

作者	生姜 DrGinger
脚本	生姜 DrGinger
视频	崔崔 CuiCui
开源学习资源	https://github.com/Visualize-ML
平台	https://www.youtube.com/@DrGinger_Jiang https://space.bilibili.com/3546865719052873 https://space.bilibili.com/513194466

2.2 二元离散随机变量



本节你将掌握的核心技能：

- ▶ 用联合概率质量函数描述二元离散随机变量的概率分布。
- ▶ 联合 PMF 所有概率和为 1。
- ▶ 通过“偏求和”方法从联合 PMF 得到边缘概率。
- ▶ 给定事件条件下计算另一随机变量的条件分布。
- ▶ 条件概率、边缘概率、联合概率之间的相互转化关系。
- ▶ 如果离散随机变量独立，边缘 PMF 乘积为联合 PMF。

在一次随机试验中，我们常常不仅关心一个随机变量的结果，还可能同时观测到两个相关的离散随机变量 X 和 Y 。这时，我们就需要一个能同时描述这两个变量取值可能性的函数：**联合概率质量函数** (joint Probability Mass Function, joint PMF)，记作 $p_{X,Y}(x, y)$ 。

联合概率质量函数

假设同一个试验中，有两个离散随机变量 X 和 Y 。二元随机变量 (X, Y) 概率取值可以用**联合概率质量函数** $p_{X,Y}(x, y)$ 刻画。

概率质量函数 $p_{X,Y}(x, y)$ 代表事件“随机变量 X 取值为 x ，同时随机变量 Y 取值为 y ”的概率，即 $\{X = x, Y = y\}$ 发生的联合概率：

$$\underbrace{p_{X,Y}(x, y)}_{\text{Joint}} = \Pr(X = x, Y = y) \quad (1)$$

⚠ 对于二元离散随机变量， $p_{X,Y}(x, y)$ 本身就是概率值。

举个例子，图 1 所示为某个二元联合概率质量函数 $p_{X,Y}(x, y)$ 的**热图** (heatmap)。

在图中，每个小方格对应一个可能的取值组合 (x, y) ，颜色的亮暗代表该组合发生的概率大小；颜色越亮，说明该事件的概率越高。这种可视化方式能够帮助我们直观地理解两个离散随机变量之间的依赖

关系：若浅色主要集中在左下到右上这条对角线附近，说明 X 与 Y 之间存在线性正相关（这便是图 1 的情况）；反之，若集中在左上到右下这条对角线附近附近，则代表线性负相关。



本书后续讲专门讲解线性相关性。

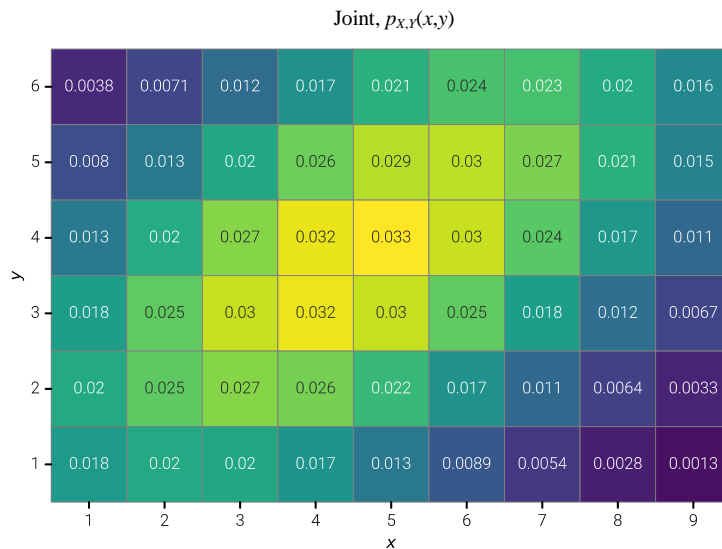


图 1. 概率质量函数 $p_{X,Y}(x, y)$ 热图

二元联合概率质量函数 $p_{X,Y}(x, y)$ 必须满足归一化性质，即所有可能取值组合的概率总和为 1：

$$\sum_x \sum_y \underbrace{p_{X,Y}(x, y)}_{\text{Joint}} = \sum_y \sum_x \underbrace{p_{X,Y}(x, y)}_{\text{Joint}} = 1, \quad 0 \leq p_{X,Y}(x, y) \leq 1 \quad (2)$$

这意味着图 1 这幅热图中所有方格的概率值加起来正好为 1，无论我们先对 x 求和还是先对 y 求和，结果都不会改变。这个性质确保了联合分布完整地描述了整个样本空间中所有可能事件的概率分布。

边缘概率：偏求和，相当于降维

在理解联合概率质量函数之后，我们常常只关注其中一个随机变量的概率分布，而暂时忽略另一个变量的影响。此时，就需要引入一个非常重要的概念：**边缘概率** (marginal probability)。

顾名思义，边缘概率描述的是一个随机变量自身的概率分布，而不考虑其他变量的取值。换句话说，它是在联合分布中“降维”或“整合”掉其他变量后所得到的结果。

假设我们有两个离散随机变量 X 和 Y ，它们的联合概率质量函数为 $p_{X,Y}(x, y)$ 。如果我们只关心随机变量 $X = x$ 的概率分布 $p_X(x)$ ，那么我们可以在 $X = x$ 条件下通过对所有可能的 Y 取值进行求和，得到：

$$\underbrace{p_X(x)}_{\text{Marginal}} = \sum_y \underbrace{p_{X,Y}(x, y)}_{\text{Joint}} \quad (3)$$

这个过程称为**边缘化** (marginalization)。

直观上可以理解为，我们在 $X = x$ 的情况下，把所有与不同 Y 值的联合概率全部加起来，相当于把二维分布在 Y 轴方向上“折叠”成一维分布。这样得到的 $p_X(x)$ 就是 X 的边缘概率质量函数——它只依赖于 x ，与 Y 无关。

