4 Discrete Random Variables 离散随机变量

取值为有限个或可数无穷个，对应概率质量函数 PMF

我，是由无数原子组成的宇宙，又是整个宇宙的一粒原子。

I, a universe of atoms, an atom in the universe.

—— 理查德·费曼 (Richard P. Feynman) | 美国理论物理学家 | 1918 ~ 1988

- ◄ numpy.sort() 排序
- ◄ seaborn.heatmap() 产生热图
- ◄ seaborn.histplot() 绘制频率/概率/概率密度直方图
- ◄ seaborn.scatterplot() 绘制散点图

4.1 随机：天地不仁，以万物为刍狗

随机试验

《数学要素》第 20 章介绍过随机现象的准确定义——在一定条件下，出现的可能结果不止一个，事前无法确切知道哪一个结果一定会出现，但大量重复试验中其结果又具有统计规律的现象称为随机现象。

随机试验 (random experiment) 是在相同条件下对某随机现象进行的大量重复观测。随机试验需要满足三个条件：

a) 可重复，在相同条件下试验可以重复进行；

b) 结果集合明确，每次试验的可能结果不止一个，并且能事先明确试验的所有可能结果；

c) 单次试验结果不确定，进行一次试验之前不能确定哪一个结果会出现，但必然出现结果集合中的一个。

两种随机变量：离散、连续

随机变量 (random variable) 是一个函数，它将样本数值赋给试验结果。换句话说，它是试验样本空间到实数集合的函数。比如上一章为了方便表达“抛三枚色子试验”中三枚色子各自点数，我们定义了 X_1、X_2、X_3，它们都是随机变量。

随机变量分为两种——**离散** (discrete)、**连续** (continuous)。

如果随机变量的所有取值能够一一列举出来，可以是为有限个或可数无穷个，这种随机变量被称作**离散随机变量** (discrete random variable)。比如，投一枚硬币结果正面为 1、反面为 0。掷一枚色子得到的点数为 1、2、3、4、5、6 中的一个值。再比如，鸢尾花的标签有三种——setosa (C_1)、versicolour (C_2)、virginica (C_3)。上一章介绍的古典概率就是针对离散型随机变量。

与之相对的是，**连续随机变量** (continuous random variable) 可能取值对应全部实数，或者数轴上某一区间内，比如温度、人的身高体重就是连续随机变量。再比如，鸢尾花数据花萼长度、花萼宽度、花瓣长度、花瓣宽度也都可以视作连续随机变量。

字母

本书用大写斜体字母表达随机变量，比如 X、Y、Z、X_1、X_2、Y_1、Y_2 等。

用小写字母表达随机变量取值，比如 x、y、x_1、x_2、x_1、x_2、i、j、k 等。其中，x、y、x_1、x_2、x_1、x_2 等通用于离散、连续随机变量，而序号 i、j、k 一般用于离散随机变量。

简单来说，X、Y、Z、X_1、X_2、Y_1、Y_2 等替代描述随机试验结果的描述性文字。而 x、y、x_1、x_2、x_1、x_2 等相当于函数的输入变量，它们主要用在**概率密度函数** (probability density function, PDF)、**概率质量函数** (probability mass function, PMF) 中。

如图 1 所示，抛一枚色子试验中，令随机变量 X 为色子点数，$X = x$，x 代表取值，也就是 X 的取值为变量 x。举个例子，$\Pr(X = x)$ 为事件 $\{X = x\}$ 的概率，x 表示随机变量 X 的取值。

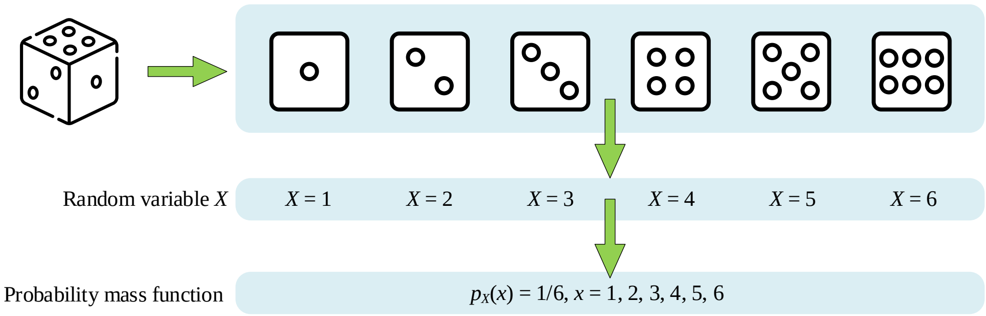


图 1. 随机试验、随机变量、概率质量函数三者关系

两种概率分布函数

研究随机变量取值的统计规律是概率论重要目的之一。概率分布函数则是对统计规律的简化和抽象。图 2 比较两种概率分布函数——概率质量函数 PMF、概率密度函数 PDF。

白话来说，概率质量函数 PMF、概率密度函数 PDF 就是两种对概率为 1“切丝切片”、“切条切丝”的不同方法。本章后续还会沿着这个思路继续讨论。

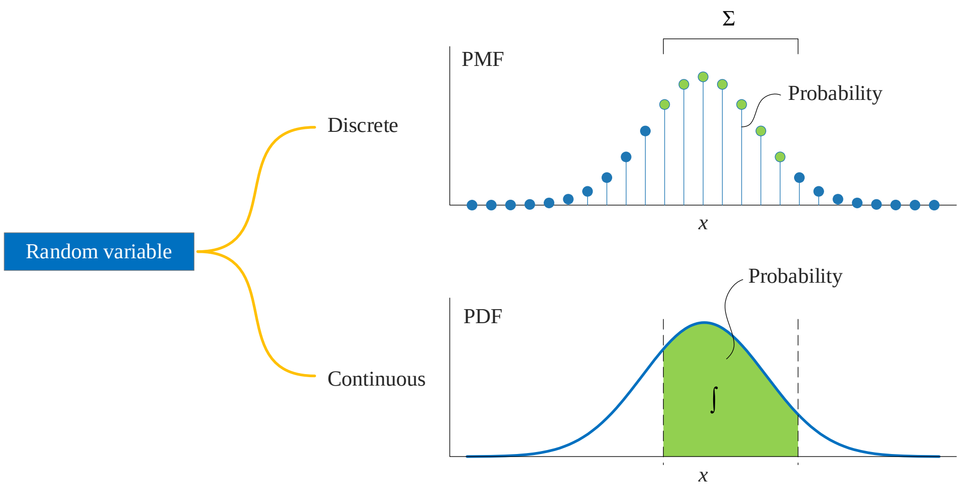


图 2. 比较概率质量函数、概率密度函数

概率质量函数 PMF

图 2 上图所示，**概率质量函数** (probability mass function, PMF) 是离散随机变量在特定取值上的概率。很多教材翻译把 PMF 翻译做“分布列”，本书则直译其为概率质量函数。

概率质量函数本质上就是概率，因此本书很多时候也直接称之为概率。此外，本书大多时候将概率质量函数直接简写为 PMF。

本书用小字斜体字母 p 表达 PMF，比如随机变量 X 的概率质量函数记做 $p_X(x)$。下角标 X 代表描述随机试验的随机变量，概率质量函数的输入为变量 x。而概率质量函数 $p_X(x)$ 的输出则为“概率”。

和函数一样，概率质量函数的输入也可以不止一个。比如，$p_{X,Y}(x, y)$ 代表 (X, Y) 的联合概率质量函数。$p_{X,Y}(x, y)$ 的输入为 (x, y)，函数的输出为“概率”值。本章后文将专门以二元、三元概率质量函数为例讲解多元概率质量函数。

有些资料为了方便，将 $p_X(x)$ 简写为 $p(x)$，$p_{X,Y}(x, y)$ 简做 $p(x, y)$。

$p_X(x)$ 本身就是“概率值”，因此计算离散随机变量 X 取不同值时的概率，我们使用求和运算。因此，$p_X(x)$ 对应的数学运算符是 Σ。

抛一枚硬币

举一个例子，抛一枚硬币试验中，令 X_1 为正面朝上数量，X_1 的 PMF 为：

$$p_{X1}(x) = \begin{cases} 1/2 & x = 0 \\ 1/2 & x = 1 \end{cases} \tag{1}$$

相信读者已经对图 3 不陌生，我们在图像上增加标注，水平轴加 x 代表函数输入，纵轴改为 PMF, $p_{X1}(x)$ 代表概率质量函数。

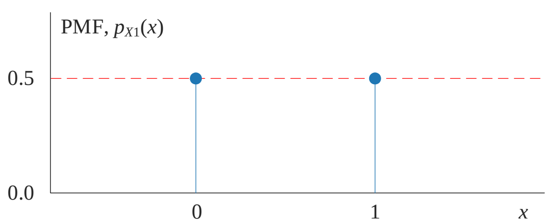


图 3. 随机变量 X_1 的 PMF

抛一个色子

再举一个例子，抛一枚色子试验，令离散随机变量 X_2 为色子点数。如图 4 所示，X_2 的 PMF 为：

$$p_{X2}(x)=\begin{cases}1/6 & x=1\sim 6\\ 0 & \text{Otherwise}\end{cases}\tag{2}$$

