癌的影响。此外，后天不可行性比如，不能将国家随机化为社会主义国家和资本主义国家以衡量社会制度对 GDP 的影响；先天不可行性如，不能在出生时改变人的 DNA 以衡量 DNA 对癌症的影响。理想情况下，处理（T）能够直接影响结果（Y），但在观测试验中，随机性无法得到保证，这会影响因果推断的准确性。在混杂因素（C）的影响下，非随机实验中的统计意义上的差异不能代表因果关系的差异，因此无法衡量因果效应。公式 $E[Y(1) - Y(0)] \neq E[Y|T = 1] - E[Y|T = 0]$ 表明了因果关系和统计关联的差异。总结而言，观测试验在混杂因素的影响下，难以准确反映因果关系。

在因果推断中，因果效应估计值是一个定量衡量某变量对结果影响的指标，它可以分为个体因果效应和平均因果效应两种，分别对应个体和群体平均两个维度。为了计算因果效应估计值，首先介绍因果分析中的一个定义，潜在结果。

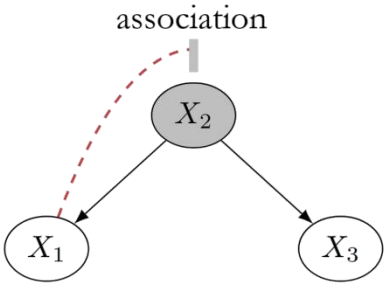
潜在结果理论认为，在某一变量生效之前，模型的最终结果可能会有不同的可能性，这取决于该变量具体的取值，这些不同可能性的结果，都是模型的潜在结果。如一个函数 y=x2，x 取值范围为 2，3，那么 y 的潜在结果就是 4，9。而对于具体的统计数据而言，结果是固定的，这时候观测的数据被称为事实结果，与之相反的，没观测到的，就叫反事实结果。基于潜在结果理论，我们有如下假设：样本数据中，每个个体都存在两个潜在结果：接受干预时的潜在结果 Y(1)和在未接受时的潜在结果 Y(0)。这里的干预指的是模型中的变量的取值，如前述例子中 x 取 2 或者 3。有了潜在结果的相关理论，我们就可以从统计数据中分析个体因果效应和平均因果效应：个体因果效应： ICE=Y(1)−Y(0)；平均因果效应：ACE=E[Y(1)−Y(0)]。通过计算 ICE 和 ACE 的数值，可以量化 1 和 0 两种干预对模型结果的影响。这里我们强调了从统计数据中计算，而不是更直截了当地去进行随机对照实验，这是因为在一些场景中对照实验存在限制，无法进行实验，并且统计数据也更易得，在因果推断中常用统计数据替代随机对照实验。

在潜在结果的介绍中，我们强调了用观测数据替代随机对照实验的做法，在因果推断中，这一做法被称为观测性研究，观测性研究相比对照实验会更不精确，但是一种符合社会伦理道德且节约成本的近似做法。所谓观测性研究，其定义为：从大量的观测数据中选取两组样本，控制两组样本只有干预手段不同而其它变量全部一致，利用这两组数据计算因果效应估计值，分析因果效应。这些其它变量统称为协变量。

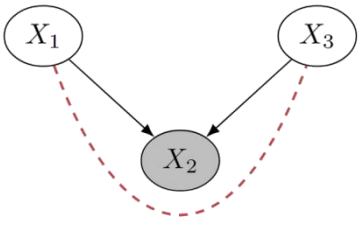
因果图的基础构建块，两个节点之间可以独立存在或有方向箭头连接。对于三个节点，展示了三种基本结构：链（Chain）、叉（Fork）和对撞机。下边简单介绍这三种结构中的阻断效应，首先是链结构中的阻断效应，如图节点X_2 阻断了节点X_1 和X_3 之间的关联。目标是证明 $P(X_1, X_3|X_2) = P(X_1|X_2)P(X_3|X_2)$，通过局部马尔可夫假设和贝叶斯定理进行推导。



叉结构中的阻断效应，如图节点X_2 阻断了节点X_1 和X_3 之间的关联。目标是证明 $P(X_1, X_3|X_2) = P(X_1|X_2)P(X_3|X_2)$，通过局部马尔可夫假设和贝叶斯定理可以进行推导。