从函数角度看，联合概率质量函数 $p_{X,Y}(x, y)$ 是一个二元函数，而边缘概率 $p_X(x)$ 是一个一元函数。也就是说，边缘化的过程相当于对函数“降维”：从二维空间中的一个分布，压缩为一维空间中的一个分布。

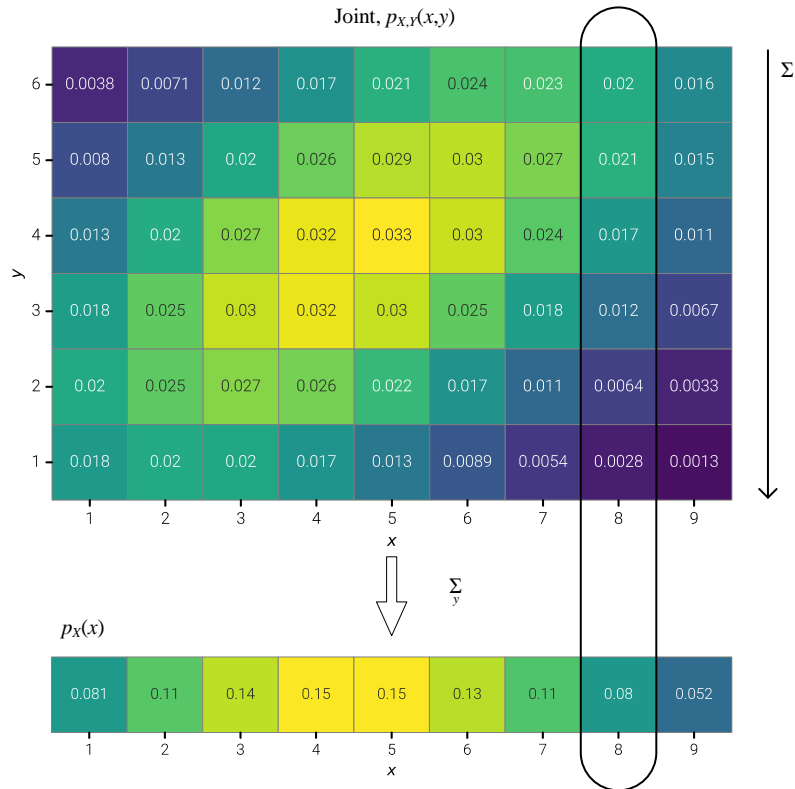


图 2. 利用联合概率 $p_{X,Y}(x, y)$ 计算边缘概率 $p_X(x)$

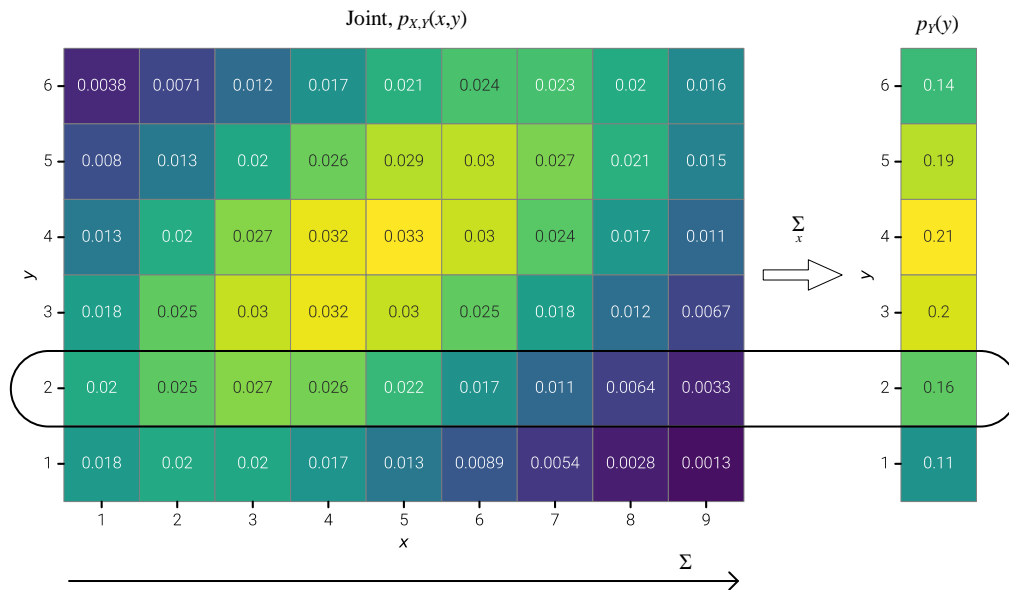
从矩阵运算角度来看， $p_{X,Y}(x, y)$ 代表矩阵 (matrix)，矩阵沿 Y 方向求和，折叠得到行向量 $p_X(x)$ 。行向量 $p_X(x)$ 进一步求和结果为标量 1 (归一化)，对应样本空间概率。反方向来看，概率 1 沿 X 和 Y 展开，相当于“切片、切丝”。

如图 2 所示，在联合概率热图中，固定某一个 X 值，比如 $X = 8$ ，然后把该列中所有 $p_{X,Y}(8, y)$ 的概率值相加，得到 $p_X(8) = 0.08$ 。类似地，对其他 X 的取值重复这个过程，就能得到完整的一维分布 $p_X(x)$ 。

? 请大家自己验算当 X 取其他值时，边缘概率 $p_X(x)$ 的具体值。

同理， $p_{X,Y}(x, y)$ 对 x “偏求和”消去 x 得到 $p_Y(y)$ ：

$$\underbrace{p_Y(y)}_{\text{Marginal}} = \sum_x \underbrace{p_{X,Y}(x, y)}_{\text{Joint}} \quad (4)$$