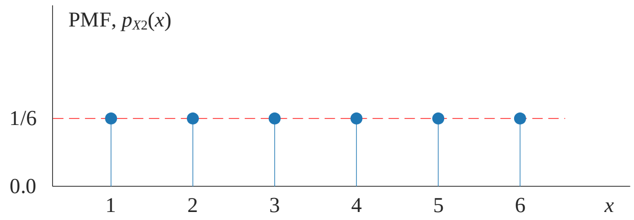


图 4. 离散随机变量 X_2 的 PMF

显然，上述两个例子中 X_1 和 X_2 分别代表不同试验的随机变量，而 x 仅代表实数值。读到这里大家可能已经意识到，在概率质量函数中引入下角标 X_1 和 X_2 能帮助我们区分 $p_{X1}(x)$、$p_{X2}(x)$ 这两个不同的 PMF。但是，一般情况，本书中随机变量和变量形式上对应，比如 $p_{X1}(x_1)$、$p_{X2}(x_2)$、$p_X(x)$、$p_Y(y)$。

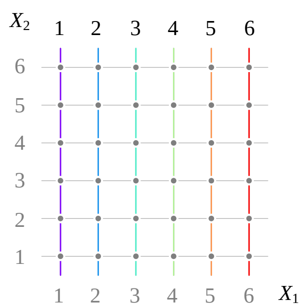
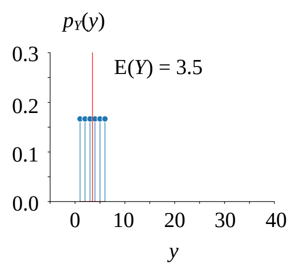
抛两个色子例子

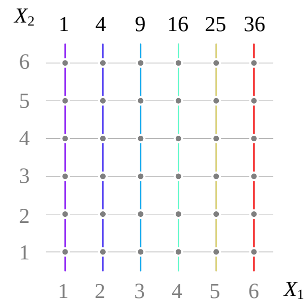
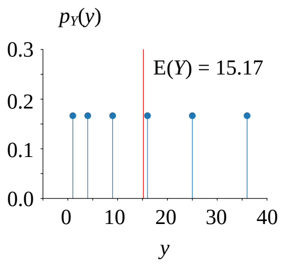
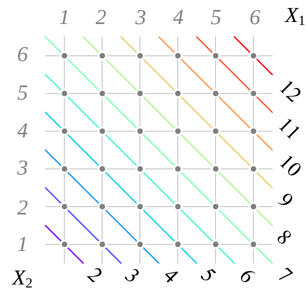
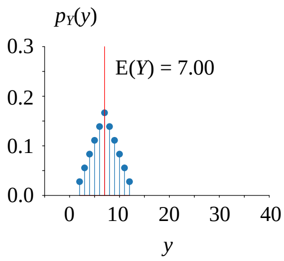
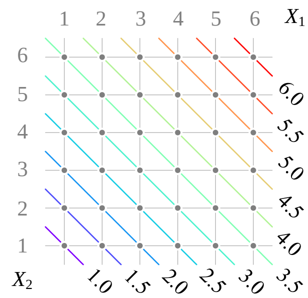
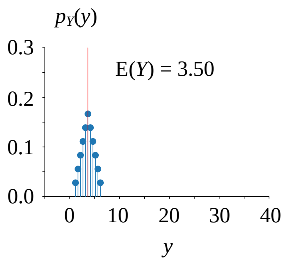
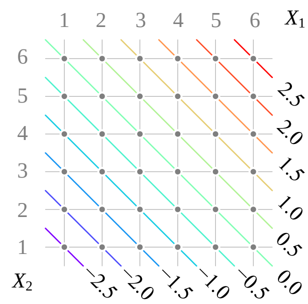
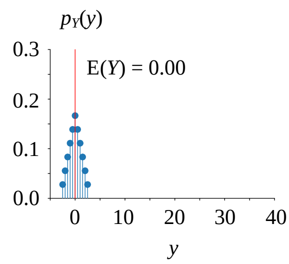
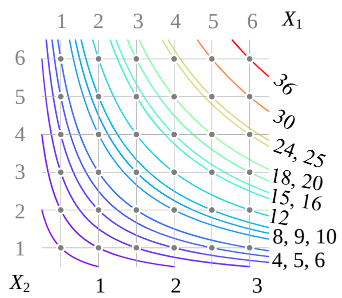
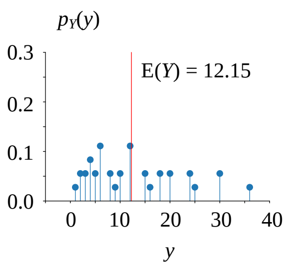
上一章讲过一个例子，一次抛两个色子，第一个色子点数设为 X_1，第二枚色子的点数为 X_2。X_1 和 X_2 可以进行各种数学运算获得随机变量 Y。

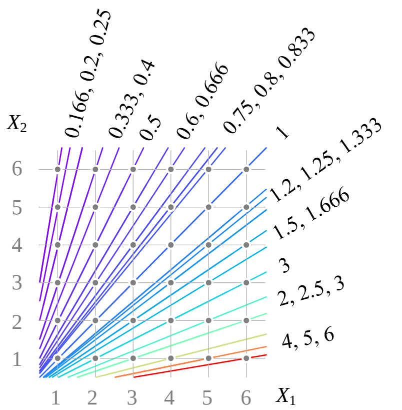
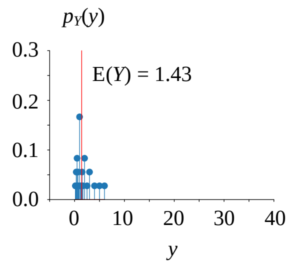
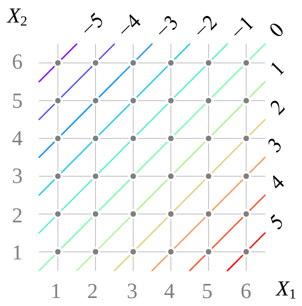
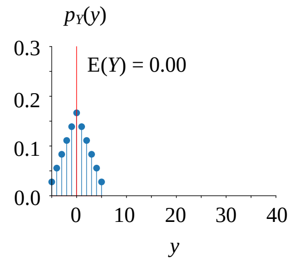
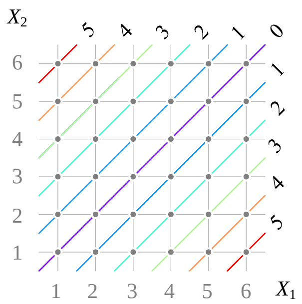
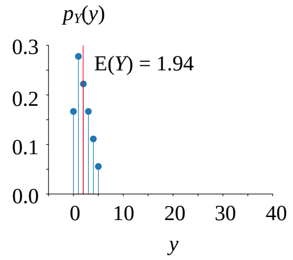
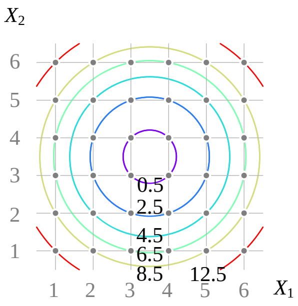
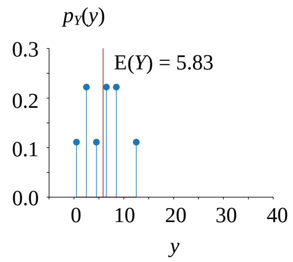
Y 本身有自己的样本空间，样本空间的每个样本都对应特定概率值。利用本章前文内容，我们可以把 $Y = y$ 的概率值写成概率密度函数 $p_Y(y)$。

表 1 总结各种“花式玩法”样本空间，以及概率质量函数 $p_Y(y)$。请大家逐个分析，特别注意概率质量函数的分布规律。

表 1. 基于抛两枚色子试验结果的更多花式玩法

随机变量的函数	样本空间	样本位置	概率质量函数
$Y = X_1$	{1, 2, 3, 4, 5, 6}		

$Y = X_1^2$	{1, 4, 9, 16, 25, 36}		
$Y = X_1 + X_2$	{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}		
$Y = \dfrac{X_1 + X_2}{2}$	{1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0}		
$Y = \dfrac{X_1 + X_2 - 7}{2}$	{−2.5, −2.0, −1.5, −1.0, −0.5, 0.0, 0.5, 1.0, 1.5, 2.0, 2.5}		
$Y = X_1 X_2$	{1, 2, 3, 4, 5, 6, 8, 9, 10, 12, 15, 16, 18, 20, 24, 25, 30, 36}		