叉结构中的阻断效应，如图节点X_2 连接了节点X_1 和X_3 之间的关联。目标是证明 $P(X_1, X_3) = P(X_1)P(X_3)$。



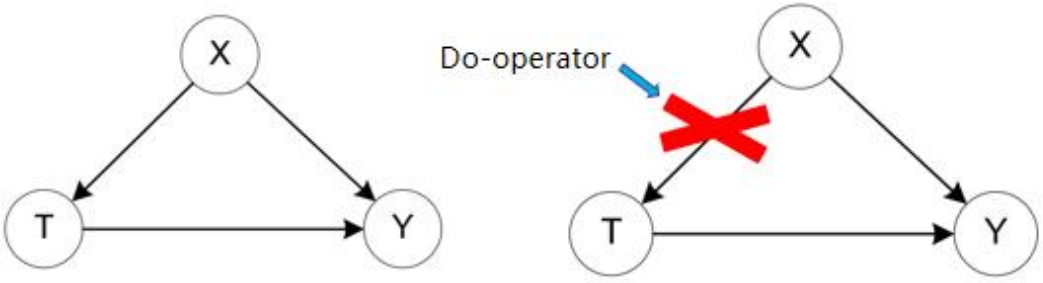
$$
\begin{aligned}
P(x_1, x_3) &= \sum_{x_2} P(x_1, x_2, x_3) \\
&= \sum_{x_2} P(x_1)\, P(x_3)\, P(x_2 \mid x_1, x_3) \\
&= P(x_1)\, P(x_3) \sum_{x_2} P(x_2 \mid x_1, x_3) \\
&= P(x_1)\, P(x_3)
\end{aligned}
$$

在因果推断中，do-算子是一个重要的工具，用于表示干预分布。具体来说，当我们进行干预 T=t 时，变量 Y 取值 y 的概率可以表示为 $P(Y(t) = y) = P(Y = y|do(T = t))$。这个定

义帮助我们理解在不同干预条件下的因果关系。为了衡量干预 T 在不同取值下对 Y 的影响，我们使用平均处理效应（Average Treatment Effect, ATE）。ATE 的计算公式为 $E[Y|do(T=1)]-E[Y|do(T=0)]$，它表示在干预 T 从 0 变为 1 时，Y 的期望值变化。

此外，我们需要区分观测分布和干预分布。在观测数据中，我们可以计算联合分布 $P(Y,T,X)$ 以及条件分布 $P(Y|T=t)$。而在干预情况下，分布则是 $P(Y|do(T=t))$ 和 $P(Y|do(T=t)，X=x)$。截断分解（Truncated Factorization）是一种用于处理干预分布的技术。假设 x 与干预 $do(S=s)$ 一致，那么截断分解可以表示为：$P(x_1,\cdots x_n|do(S=s))=\prod_{i\notin S}P(x_i|pa_i)$ 如果 x 与干预不一致，则该概率为 0。

为了更加精准地刻画变量之间的因果关系，Judea Pearl 等学者提出了因果图模型。因果图模型是一个用于表示变量之间依赖关系的有向无环图，可以通过概率分布来描述变量之间的条件独立性。一个简单的因果图如图所示，图中节点表示变量，有向边表示变量之间的因果关系，观察这个简单的因果图可以发现：T 和 X 之间有因果关系，X 同时在 T 和 Y 之间都有因果关系，此时如果我们想分析 T 跟 Y 之间的因果效应，直接利用 X，T，Y 三者的数据显然是行不通的，因为 X 同时影响了 T 和 Y，我们无法判断因果效应估计值的变化是因为 T 还是 T 和 X 的不同取值导致的，在这个模型中，X 就是协变量。



此时我们需要一个叫做 do 算子的工具，这个工具切断了 T 的所有父亲节点与其的联系，确保 T 不受其它变量的影响，这一步操作也叫因果干预。Do 算子的机制如图所示，它保证了 T 不受任何其它因素的影响。对 T 施加干预后，我们可以计算平均因果效应，其计算公式如下式所示，与之前的公式相比，Y(1)和 Y(0)分别被替换为了 Y|do(T=1)，Y|do(T=0)，用于表明干预机制。