图 3. 利用联合概率 $p_{X,Y}(x, y)$ 计算边缘概率 $p_Y(y)$

举个例子，如图 3 所示，当 $Y=2$ 时，将整个一行的 $p_{X,Y}(x, 2)$ 相加得到 $p_Y(2) = 0.16$ 。

从函数角度来看， $p_Y(y)$ 也是个一元离散函数。

从矩阵运算角度来看，矩阵 $p_{X,Y}(x, y)$ 沿 X 方向求和，折叠得到列向量 $p_Y(y)$ 。这相当于从二维降维到一维。

列向量 $p_Y(y)$ 进一步折叠结果同样为标量 1。

因此，边缘概率的核心思想就是“通过求和折叠不关心的维度”，实现从联合分布到单变量分布的降维。这不仅是理解多维概率分布的关键步骤，也为后续计算条件概率、贝叶斯推断等概念奠定了基础。

条件概率：引入贝叶斯定理

在理解了联合概率和边缘概率之后，我们就可以进一步探讨**条件概率** (conditional probability)。本书前文提过，条件概率用于描述在一个事件已知发生的前提下，另一个事件发生的可能性。

换句话说，条件概率回答了这样的问题：“如果我已经知道某个事件发生了，那么另一个事件发生的概率会是多少？”

假设我们有两个离散随机变量 X 和 Y 。给定事件 $\{X=x\}$ 条件下，当 $p_X(x) > 0$ ，事件 $\{Y=y\}$ 发生的概率可以用**条件概率质量函数** (conditional probability mass function, conditional PMF) $p_{Y|X}(y|x)$ 表达：

$$\underbrace{p_{Y|X}(y|x)}_{\text{Conditional}} = \frac{\overbrace{p_{X,Y}(x,y)}^{\text{Joint}}}{\underbrace{p_X(x)}_{\text{Marginal}}} \quad (5)$$

其中分母 $p_X(x)$ 表示事件 $\{X=x\}$ 的概率。

举个例子，当 $X=8$ 时 (条件)，计算 $Y=6$ 的条件概率，

$$p_{Y|X}(6|8) = \frac{p_{X,Y}(6,8)}{p_X(8)} = \frac{0.02}{0.08} = 0.25 \quad (6)$$