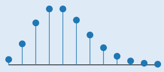
$Y=\frac{X_1}{X_2}$	{0.166, 0.2, 0.25, 0.333, 0.4, 0.5, 0.6, 0.666, 0.75, 0.8, 0.833, 1.0, 1.2, 1.25, 1.333, 1.5, 1.666, 2.0, 2.5, 3.0, 4.0, 5.0, 6.0}		
$Y=X_1-X_2$	{−5, −4, −3, −2, −1, 0, 1, 2, 3, 4, 5}		
$Y=\left\|X_1-X_2\right\|$	{0, 1, 2, 3, 4, 5}		
$Y=(X_1-3.5)^2+(X_2-3.5)^2$	{0.5, 2.5, 4.5, 6.5, 8.5, 12.5}		

代码 Bk5_Ch04_01.py 绘制表 1 中图像。学完本章后续内容后，请大家修改代码计算标准差 std(Y)，并在火柴梗图上展示 E(Y) ± std(Y)。

归一律

一元离散随机变量 X 的概率质量函数 $p_X(x)$ 如下重要性质：

$$\sum_{x} p_X(x) = 1, \quad 0 \le p_X(x) \le 1 \tag{3}$$



上式实际上就是“穷举法”，即遍历所有的可能性，将它们的概率值求和，结果为 1，也叫归一律。值得强调的是，概率质量函数 $p_X(x)$ 最大取值为 1。

概率密度函数 PDF

与 PMF 相对的是，**概率密度函数** (probability density function, PDF)。PDF 对应连续随机变量，本书用小写斜体字母 f 表达 PDF，比如连续随机变量 X 的概率密度函数记做 $f_X(x)$。

当连续随机变量取不同值时，概率密度函数 $f_X(x)$ 用积分方式得到概率值。因此，$f_X(x)$ 对应的数学运算符是积分 $\int$。

注意，联合概率密度函数 $f_{X1,X2,X3}(x_1,x_2,x_3)$ “偏积分”结果还是概率密度。$f_{X1,X2,X3}(x_1,x_2,x_3)$ 三重积分结果才是密度。

举个例子，连续随机变量 X 服从标准正态分布 $N(0, 1)$，其 PDF 为：

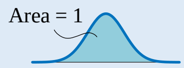
$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \tag{4}$$

举个例子，当 $x = 0$ 时，$f_X(x)$ 约为 0.4，这个值本身是概率密度，不是概率。只有对连续随机变量 PDF 在指定区间内进行积分后才可能是概率。

值得反复强调的是，PMF 本身就是概率，对应的数学工具为 Σ 求和。PDF 积分后才可能是概率，对应的数学工具为 $\int$ 积分。

一元连续随机变量 X 的概率密度函数 $f_X(x)$ 也有如下重要性质：

$$\int_{-\infty}^{+\infty} f_X(x)\,\mathrm{d}x = 1, \quad f_X(x) \ge 0 \tag{5}$$



上式也相当于是“穷举法”，注意概率密度函数 $f_X(x)$ 取值非负，但是不要求小于 1。本书后续将给出具体示例。

概率质量函数 PMF、概率密度函数 PDF 是特殊的函数。特殊之处在于它们的输入为随机变量的取值，输出为概率质量、概率密度。但是，本质上，它们又都是函数。所以，我们可以把函数的分析工具用在概率质量函数 PMF、概率密度函数 PDF 上。

本章和下一章首先讲解离散随机变量。本书第 6、7 章讲解连续随机变量。

区分符号

在此有必要再次区分本系列丛书的容易混淆的代数、线性代数和概率统计符号。以下内容来自《矩阵力量》第 23 章，几乎原封不动拿来。

粗体、斜体、小写 $\boldsymbol{x}$ 为列向量。从概率统计的角度，$\boldsymbol{x}$ 可以代表随机变量 X 采样得到的样本数据，偶尔也代表 X 总体数据。随机变量 X 样本“无序”集合为 $X=\left\{x^{(1)},x^{(2)},\ldots,x^{(n)}\right\}$。很多时候，随机变量 X 本身也可以看成“有序”的数组，即向量。

粗体、斜体、小写、加下标序号的 $\boldsymbol{x}_1$ 为列向量，下角标仅仅是序号，以便区分 $\boldsymbol{x}_1$、$\boldsymbol{x}_2$、$\boldsymbol{x}_j$、$\boldsymbol{x}_D$ 等等。从概率统计的角度，$\boldsymbol{x}_1$ 可以代表随机变量 X_1 样本数据，也可以表达 X_1 总体数据。

行向量 $\boldsymbol{x}^{(1)}$ 代表一个具有多个特征的样本点。注意，在机器学习算法中，为了方便，$\boldsymbol{x}^{(1)}$ 偶尔也代表列向量。

从代数角度，斜体、小写、非粗体 x_1 代表变量，下角标代表变量序号。这种记法常用在函数解析式中，比如线性回归解析式 $y = x_1 + x_2$。在概率质量函数、概率密度函数中，我们看到也用做函数输入，比如 $p_{X1}(x_1)$、$f_{X2}(x_2)$。

$x^{(1)}$ 代表变量 x 的一个取值，或代表随机变量 X 的一个取值。

而 $x_1^{(1)}$ 代表变量 x_1 的一个取值，或代表随机变量 X_1 的一个取值，比如 $X_1=\left\{x_1^{(1)},x_1^{(2)},\ldots,x_1^{(n)}\right\}$。

粗体、斜体、大写 $\boldsymbol{X}$ 则专门用来表达多行、多列的数据矩阵，$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_D]$。数据矩阵 $\boldsymbol{X}$ 中第 i 行、第 j 列元素则记做 $x_{i,j}$。多元线性回归中，$\boldsymbol{X}$ 也叫**设计矩阵** (design matrix)。

我们还会用粗体、斜体、小写希腊字母 $\boldsymbol{\chi}$ (chi，读作/ˈkaɪ/) 代表 D 维随机变量构成的列向量，$\boldsymbol{\chi} = [X_1, X_2, \ldots, X_D]^{\mathrm{T}}$。希腊字母 $\boldsymbol{\chi}$ 主要用在多元概率统计中，比如，多元概率密度函数 $f_{\boldsymbol{\chi}}(\boldsymbol{x})$。

4.2 期望值

期望值

离散随机变量 X 有 n 个取值 $\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$，$X$ 的**期望** (expectation)，也叫**期望值** (expected value)，$\mathrm{E}(X)$ 为：

$$\underbrace{\mathrm{E}(X)}_{\text{Scalar}}=\mu_X=x^{(1)}p_X\left(x^{(1)}\right)+x^{(2)}p_X\left(x^{(2)}\right)+\cdots+x^{(n)}p_X\left(x^{(n)}\right)=\sum_{i=1}^{n}x^{(i)}\cdot\underbrace{p_X\left(x^{(i)}\right)}_{\text{Weight}}\tag{6}$$

上式相当于加权平均数，边缘 PMF $p_X(x)$ 就代表权重。期望值把随机变量一系列取值转化成了一个标量数值，这相当于降维。如图 5 所示，从矩阵乘法角度，计算期望值相当于将 X 这个维度折叠。

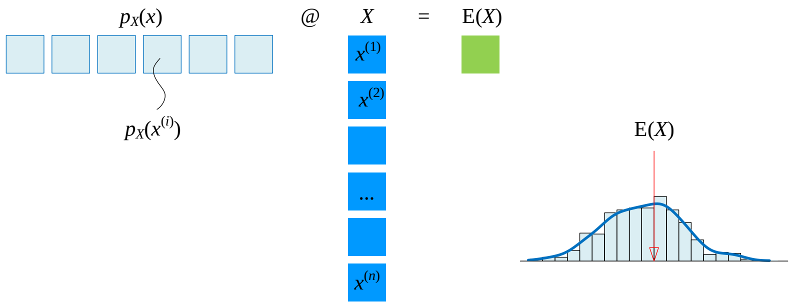
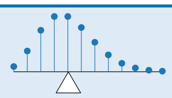


图 5. 计算离散随机变量 X 期望值/均值

为了方便，我们经常把 (6) 简写作：

$$\mathrm{E}(X) = \sum_{x} x \cdot p_X(x) \tag{7}$$



$\sum_{x}(\cdot)$ 代表对 x 的遍历，也就是穷举。我们知道求加权平均值时，权重之和为 1，也就是说边缘 PMF $p_X(x)$ 满足 $\sum_{x} p_X(x) = 1$。我们也经常把期望值 (均值) 叫做**质心** (centroid)。

举个例子

图 4 中随机变量 X_2 的期望值为：

$$\mathrm{E}(X_2) = \sum_{x_2} x_2 \cdot \underbrace{p_{X2}(x_2)}_{\text{Weight}} = \sum_{x_2} x_2 \cdot \underbrace{\frac{1}{6}}_{\text{Weight}} = 1\times\frac{1}{6}+1\times\frac{1}{6}+1\times\frac{1}{6}+1\times\frac{1}{6}+1\times\frac{1}{6}+1\times\frac{1}{6} = 3.5 \tag{8}$$