使用 Do 算子可以切断模型中的某个节点的父亲节点(如混淆变量)对其的影响，将其固定为某个值，此时，可以重新计算其他变量的条件概率分布或期望值。首先，变量间的联合分布为 $P(x,y,t)$。这是由执行 do 算子之前的有向图决定的。对 T 施加干预后，将 X 的值固定，上式可以进行改写，此时，T 与 X 不再相关。然后可以推出在对 T 进行干预时，Y 的概率分布。通过上述计算，施加因果干预后的平均因果效应可以被计算出来，也即，我们定

量分析出了不受协变量影响的，T 与 Y 之间的因果性。

后门调整是因果推断中的一种方法，用于控制或消除因果路径上的潜在混淆因素，以推断变量之间的因果效应。在下面的因果图中，X 是 Y 的成因，而 C 是变量间的混淆因素 (或协变量)。前面我们已经明确了协变量会对真正的因果效应计算产生干扰，因此必须利用 do 算子将其排除。满足后门调整的条件后，在具体的因果分析中，就可以利用后门调整公式消除协变量的影响，计算因果效应。后门公式如下式所示，可以发现，后门调整的公式与前面我们基于 do 算子所分析的变量的条件概率分布计算方式是一致的。这是因为，后门调整的最重要操作就是基于 do 算子的 do 运算。

辛普森悖论（Simpson's Paradox）是一种统计现象，指在分组数据中存在的趋势在合并后可能会消失或反转。例如，某疾病的两类药品 A 和 B 在老人和年轻人两组中的康复率均显示 A 药更有效，但合并分析所有数据后却显示 B 药的总体康复率更高。这表明在数据分析时需谨慎处理分组和合并，以避免得出误导性结论。并且我们要明确关联性不等于因果性，通过 1999-2009 年间泳池溺水人数与尼古拉斯凯奇每年上映影片数量的变化趋势一致，说明虽然两者存在关联，但并不意味着尼古拉斯凯奇的影片数量对泳池溺水人数负责，从而强调了关联性不等于因果性。

经过前面对因果推断中相关理论的学习，我们再次讨论经典的辛普森悖论问题。首先，将该问题抽象成一个因果图模型，我们的目标是分析药品对疾病的疗效，那么其它所有条件或变量都被视为协变量。根据表中的数据，这里的协变量只有一个，就是患者的年龄。显然，年龄既会影响药物选择也会影响康复率。因此，在最初的分析计算中，之所以会得出两个矛盾的结论，是因为忽略了年龄这个协变量对结果的影响。在学习了因果推断的分析方法和计算工具后，我们现在要做的就是消除协变量的影响，计算真正的平均因果效应。可以简单套用一下后门调整的各项条件，模型中存在后门路径 X←C→Y，协变量是患者年龄，条件符合，那么直接利用后门调整公式计算即可。

根据下表所示，利用后门调整公式再次计算使用 A、B 药后的康复概率：

$P(Y=H|do(X=A))=P(Y=H|do(X=A,C=old))P(C=old)+P(Y=H|do(X=A,C=young))P(C=young)=11/30$

$P(Y=H|do(X=B))=P(Y=H|do(X=B,C=old))P(C=old)+P(Y=H|do(X=B,C=young))P(C=young)=8/30$

$ACE=E[Y=H|do(X=A)]-E[Y=H|do(X=B)]=11/30-8/30=3/30$

由于 X 有两个取值可能，所以我们分别对 X 进行干预，将其固定为 A 和 B，将后门调

整公式按不同的 C 的取值展开，最后得到 A 药的真正疗效就是 11/30，B 的真正疗效是 8/30。A、B 药品的平均因果效应即为 3/30。根据基于后门调整公式的计算，A、B 两种药物与某疾病之间的因果效应被量化出来。因此 A 药疗效更强，这也符合最初对特定人群单独分析时得出的结论。

服用 A、B 两类药的不同人群的康复率

	老人	年轻人	总计
A 药	30% 200/1000	60% 30/50	31.4% 330/1050
B 药	20% 4/50	50% 125/250	43% 129/300