如图 4 所示，当 $X = 8$ 时，改变 Y 的取值我们可以计算得到一组条件概率值 $p_{Y|X}(y | 8)$ 。

? 当 Y 取不同值时，计算 $p_{Y|X}(y | 8)$ 之和，并试着解释求和结果。

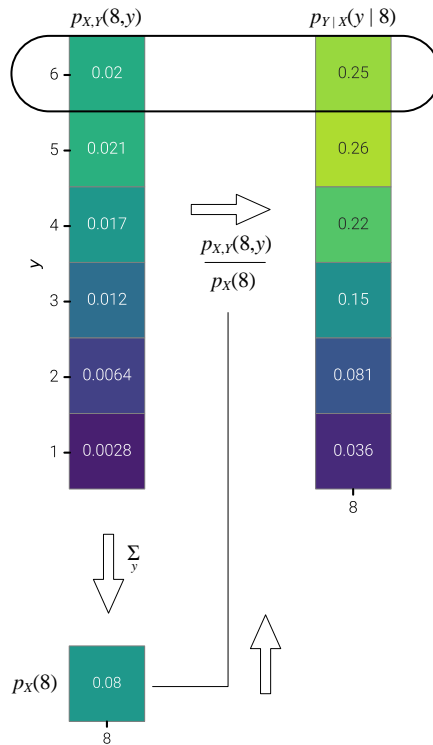


图 4. 计算条件概率 $p_{Y|X}(y | 8)$

当 X 取不同值时 (即不同条件时)，不断重复图 4 类似计算，我们可以得到图 5 这张热图。

如图 5 所示，从函数角度来看， $p_{Y|X}(y | x)$ 也是个二元函数。

⚠ 注意，对于条件概率 $p_{Y|X}(y | x)$ ， $\{X = x\}$ 是新的样本空间。

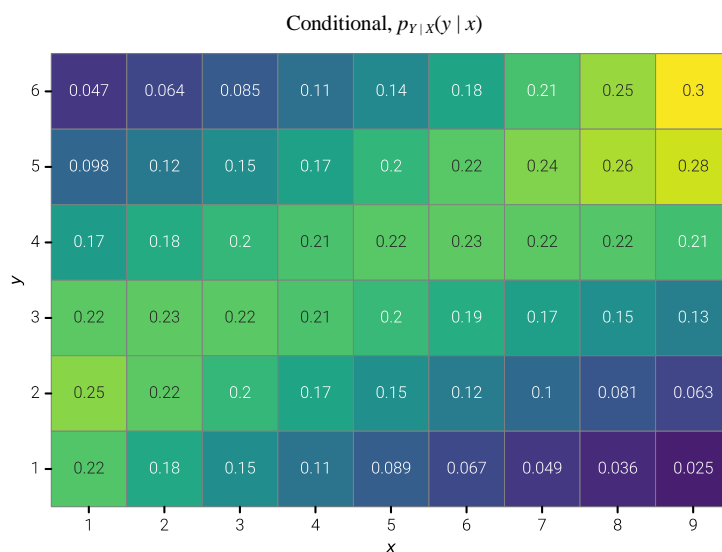
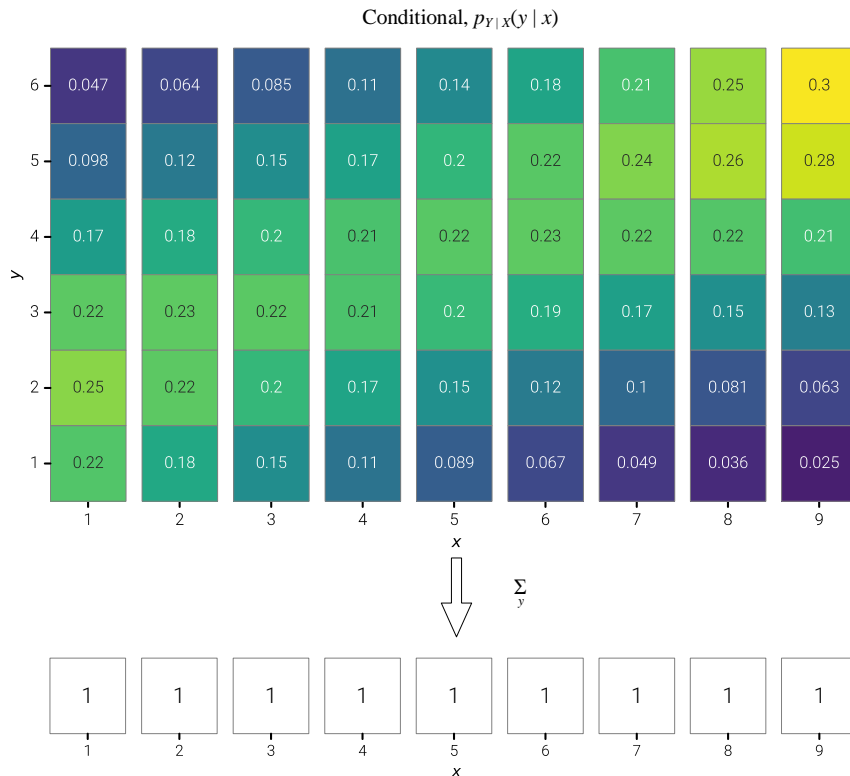


图 5. 条件概率 $p_{Y|X}(y|x)$ 热图

如图 6 所示，我们可以将 $p_{Y|X}(y|x)$ 理解为在固定 $X=x$ 的情况下，观察 Y 的不同取值时对应的概率分布。此时，条件 $\{X=x\}$ 就定义了一个新的样本空间——所有的概率都限定在这一条件之下，并且在这个新的样本空间中，所有 Y 的概率之和等于 1：

$$\sum_y p_{Y|X}(y|x) = 1 \quad (7)$$

图 6. 条件概率 $p_{Y|X}(y|x)$ 热图，每一列独立求和

(5) 也可以用来反求联合概率 $p_{X,Y}(x,y)$ ：

$$\underbrace{p_{X,Y}(x,y)}_{\text{Joint}} = \underbrace{p_{Y|X}(y|x)}_{\text{Conditional}} \cdot \underbrace{p_X(x)}_{\text{Marginal}} \quad (8)$$

这一定义揭示了联合分布可以看作“条件分布 \times 边缘分布”，即“整体概率 = 条件下的局部概率 \times 条件的权重”。

边缘概率 $p_Y(y)$ 也是条件概率 $p_{Y|X}(y|x)$ 的加权平均：

$$\underbrace{p_Y(y)}_{\text{Marginal}} = \sum_x p_{X,Y}(x,y) = \sum_x \underbrace{p_{Y|X}(y|x)}_{\text{Conditional}} \underbrace{p_X(x)}_{\text{Marginal}} \quad (9)$$

这其实是全概率公式的离散形式，表达了“边缘概率 = 各条件下的概率 \times 各条件发生的权重”。

假设事件 $\{Y=y\}$ 已经发生，即 $p_Y(y) > 0$ 。在给定事件 $\{Y=y\}$ 条件下，事件 $\{X=x\}$ 发生的概率可以用条件概率质量函数 $p_{X|Y}(x|y)$ 表达。

⚠注意，对于 $p_{X|Y}(x|y)$ ， $\{Y=y\}$ 是新的样本空间。

利用贝叶斯定理，条件概率 $p_{X|Y}(x|y)$ 可以用联合概率 $p_{X,Y}(x,y)$ 除以边缘概率 $p_Y(y)$ 得到：

$$\underbrace{p_{X|Y}(x|y)}_{\text{Conditional}} = \frac{\overbrace{p_{X,Y}(x,y)}^{\text{Joint}}}{\underbrace{p_Y(y)}_{\text{Marginal}}} \quad (10)$$

如图 7 所示，从函数角度来看， $p_{X|Y}(x|y)$ 本质上也是个二元函数。

当我们固定 $Y=y$ 时，就在联合分布的矩阵中“取出一行”，并对该行进行归一化，使其和为 1。换句话说，条件概率的计算本质上是一个归一化过程：通过除以边缘概率，把某一行或某一列的联合概率标准化为新的概率分布。

$p_{X|Y}(x|y)$ 显然随着 $X=x$ 变化。虽然 $Y=y$ 为条件，但是这个条件也可以变动。 $Y=y$ 变动就会导致概率质量函数 $p_{X|Y}(x|y)$ 变化。