大家已经发现上式中随机变量 X_2 的概率密度函数为定值。这和求样本均值的情况类似。求样本均值时，我们用到的权重为 $1/n$，即每个样本赋予相同的权重。

图 6 所示为投色子实验均值随试验次数变化。随着重复次数接近无穷大，试验结果的算术平均值不断地靠近期望值 (理论值)。

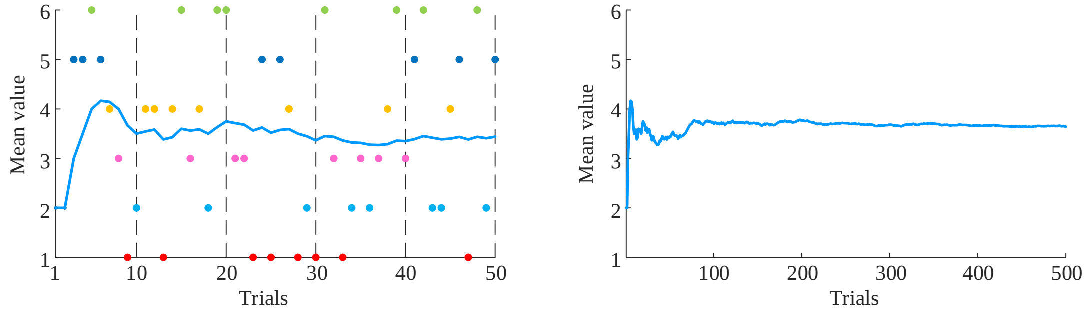


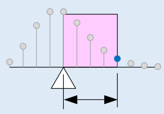
图 6. 投色子试验均值随试验次数变化

4.2 方差

方差

随机变量 X 另外一个重要特征是**方差** (variance)，记做 var(X)。对于离散随机变量 X，方差用来度量 X 和数学期望 E(X) 之间的偏离程度。具体定义为：

$$\operatorname{var}(X) = \overbrace{\mathrm{E}\left[\left(\underbrace{X - \mathrm{E}(X)}_{\text{Deviation}}\right)^2\right]}^{\text{Expectation}} = \sum_x \left(\underbrace{x - \mathrm{E}(X)}_{\text{Demean}}\right)^2 \cdot \underbrace{p_X(x)}_{\text{Weight}} \tag{9}$$



上式中 $x - \mathrm{E}(X)$ 代表以期望值 $\mathrm{E}(X)$ 为参照，样本点 x 的偏离量。如图 7 所示，$X - \mathrm{E}(X)$ 就是**去均值** (demean)，也叫**中心化** (centralize)。

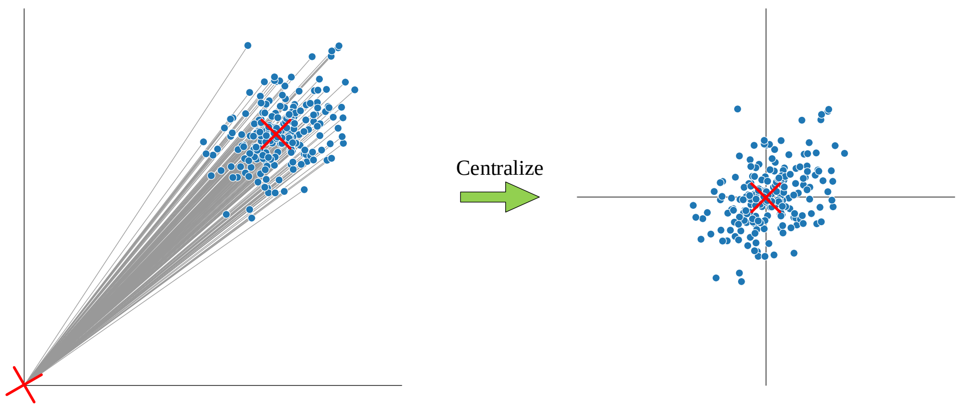


图 7. 样本去均值

观察 (9)，容易发现方差实际上是 $\left(X-\mathrm{E}(X)\right)^2$ 的期望值。上式就是求 $\left(x-\mathrm{E}(X)\right)^2$ 的加权平均数，权重为 $p_X(x)$。从几何角度，$\left(X-\mathrm{E}(X)\right)^2$ 代表以 $\left|X-\mathrm{E}(X)\right|$ 为边长的正方形的面积。而对于离散随机变量，$p_X(x)$ 就是投票权，体现不同样本重要性。

举个例子

图 4 对应的方差为：

$$
\begin{aligned}
\operatorname{var}(X_2) &= \frac{1}{6}\times(1-3.5)^2+\frac{1}{6}\times(2-3.5)^2+\frac{1}{6}\times(3-3.5)^2+\frac{1}{6}\times(4-3.5)^2+\frac{1}{6}\times(5-3.5)^2+\frac{1}{6}\times(6-3.5)^2 \\
&= \frac{1}{6}\times\left(\frac{25}{4}+\frac{9}{4}+\frac{1}{4}+\frac{1}{4}+\frac{9}{4}+\frac{25}{4}\right)=\frac{35}{12}\approx 2.9167
\end{aligned}
\tag{10}
$$

注意，本书前文在计算样本方差时，分母除以 $n-1$。而 (10) 分母相当于除以 n，这是因为对随机变量求方差，相当于对总体求方差。而且，恰好 X_2 取 1 ~ 6 这 6 个不同值是对应的概率相等。

也就是说，特别地离散随机变量 X 的概率质量函数为等概率时，即：

$$
p_X(x)=\frac{1}{n} \tag{11}
$$

(9) 可以写成：

$$
\operatorname{var}(X)=\frac{1}{n}\sum_x\left(x-\mathrm{E}(X)\right)^2 \tag{12}
$$

再次强调，上式是求离散随机变量方差的一种特殊情况。统计中，样本的方差计算方法类似上式，不过要将分母中的 n 换成 $n-1$。

技巧：方差计算

方差有个简便算法：

$$
\operatorname{var}(X)=\underbrace{\mathrm{E}(X^2)}_{\text{Expectaton of } X^2}-\underbrace{\mathrm{E}(X)^2}_{\text{Square of } \mathrm{E}(X)} \tag{13}
$$

其中，为：

$$
\underbrace{\mathrm{E}(X^2)}_{\text{Expectaton of } X^2}=\sum_x x^2\cdot\underbrace{p_X(x)}_{\text{Weight}} \tag{14}
$$

(13) 的推导过程如下所示：

$$\begin{aligned}\operatorname{var}(X)&=\mathrm{E}\left(\left(X-\mathrm{E}(X)\right)^2\right)\\&=\mathrm{E}\left(X^2-2X\cdot\mathrm{E}(X)+\mathrm{E}(X)^2\right)\\&=\mathrm{E}\left(X^2\right)-2\mathrm{E}(X)\cdot\mathrm{E}(X)+\mathrm{E}(X)^2\\&=\mathrm{E}\left(X^2\right)-\mathrm{E}(X)^2\end{aligned}\tag{15}$$

注意，(13) 也适用于连续随机变量。请大家用 (15) 计算 (10) 的方差。

几何意义

下面我们聊聊 (15) 的几何含义。

方差度量离散程度，本质上来说是"自己"和"自己"比较的产物。前一个"自己"是 X 每个样本，后一个"自己"是代表 X 整体位置的期望值 $\mathrm{E}(X)$。

下式，计算方差 $\operatorname{var}(X)$ 有 $\mathrm{E}\left(X^2\right)$ 和 $-\mathrm{E}(X)^2$ 两部分：

$$\operatorname{var}(X)=\mathrm{E}\left(X^2\right)-\mathrm{E}(X)^2\tag{16}$$

利用图 8 解剖来看，方差 $\operatorname{var}(X)$ 代表样本以质心 (centroid) 为基准的离散程度。

$\mathrm{E}\left(X^2\right)$ 度量样本以原点 (origin) 为基准的离散程度。

$-\mathrm{E}(X)^2$ 则代表消除"原点"影响，也就是方差中的"去均值"。

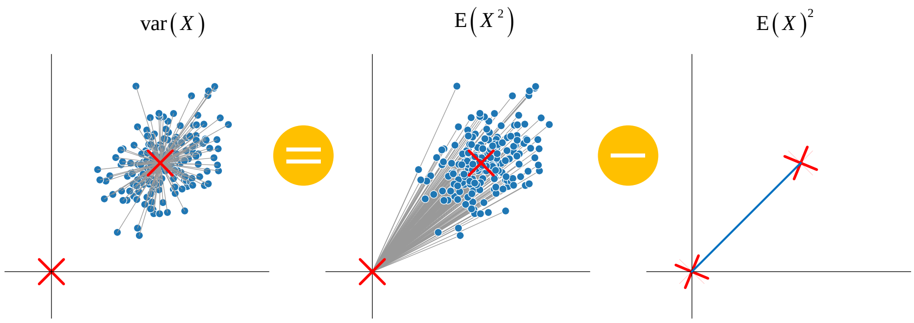


图 8. 几何视角理解计算方差技巧

标准差

标准差 (standard deviation) 是方差的平方根：

$$\operatorname{std}(X)=\sigma_X=\sqrt{\operatorname{var}(X)}\tag{17}$$

方差既然可以用来度量“离散程度”，为什么我们还需要标准差？简单来说，标准差 σ_X、期望值 E(X)、随机变量 X 为同一量纲。比如，鸢尾花花萼长度 X 的单位是 cm，σ_X 的单位也对应是 cm。但是，var(X) 的量纲是 cm^2。