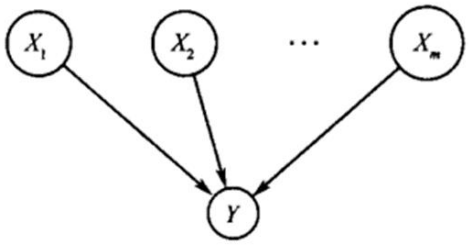
贝叶斯网的参数，即各变量的概率分布，一般是通过数据分析获得的，有时也可以从问题的特性直接得到。列举一个从问题特性得到参数的例子，假设用一个噪音信道传递信息位 U，信道的出错概率为 0.1。为提高传输准确度，可以简单地把 U 重复传输 3 次，然后在接收方根据接收结果进行解码，推测正确的 U 的取值。这个过程可以用如下图中所示的贝叶斯网来表示。网络中所有变量均取二值，其中传输位 X1,X2,X3 是 U 的简单重复，接收位 Y1,Y2,Y3 是接收方接收到的 3 个信息位。可以根据问题的特性来确定网络参数:对于信息位 U 和传输位 X1,X2,X3，有 P(U)=(0.5,0.5)，$P(X_i|U)=\begin{cases}1, \text{如果 U 和} X_i \text{状态相同}\\0, \text{如果 U 和} X_i \text{状态相同}\end{cases}$，而接收位 Yi 的条件概率分布由噪音信道的特性决定：$P(X_i|U)=\begin{cases}0.9, \text{如果} Y_i \text{和} X_i \text{状态相同}\\0.1, \text{如果} Y_i \text{和} X_i \text{状态相同}\end{cases}$。种马农场是通过数据分析来获取参数的。一个种马农场中的公马、母马和它们所生育的后代之间的基因遗传关系。假设基因 a 是一个隐性致病基因，对应显性基因是 A。当没有任何信息时，任意一匹马关于该疾病的基因型可能是以下三者之一:aa(患病)，aA(携带者)或 AA(正常)，根据基因遗传学，可以直接确定任意一匹马的基因型 GC 与它的父母的基因型 GF 和 GM 之间的概率关系 P(GC｜GF,GM)。如表所示，考察表中第一行，当公马的基因是 aa 时，如果母马也是 aa，根据基因遗传交换规律，后代只可能是 aa，所以其分布是(1,0,0);如果母马是 aA，则后代不可能是 AA，而有 50%的可能是 aa，50%的可能是 aA，因此其分布是(0.5,0.5,0);如果母马是 AA,则后代一定是 aA，因此其分布是(0,1,0)。其余概率分布之意义可依次类推。

种马基因遗传的概率分布 $P(G_C|G_F,G_M)$

G_F \ G_M	aa	aA	AA
aa	(1，0，0)	(0.5，0.5，0)	(0，1，0)
aA	(0.5，0.5，0)	(0.25，0.5，0.25)	(0，0.5，0.5)
AA	(0，1，0)	(0，0.5，0.5)	(0，0，1)

注：数字（α，β，γ）分别表示后代基因型是（aa，aA，AA）的概率.

有时候，网络参数是通过专家咨询得到，这在贝叶斯网发展的早期尤为常见，专家咨询耗费人力和时间，因此有必要尽量减少参数个数，通常这也有利于提高咨询结果的质量，下面介绍两个减少参数的方法。设变量 Y 有 m 个父节点 $X_1,\cdots,X_m$，条件分布 $P(Y \mid X_1,\cdots,X_m)$ 刻画 Y 对其父节点的依赖关系，如下图所示。当所有变量均取二值时，它包含 2^m 个独立参数。为减少参数个数，人们往往假设条件分布具有某种规律，称为局部结构。常见的局部结构有两种：因果机制独立和环境独立。



1）因果机制独立：指的是多个原因独立地影响同一个结果。在 Alarm 问题中，地震(E)和盗窃(B)都可以触发警铃(A)，但是两者的机制不同，因此可以假设地震和盗窃独立地影响警铃响。

2）环境独立：环境独立是指在特定环境下才成立的条件独立关系，一个环境是一组变量及其取值的组合。例一个人的收入一般来说取决于其职业、受教育程度以及气候。然而对于程序员，其收入不依赖于气候；而对于农民，则收入不依赖于受教育程度。这种环境独立关系使得在建模过程中参数数量大幅减少，从而简化了模型的复杂度。