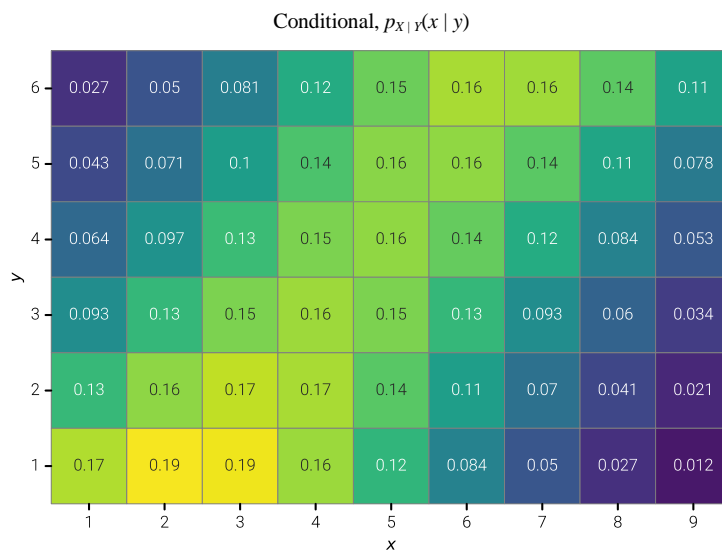


图 7. 条件概率 $p_{X|Y}(x|y)$ 热图

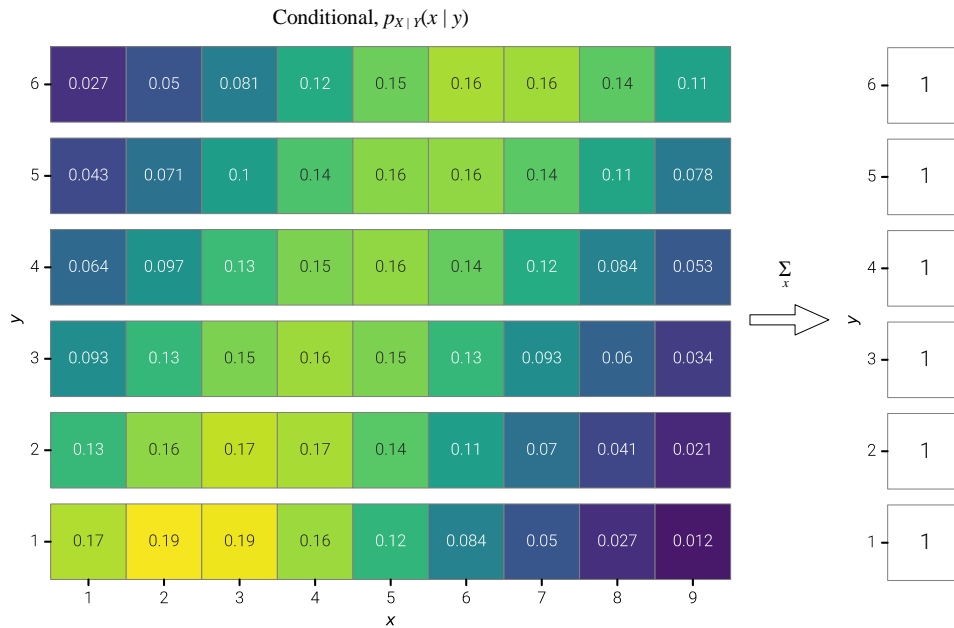
如图 8 所示， $p_{X|Y}(x|y)$ 对 x 求和等于 1：

$$\sum_x p_{X|Y}(x|y) = 1 \quad (11)$$

也就是说， $p_{X|Y}(x|y)$ 矩阵的每一行求和结果为 1。

也就是说，每一行代表一个不同的“样本空间”。

换个视角来看，条件概率的“条件”就是“新的样本空间”，这个新的样本空间对应概率为 1。

图 8. 条件概率 $p_{X|Y}(x|y)$ 热图，每一行独立求和

如图 3 所示， $Y=2$ 时，边缘概率 $p_Y(Y=2)$ 可以通过求和得到：

$$p_Y(2) = \sum_x p_{X,Y}(x, 2) \quad (12)$$

$p_Y(2)$ 为一定值。给定 $Y=2$ 作为条件时，条件概率 $p_{X|Y}(x|2)$ 通过下式得到：

$$\underbrace{p_{X|Y}(x|2)}_{\text{Conditional}} = \frac{\overbrace{p_{X,Y}(x, 2)}^{\text{Joint}}}{\underbrace{p_Y(2)}_{\text{Marginal}}} \quad (13)$$

进一步，条件概率 $p_{X|Y}(x|2)$ 对 x 求和得到 1：

$$\sum_x p_{X|Y}(x|2) = \frac{\sum_x p_{X,Y}(x, 2)}{p_Y(2)} = \frac{p_Y(2)}{p_Y(2)} = 1 \quad (14)$$

$p_{X,Y}(x, 2)$ 到 $p_{X|Y}(x|2)$ 是一个归一化过程。也就是说，上式分母中的 $p_Y(y)$ 是一个归一化系数。

引入贝叶斯定理，边缘概率 $p_X(x)$ 相当于是条件概率的加权平均：

$$\underbrace{p_X(x)}_{\text{Marginal}} = \sum_y \underbrace{p_{X,Y}(x, y)}_{\text{Joint}} = \sum_y \underbrace{p_{X|Y}(x|y)}_{\text{Conditional}} \underbrace{p_Y(y)}_{\text{Marginal}} \quad (15)$$

相反，条件概率 $p_{X|Y}(x|y)$ 到联合概率 $p_{X,Y}(x, y)$ 相当于，以边缘概率 $p_Y(y)$ 作为系数缩放 $p_{X|Y}(x|y)$ 的过程：

$$\underbrace{p_{X,Y}(x, y)}_{\text{Joint}} = \underbrace{p_{X|Y}(x|y)}_{\text{Conditional}} \underbrace{p_Y(y)}_{\text{Marginal}} \quad (16)$$

独立：条件概率等于边缘概率

在研究联合概率和条件概率之后，我们就可以进一步讨论一个更为特殊但非常重要的概念：随机变量的**独立性** (independence)。独立性是描述两个随机变量之间是否存在统计关联的核心思想。如果两个变量相互独立，那么一个变量的取值不会对另一个变量的分布产生任何影响，反之亦然。

如果离散随机变量 X 和 Y 独立，联合概率 $p_{X,Y}(x,y)$ 等于 $p_Y(y)$ 和 $p_X(x)$ 两个边缘概率质量函数 PMF 乘积：

$$\underbrace{p_{X,Y}(x,y)}_{\text{Joint}} = \underbrace{p_Y(y)}_{\text{Marginal}} \cdot \underbrace{p_X(x)}_{\text{Marginal}} \quad (17)$$