汇总

期望值、方差、标准差本身是一个个特定标量值。折叠，总结，汇总，降维，压扁，本章及本书后文会用这些字眼形容期望值、方差、标准差。这是因为，计算期望值、方差、标准差时，我们不再关注一个个样本数据的具体取值，而是在乎某种方式的汇总 (aggregation)。

如果汇总的形式为期望，它相当于计算平均数、质心。如果汇总的形式为方差，方差度量离散程度。其他常用的汇总形式还包括：计数 (count)、求和、百分位 (percentile)、均方差 (standard deviation)、最大值 (maximum)、最小值 (minimum)、中位数 (median)、众数 (mode) 等等。

4.3 累积分布函数 CDF：累加

如果 X 是离散的，**累积分布函数** (Cumulative Distribution Function, CDF)，又叫分布函数，本书记做 $F_X(x)$。对于离散随机变量，累积分布函数对应概率质量函数的求和。

$F_X(x)$ 的定义为：

$$F_X(x) = \Pr(X \le x) = \sum_{t \le x} p_X(t) \tag{18}$$

上式相当于累加概念，累加截止于 $X = x$。

离散随机变量 X 的取值范围为 $a < X \le b$ 时，对应的概率可以利用 CDF 计算：

$$\Pr(a < X \le b) = F_X(b) - F_X(a) \tag{19}$$

图 4 对应的 CDF 图像为图 9。

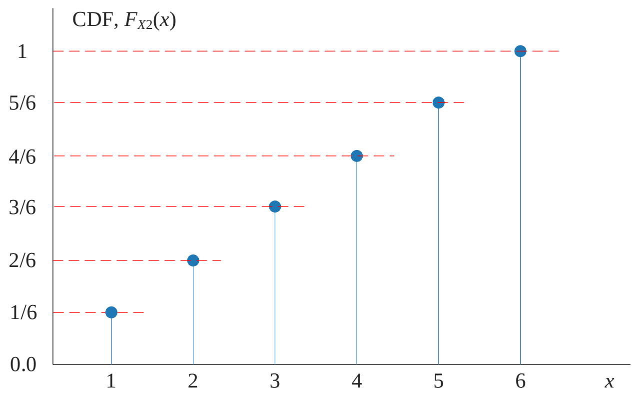


图 9. 随机变量 X_2 的 CDF

4.4 二元离散随机变量

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

概率质量函数 $p_{X,Y}(x, y)$ 代表事件 $\{X = x, Y = y\}$ 发生的联合概率：

$$\underbrace{p_{X,Y}(x,y)}_{\text{Joint}} = \Pr(X = x, Y = y) = \Pr(X = x \cap Y = y) \tag{20}$$

再次强调，对于离散随机变量，$p_{X,Y}(x, y)$ 本身就是概率值。图 10 所示为二元离散随机变量 (X, Y) 的样本空间 Ω，空间中共有 81 个点。从函数角度来看，$p_{X,Y}(x, y)$ 是个二元函数。因此，我们可以用二元函数的分析方法来讨论 $p_{X,Y}(x, y)$。

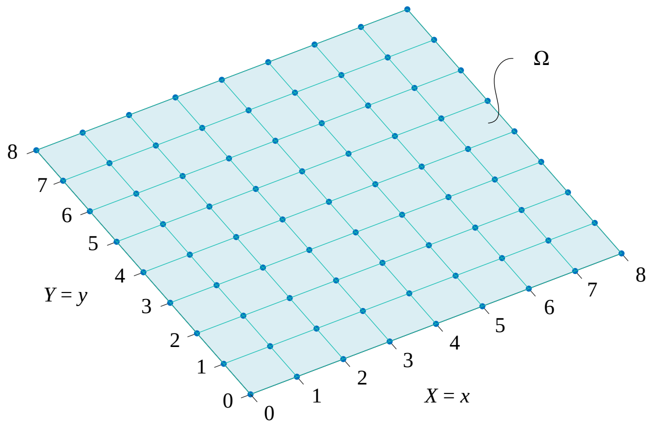


图 10. 二元随机变量的样本空间

取值

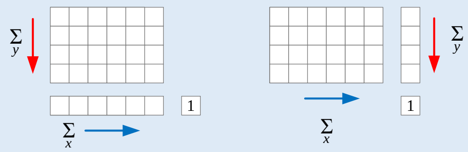
图 11 所示为二元联合概率密度函数 $p_{X,Y}(x, y)$ 的取值表格。

Joint, $p_{X,Y}(x,y)$	$X = x$: 0	1	2	3	4	5	6	7	8
$Y = y$: 8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003	0.0004	0.0002	0.0001
6	0.0000	0.0000	0.0001	0.0005	0.0014	0.0025	0.0030	0.0020	0.0006
5	0.0000	0.0001	0.0005	0.0022	0.0064	0.0119	0.0138	0.0092	0.0027
4	0.0000	0.0002	0.0014	0.0064	0.0185	0.0346	0.0404	0.0269	0.0078
3	0.0000	0.0003	0.0025	0.0119	0.0346	0.0646	0.0753	0.0502	0.0146
2	0.0000	0.0004	0.0030	0.0138	0.0404	0.0753	0.0879	0.0586	0.0171
1	0.0000	0.0002	0.0020	0.0092	0.0269	0.0502	0.0586	0.0391	0.0114
0	0.0000	0.0001	0.0006	0.0027	0.0078	0.0146	0.0171	0.0114	0.0033

图 11. 概率质量函数 $p_{X,Y}(x, y)$ 取值

我们同时用热图来可视化 $p_{X,Y}(x, y)$。二元联合概率密度函数 $p_{X,Y}(x, y)$ 也有一条重要的性质：

$$\sum_{x}\sum_{y}\underbrace{p_{X,Y}(x,y)}_{\text{Joint}}=\sum_{y}\sum_{x}\underbrace{p_{X,Y}(x,y)}_{\text{Joint}}=1,\quad 0\le p_{X,Y}(x,y)\le 1 \tag{21}$$



也就是说，图 11 这幅热图中所有数值求和的结果为 1，和求和顺序无关。

火柴梗图

二元联合概率密度函数 $p_{X,Y}(x, y)$ 可以长成什么样子呢？火柴梗图最适合可视化概率质量函数，如图 12 所示。注意，为了展示火柴梗图分别沿 X、Y 方向变化趋势，图 12 将火柴梗散点连线。一般情况，火柴梗图不存在连线。

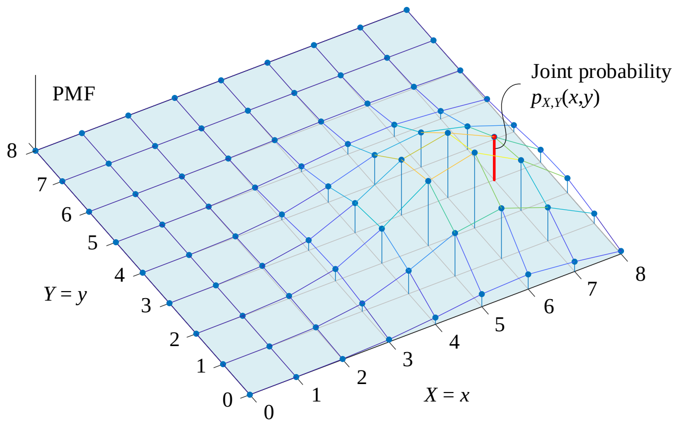


图 12. $p_{X,Y}(x, y)$ 对应的二维火柴梗图

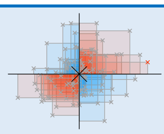
4.5 协方差、相关性系数

本书读者对协方差、相关性系数这两个概念应该不陌生，本节简要介绍如何求解离散随机变量的协方差和相关性系数。

协方差

一对离散随机变量 (X, Y) 的协方差定义为：

$$\operatorname{cov}(X,Y)=\mathrm{E}\Big(\big(X-\mathrm{E}(X)\big)\big(Y-\mathrm{E}(Y)\big)\Big) \tag{22}$$