在图 9 中，我们可以看到，若 X 和 Y 独立，则它们的联合分布的形状正好是两个一维边缘分布相乘后的结果。

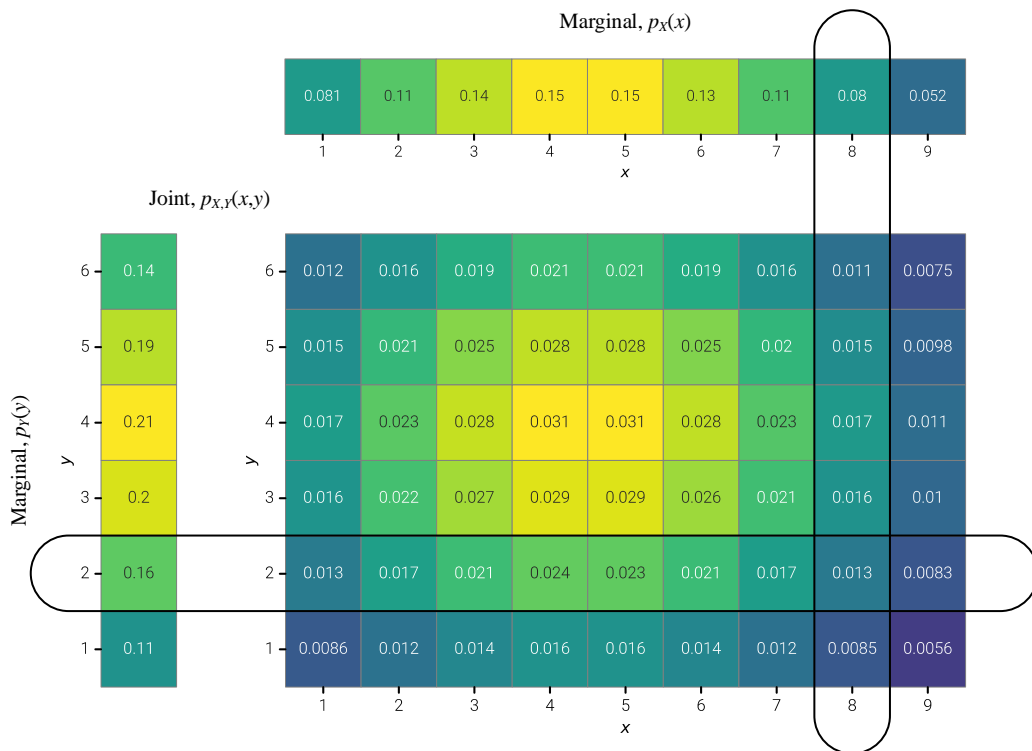


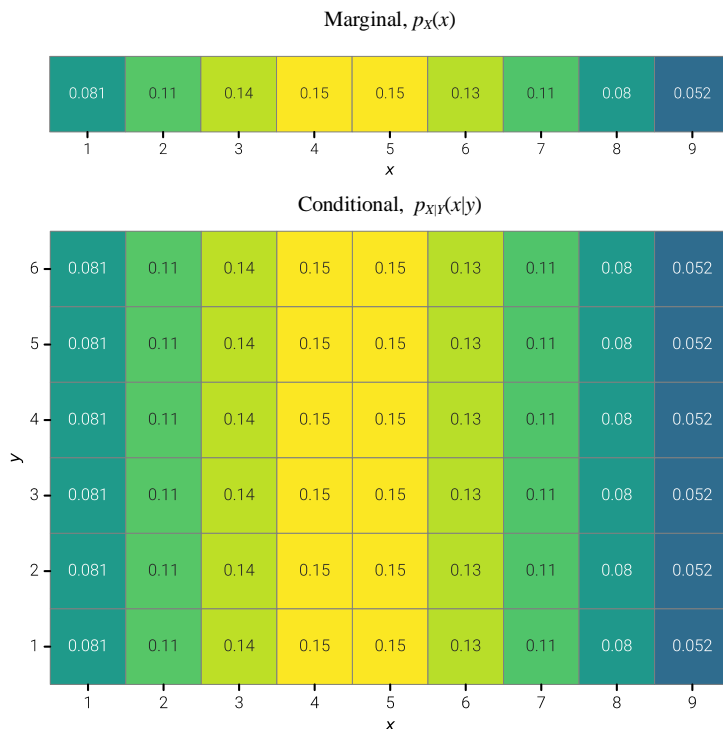
图 9. 联合概率 $p_{X,Y}(x,y)$ 等于 $p_Y(y)$ 和 $p_X(x)$ 两个边缘概率乘积， X 、 Y 独立

独立性还可以从条件概率的角度来理解。如果两个离散变量 X 和 Y 独立，条件概率 $p_{X|Y}(x|y)$ 等于边缘概率 $p_X(x)$ ，下式成立：

$$\underbrace{p_{X|Y}(x|y)}_{\text{Conditional}} = \underbrace{p_X(x)}_{\text{Marginal}} \quad (18)$$

这意味着无论 Y 取什么值， X 的概率分布都不会变化。换句话说，知道 Y 的信息对 X 的预测没有任何帮助。

如图 10 所示，离散变量 X 和 Y 独立，不管 y 取任何值 ($0 \sim 8$)， $p_X(x)$ 的取值和 $p_{X|Y}(x|y)$ 完全相同。