如果 (X, Y) 的二元 PMF 为 $p_{X,Y}(x, y)$，X 的取值为 $x^{(i)}$ $(i = 1, 2, \dots, n)$，Y 的取值为 $y^{(j)}$ $(j = 1, 2, \dots, m)$。(22) 可以展开写成：

$$\begin{aligned}\operatorname{cov}(X,Y) &= \mathrm{E}\Big(\big(X-\mathrm{E}(X)\big)\big(Y-\mathrm{E}(Y)\big)\Big) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} p_{X,Y}\left(x^{(i)},y^{(j)}\right)\left(x^{(i)}-\mathrm{E}(X)\right)\left(y^{(j)}-\mathrm{E}(Y)\right)\end{aligned} \tag{23}$$

其中，

$$\mathrm{E}(X) = \sum_{x} x\cdot p_X(x),\quad \mathrm{E}(Y) = \sum_{y} y\cdot p_Y(y) \tag{24}$$

(23) 常简写为：

$$\operatorname{cov}(X,Y) = \sum_{x}\sum_{y} p_{X,Y}(x,y)\big(x-\mathrm{E}(X)\big)\big(y-\mathrm{E}(Y)\big) \tag{25}$$

类似方差，协方差运算也有如下技巧：

$$\begin{aligned}\operatorname{cov}(X,Y) &= \mathrm{E}(XY)-\mathrm{E}(X)\cdot\mathrm{E}(Y) \\ &= \sum_{x}\sum_{y} x\cdot y\cdot p_{X,Y}(x,y) - \left(\sum_{x} x\cdot p_X(x)\right)\cdot\left(\sum_{y} y\cdot p_Y(y)\right)\end{aligned} \tag{26}$$

推导过程如下所示：

$$\begin{aligned}\operatorname{cov}(X,Y) &= \mathrm{E}\Big(\big(X-\mathrm{E}(X)\big)\big(Y-\mathrm{E}(Y)\big)\Big) \\ &= \mathrm{E}\Big(XY-\mathrm{E}(X)Y-X\,\mathrm{E}(Y)+\mathrm{E}\big(X\,\mathrm{E}(Y)\big)\Big) \\ &= \mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y)-\mathrm{E}(X)\mathrm{E}(Y)+\mathrm{E}(X)\mathrm{E}(Y) \\ &= \mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y)\end{aligned} \tag{27}$$

相关性

相关性的定义：

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y} \tag{28}$$

展开得到：

$$\rho_{X,Y} = \frac{\mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y)}{\sqrt{\mathrm{E}(X^2)-\big(\mathrm{E}(X)\big)^2}\sqrt{\mathrm{E}(Y^2)-\big(\mathrm{E}(Y)\big)^2}} \tag{29}$$

相关性的取值范围 [−1, 1]。相比于协方差，相关性更适合横向比较。本书第 15 章将专门讲解相关性。

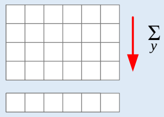
4.6 边缘概率：偏求和，相当于降维

边缘概率 (marginal probability) 是某个事件发生的概率，而与其它事件无关。对于离散随机变量来说，利用全概率定理，也就是穷举法，我们可以把联合概率结果中不需要的那些事件全部合并。合并的过程叫做**边缘化** (marginalization)，用到的数学工具为《数学要素》第 14 章讲到的“偏求和”。

边缘概率 $p_X(x)$

根据全概率公式，对于二元联合概率密度函数 $p_{X,Y}(x, y)$，求解边缘概率 $p_X(x)$ 相当于利用“偏求和”消去 y：

$$\underbrace{p_X(x)}_{\text{Marginal}} = \sum_{y} \underbrace{p_{X,Y}(x, y)}_{\text{Joint}} \tag{30}$$



也就是说，在 $X = x$ 取值条件下，概率质量函数 PMF 对所有 y 的求和。

从函数角度来看，$p_X(x)$ 是个一元函数。

从矩阵运算角度来看，$p_{X,Y}(x, y)$ 代表矩阵，矩阵沿 Y 方向求和，折叠得到行向量 $p_X(x)$。行向量 $p_X(x)$ 进一步求和折叠结果为标量 1。

举个例子

如图 13 所示，当 $X = 6$ 时，将整个一列的 PMF 求和得到 $p_X(6) = 0.2965$。

Joint, $p_{X,Y}(x,y)$

$Y = y$ \ $X = x$	0	1	2	3	4	5	6	7	8	Marginal, $p_Y(y)$
8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
7	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003	0.0004	0.0002	0.0001	0.0012
6	0.0000	0.0000	0.0001	0.0005	0.0014	0.0025	0.0030	0.0020	0.0006	0.0100
5	0.0000	0.0001	0.0005	0.0022	0.0064	0.0119	0.0138	0.0092	0.0027	0.0467
4	0.0000	0.0002	0.0014	0.0064	0.0185	0.0346	0.0404	0.0269	0.0078	0.1361
3	0.0000	0.0003	0.0025	0.0119	0.0346	0.0646	0.0753	0.0502	0.0146	0.2541
2	0.0000	0.0004	0.0030	0.0138	0.0404	0.0753	0.0879	0.0586	0.0171	0.2965
1	0.0000	0.0002	0.0020	0.0092	0.0269	0.0502	0.0586	0.0391	0.0114	0.1977
0	0.0000	0.0001	0.0006	0.0027	0.0078	0.0146	0.0171	0.0114	0.0033	0.0576
Marginal, $p_X(x)$	0.0001	0.0012	0.0100	0.0467	0.1361	0.2541	0.2965	0.1977	0.0576	

Σ_x Σ_y

图 13. 利用联合概率计算边缘概率

边缘概率 $p_Y(y)$

同理，随机变量 Y 的边缘分布 $p_Y(y)$ 通过“偏求和”和消去 x 得到：

$$\underbrace{p_Y(y)}_{\text{Marginal}} = \sum_{x} \underbrace{p_{X,Y}(x,y)}_{\text{Joint}} \tag{31}$$

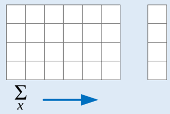


图 13 所示，当 $Y = 2$ 时，将整个一行的 PMF 相加得到 $p_Y(2) = 0.2965$。

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 $p_Y(y)$。列向量 $p_Y(y)$ 进一步折叠结果同样为标量 1。

几何视角：叠加

显然，边缘分布 $p_X(x)$ 和 $p_Y(y)$ 本身也是概率质量函数。从图像上来看，$p_X(x)$ 相当于 $p_{X,Y}(x, y)$ 中 y 在取不同值时对应的火柴梗图叠加得到，具体如图 14 所示。同理，图 15 所示为边缘分布 $p_Y(y)$ 求解过程。

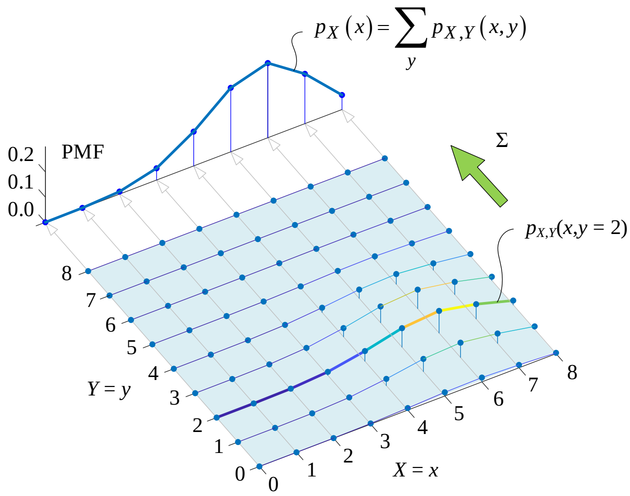


图 14. 边缘分布 $p_X(x)$ 求解过程

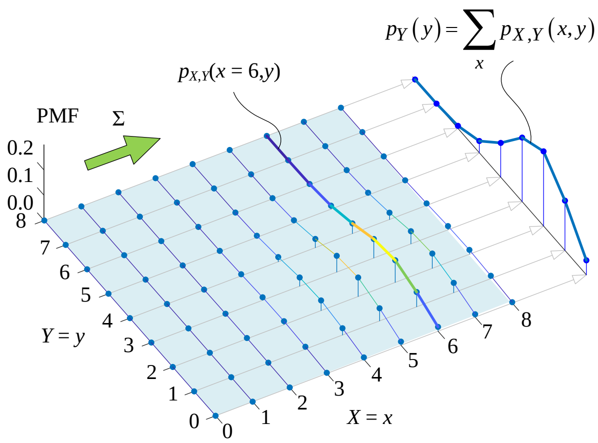


图 15. 边缘分布 $p_Y(y)$ 求解过程

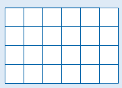
4.7 条件概率：引入贝叶斯定理

本节利用贝叶斯定理，介绍如何求解离散随机变量的条件概率质量函数。

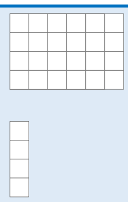
联合概率 $p_{X,Y}(x,y)$ → 条件概率 $p_{X|Y}(x|y)$