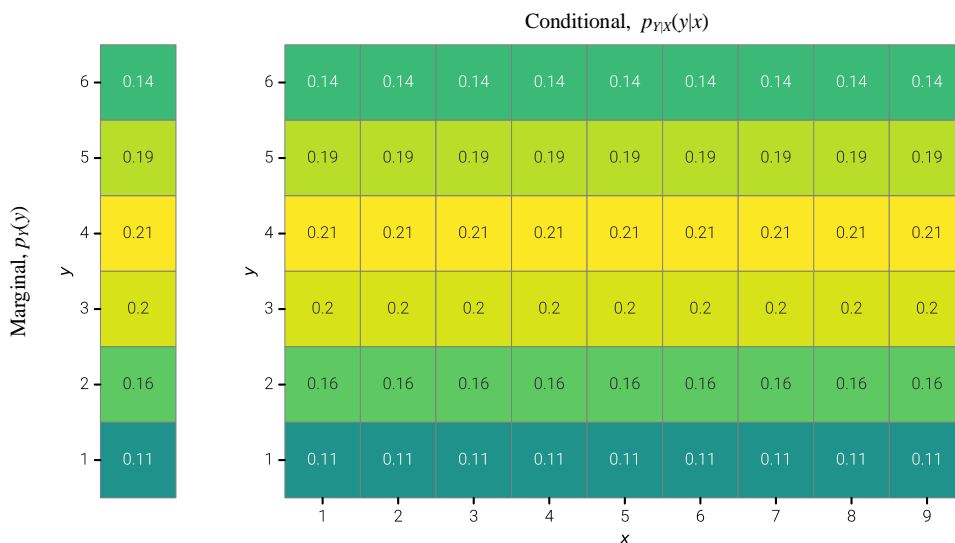
图 10. X 和 Y 独立，条件概率 $p_{X|Y}(x|y)$ 等于边缘概率 $p_X(x)$

(18) 等价于下式：

$$\underbrace{p_{Y|X}(y|x)}_{\text{Conditional}} = \underbrace{p_Y(y)}_{\text{Marginal}} \quad (19)$$

同理，如图 11 所示， X 和 Y 独立时， $p_Y(y)$ 的取值和 $p_{Y|X}(y|x)$ 相同。

这恰恰说明， X 的取值和 Y 无关，也就是为什么条件概率 $p_{Y|X}(y|x)$ 的形状不受 $X = x$ 影响，都和 $p_Y(y)$ 相同。

图 11. X 和 Y 独立，条件概率 $p_{Y|X}(y|x)$ 等于边缘概率 $p_Y(y)$

我们可以从另一个角度理解这一点：当 X 与 Y 独立时，它们的联合分布实际上是“可分离”的。二维概率表可以分解为两个一维概率向量的外积，每一个维度的变化不会对另一个维度造成影响。

不独立

在理解了独立性的定义之后，我们自然要讨论它的反面：不独立。当两个随机变量 X 和 Y 不独立时，意味着一个变量的取值会影响另一个变量的分布。换句话说，已知一个变量的信息，会改变我们对另一个变量发生概率的判断。

本节前文已经介绍，如果 X 和 Y 不独立，如果 $p_Y(y) > 0$ ，条件概率 $p_{X|Y}(x|y)$ 公式如下：

$$\underbrace{p_{X|Y}(x|y)}_{\text{Conditional}} = \frac{\overbrace{p_{X,Y}(x,y)}^{\text{Joint}}}{\underbrace{p_Y(y)}_{\text{Marginal}}} = \frac{\overbrace{p_{X,Y}(x,y)}^{\text{Joint}}}{\sum_x \overbrace{p_{X,Y}(x,y)}^{\text{Joint}}} \quad (20)$$

如图 12 所示，当 X 和 Y 不独立，条件概率 $p_{X|Y}(x|y)$ 取值不同于边缘概率 $p_X(x)$ 。也就是说，知道 Y 的取值之后，我们对 X 的分布的认识会发生改变。

例如，如果 Y 代表一个学生的学习时间，而 X 表示考试成绩，那么当我们知道学生学习时间较长时，考试成绩分布很可能会整体向高分段偏移；这正是变量之间不独立的体现。

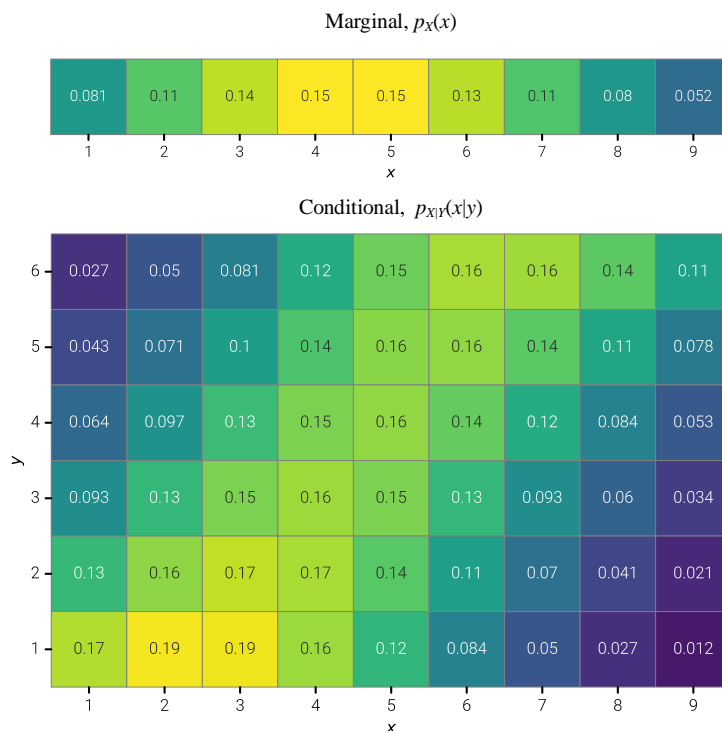


图 12. X 和 Y 不独立，条件概率 $p_{X|Y}(x|y)$ 不同于边缘概率 $p_X(x)$

如果 $p_X(x) > 0$ ，条件概率 $p_{Y|X}(y|x)$ 需要利用贝叶斯定理计算：

$$\underbrace{p_{Y|X}(y|x)}_{\text{Conditional}} = \frac{\overbrace{p_{X,Y}(x,y)}^{\text{Joint}}}{\underbrace{p_X(x)}_{\text{Marginal}}} = \frac{\overbrace{p_{X,Y}(x,y)}^{\text{Joint}}}{\sum_y p_{X,Y}(x,y)} \quad (21)$$