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。也就是说，对于 $p_{X|Y}(x|y)$ 事件 $\{Y = y\}$ 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：



$$\underbrace{p_{X|Y}(x|y)}_{\text{Conditional}} = \frac{\overbrace{p_{X,Y}(x,y)}^{\text{Joint}}}{\underbrace{p_Y(y)}_{\text{Marginal}}} \tag{32}$$

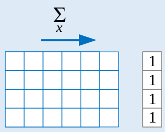


从函数角度来看，$p_{X|Y}(x \mid y)$ 本质上也是个二元函数。也就是虽然，$Y = y$ 为条件，但是这个条件也可以变动。而的变动就会导致概率质量函数 $p_{X|Y}(x \mid y)$ 变化。

从矩阵运算角度来看，$p_{X,Y}(x,y)$ 相当于矩阵，$p_Y(y)$ 相当于列向量；两者相除用到《矩阵力量》第 4 章讲的广播原则 (broadcasting)。得到的条件概率 $p_{X|Y}(x|y)$ 本质上也是个矩阵。

$p_{X|Y}(x|y)$ 对 x 求和等于 1：

$$\sum_x p_{X|Y}(x|y) = 1 \tag{33}$$



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”。

举个例子

如图 16 所示，$Y = 2$ 时，边缘概率 $p_Y(Y = 2)$ 可以通过求和得到：

$$p_Y(2) = \sum_x p_{X,Y}(x,2) \tag{34}$$

$p_Y(2)$ 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：

$$\underbrace{p_{X|Y}(x|2)}_{\text{Conditional}} = \frac{\overbrace{p_{X,Y}(x,2)}^{\text{Joint}}}{\underbrace{p_Y(2)}_{\text{Marginal}}} \tag{35}$$

观察图 16，发现 $p_{X,Y}(x,2)$ 到 $p_{X|Y}(x|2)$ 相当于曲线缩放过程。

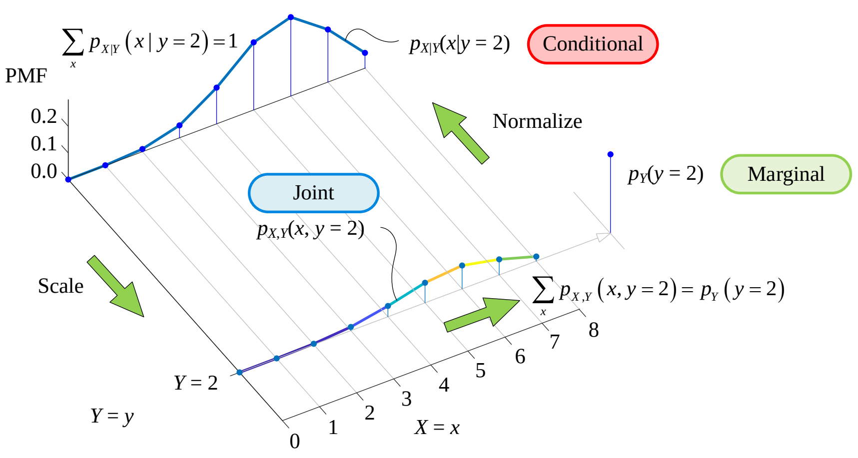


图 16. 求解条件概率 $p_{X|Y}(x|y)$ 的过程

进一步，条件概率 $p_{X|Y}(x|2)$ 对 x 求和得到 1：

$$\sum_x p_{X|Y}(x|2) = \frac{\sum_x p_{X,Y}(x,2)}{p_Y(2)} = \frac{p_Y(2)}{p_Y(2)} = 1 \tag{36}$$

$p_{X,Y}(x,2)$ 到 $p_{X|Y}(x|2)$ 是一个**归一化** (normalization) 过程。也就是说，$p_Y(y)$ 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：

$$\underbrace{p_X(x)}_{\text{Marginal}} = \sum_y \underbrace{p_{X,Y}(x,y)}_{\text{Joint}} = \sum_y \underbrace{p_{X|Y}(x|y)}_{\text{Conditional}} \underbrace{p_Y(y)}_{\text{Marginal}} \tag{37}$$

也就是说，当给定 $Y = y$ 条件下，通过“穷举法”，所有的条件概率值之和为 1。

条件概率 $p_{X|Y}(x|y) \to$ 联合概率 $p_{X,Y}(x,y)$

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：

$$\underbrace{p_{X,Y}(x,y)}_{\text{Joint}} = \underbrace{p_{X|Y}(x|y)}_{\text{Conditional}} \underbrace{p_Y(y)}_{\text{Marginal}} \tag{38}$$

条件概率 $p_{Y|X}(y|x)$

同理，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 $p_{Y|X}(y|x)$ 表达：

$$\underbrace{p_{Y|X}\left(y|x\right)}_{\text{Conditional}} = \frac{\overbrace{p_{X,Y}\left(x,y\right)}^{\text{Joint}}}{\underbrace{p_X\left(x\right)}_{\text{Marginal}}} \tag{39}$$

图 17 展示求解条件概率 $p_{Y|X}(y|x)$ 过程。同样，从函数角度来看，$p_{Y|X}(y \mid x)$ 也是个二元函数。从矩阵运算角度，上式也用到了广播原则，结果 $p_{Y|X}(y \mid x)$ 同样是个矩阵。

$p_{Y|X}(y \mid x)$ 对 y 求和等于 1：

$$\sum_y p_{Y|X}\left(y|x\right) = 1 \tag{40}$$

$\sum_y$

1 1 1 1 1 1

(39) 也可以用来反求联合概率 $p_{Y,X}(y,x)$：

$$\underbrace{p_{X,Y}\left(x,y\right)}_{\text{Joint}} = \underbrace{p_{Y|X}\left(y|x\right)}_{\text{Conditional}} \cdot \underbrace{p_X\left(x\right)}_{\text{Marginal}} \tag{41}$$

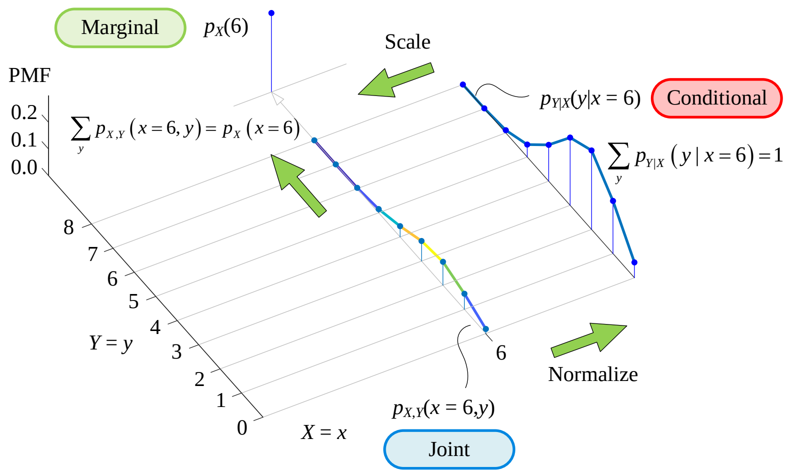


图 17. 求解条件概率 $p_{Y|X}(y|x)$ 的过程

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

$$\underbrace{p_Y\left(y\right)}_{\text{Marginal}} = \sum_x p_{X,Y}\left(x,y\right) = \sum_y \underbrace{p_{Y|X}\left(y|x\right)}_{\text{Conditional}} \underbrace{p_X\left(x\right)}_{\text{Marginal}} \tag{42}$$

上式也是一个“偏求和”过程。

4.8 独立性：条件概率等于边缘概率

独立

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：

$$\underbrace{p_{X|Y}(x|y)}_{\text{Conditional}} = \underbrace{p_X(x)}_{\text{Marginal}} \tag{43}$$

如图 18 所示，X 和 Y 独立，y 不会影响到条件概率；不管 y 取任何值 $(0 \sim 8)$，$p_X(x)$ 的形状和 $p_{X|Y}(x|y)$ 相同。(43) 等价下式：

$$\underbrace{p_{Y|X}(y|x)}_{\text{Conditional}} = \underbrace{p_Y(y)}_{\text{Marginal}} \tag{44}$$

同理，如图 19 所示，X 和 Y 独立时，$p_Y(y)$ 的形状和 $p_{Y|X}(y|x)$ 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。