如图 13 所示, X 和 Y 不独立, 条件概率 $p_{Y|X}(y|x)$ 取值不同于边缘概率 $p_Y(y)$ 。

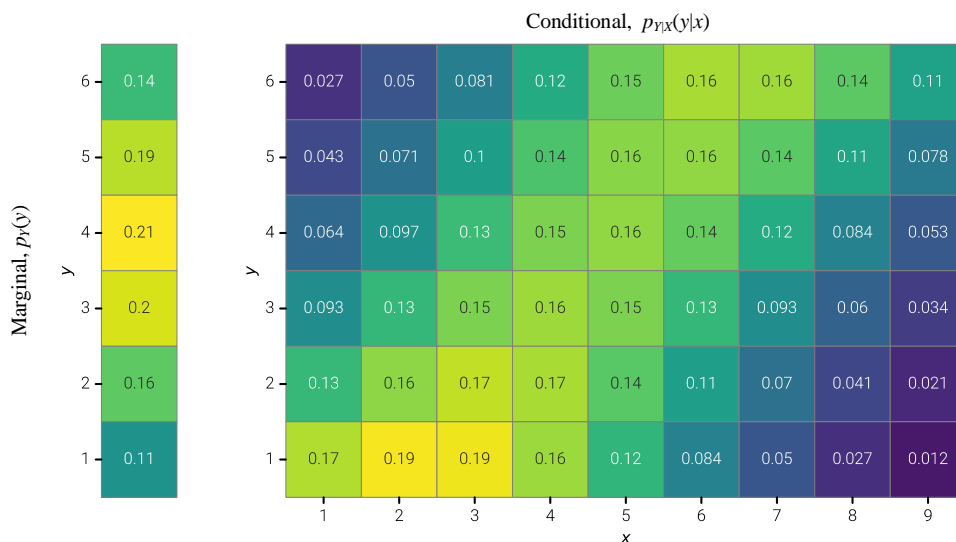
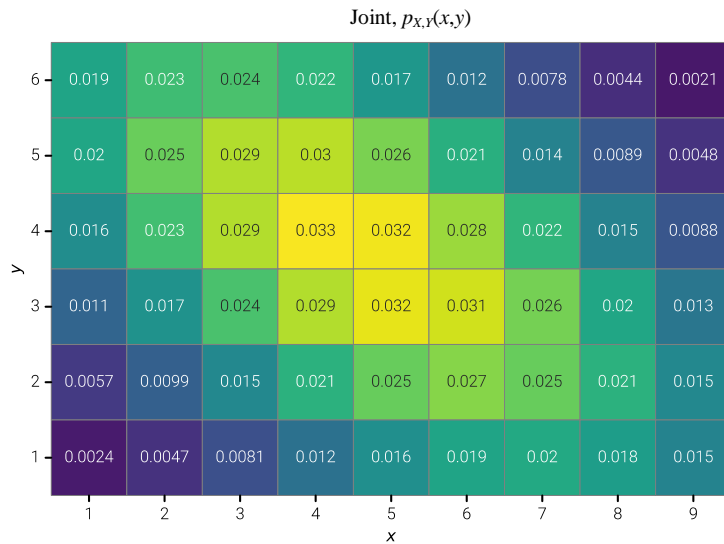


图 13. X 和 Y 不独立, 条件概率 $p_{Y|X}(y|x)$ 不同于边缘概率 $p_Y(y)$

当我们观察图 1 所示的 (X, Y) 的联合概率质量函数热图时, 如果发现当 X 取较大值时, Y 也更可能取较大值; 而当 X 取较小值时, Y 也更可能取较小值, 这种“同向变化”的趋势就是**线性正相关** (positive linear correlation) 的表现。

相反, 如图 14 所示, 当 X 取较大值时, Y 更可能取较小值; 而当 X 取较小值时, Y 更可能取较大值, 这种“反向变化”的趋势就是**线性负相关** (negative linear correlation) 的表现。

如图 9 所示, 而当热图分布均匀、没有明显的斜向聚集趋势时, 说明 X 和 Y 之间没有相关性, 即它们相互独立或几乎不相关。

图 14. X 和 Y 呈现线性负相关

本章后面将专门讲解协方差、线性相关性这两个重要的概率统计概念。



请大家用 DeepSeek/ChatGPT 等工具完成本节如下习题。

Q1. 给定如下联合概率 PMF，请首先验证所有联合概率值之和是否为 1。然后，分别计算：

		X		
		1	2	3
Y	1	0.2	0.1	0.1
	2	0.0	0.25	0.05
	3	0.05	0.1	0.15

- ▶ 边缘概率 $p_X(2)$ 、 $p_Y(2)$
- ▶ 边缘概率 $p_X(x)$ 、 $p_Y(y)$
- ▶ 条件概率 $p_{X|Y}(2|2)$ 、 $p_{Y|X}(2|2)$
- ▶ 条件概率 $p_{X|Y}(2|y)$ 、 $p_{Y|X}(2|x)$
- ▶ 条件概率 $p_{X|Y}(x|2)$ 、 $p_{Y|X}(y|2)$
- ▶ 条件概率 $p_{X|Y}(x|y)$ 、 $p_{Y|X}(y|x)$

Q2. 请编写 Python 代码完成 Q1. 运算，并用 Python 可视化。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com