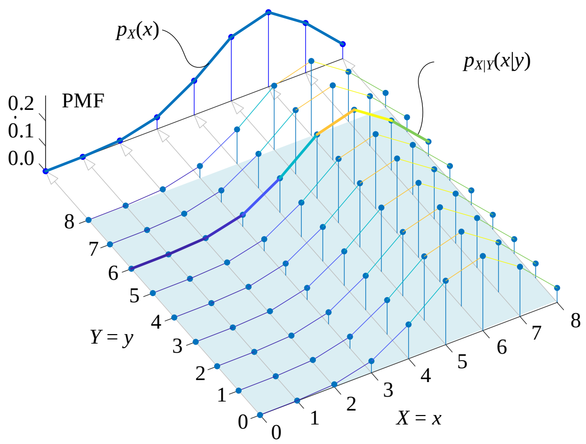


图 18. X 和 Y 独立，条件概率 $p_{X|Y}(x|y)$ 等于边缘概率 $p_X(x)$

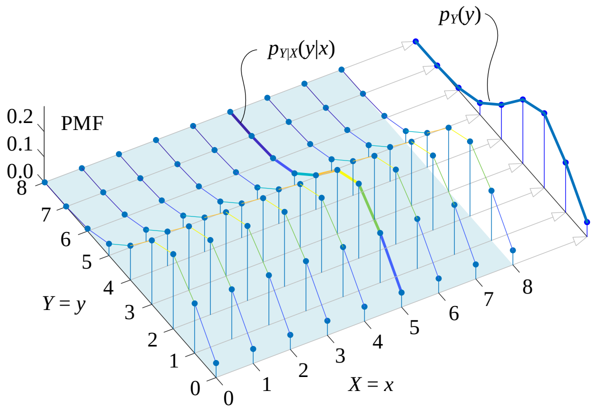
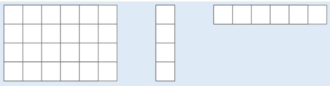


图 19. X 和 Y 独立，条件概率 $p_{Y|X}(y|x)$ 等于边缘概率 $p_Y(y)$

独立：计算联合概率 $p_{X,Y}(x,y)$

另外一个角度，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：



$$\underbrace{p_{X,Y}(x,y)}_{\text{Joint}} = \underbrace{p_Y(y)}_{\text{Marginal}} \cdot \underbrace{p_X(x)}_{\text{Marginal}} \tag{45}$$

从向量角度来看，把 $p_Y(y)$ 和 $p_X(x)$ 看成是两个向量，上式相当于 $p_Y(y)$ 和 $p_X(x)$ 的张量积。

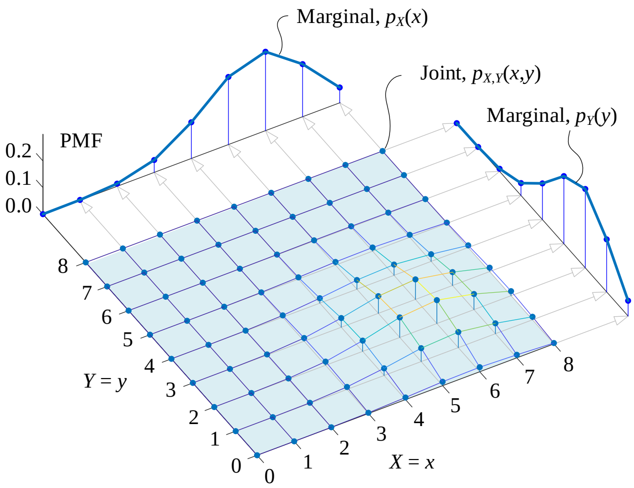


图 20. 联合概率 $p_{X,Y}(x,y)$ 等于 $p_Y(y)$ 和 $p_X(x)$ 两个边缘概率乘积

不独立

我们再来看一下，在离散随机变量 X 和 Y 不独立的情况下，$p_{Y|X}(y|x)$ 和 $p_Y(y)$ 图像可能存在的某种关系。图 21 给出另一个联合概率 $p_{X,Y}(x, y)$ 的图像。

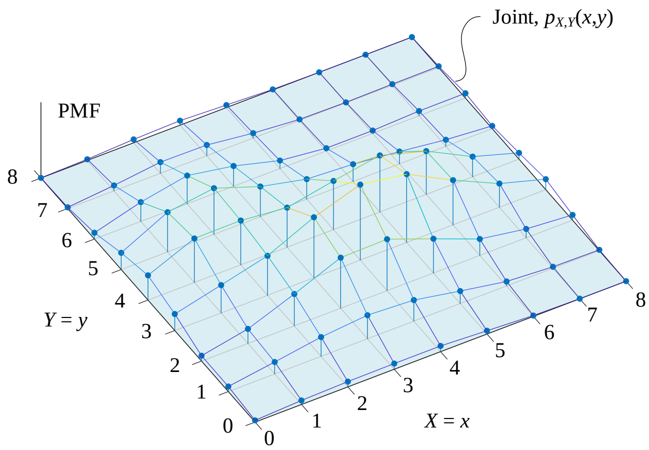


图 21. 离散随机变量 X 和 Y 不独立情况下，联合概率 $p_{X,Y}(x,y)$

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：

$$\underbrace{p_{X|Y}\left(x|y\right)}_{\text{Conditional}} = \frac{\overbrace{p_{X,Y}\left(x,y\right)}^{\text{Joint}}}{\underbrace{p_Y\left(y\right)}_{\text{Marginal}}} = \frac{\overbrace{p_{X,Y}\left(x,y\right)}^{\text{Joint}}}{\sum_x p_{X,Y}\left(x,y\right)} \tag{46}$$

如图 22 所示，当 X 和 Y 不独立，条件概率 $p_{X|Y}(x|y)$ 不同于边缘概率 $p_X(x)$。

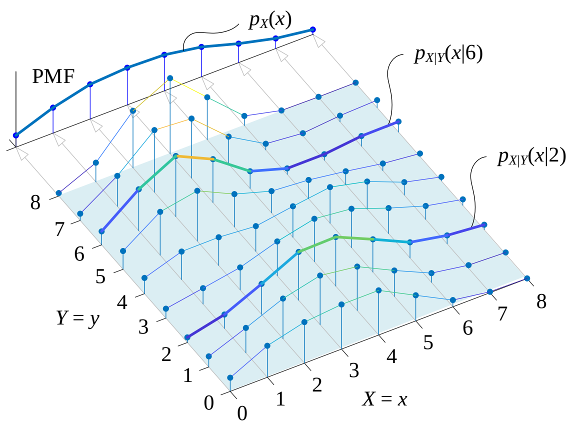


图 22. X 和 Y 不独立，条件概率 $p_{X|Y}(x|y)$ 不同于边缘概率 $p_X(x)$

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：

$$\underbrace{p_{Y|X}(y|x)}_{\text{Conditional}} = \frac{\overbrace{p_{X,Y}(x,y)}^{\text{Joint}}}{\underbrace{p_X(x)}_{\text{Marginal}}} = \frac{\overbrace{p_{X,Y}(x,y)}^{\text{Joint}}}{\sum_y p_{X,Y}(x,y)} \tag{47}$$

如图 23 所示，X 和 Y 不独立，条件概率 $p_{Y|X}(y|x)$ 不同于边缘概率 $p_Y(y)$。

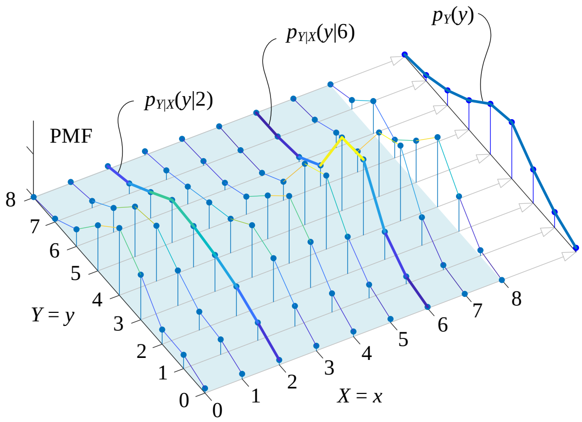


图 23. X 和 Y 不独立，条件概率 $p_{Y|X}(y|x)$ 不同于边缘概率 $p_Y(y)$

4.9 以鸢尾花数据为例：不考虑分类标签

本章下两节用鸢尾花数据集花萼长度 (X_1)、花萼宽度 (X_2)、分类标签 (Y) 样本数据为例，讲解本章前文介绍离散随机变量主要知识点。

分类标签 (Y) 本身就是离散随机变量，因为 Y 的取值只有三个，对应鸢尾花三个类别——versicolor、setosa、virginica。

而花萼长度 (X_1)、花萼宽度 (X_2) 两者取值都是连续数值，大家可能好奇，X_1 和 X_2 怎么可能变成离散随机变量？

两把直尺

这里只需要做一个很小的调整，给定鸢尾花花萼长度或宽度 d，然后进行 round($2 \times d$)/2 运算。比如，鸢尾花花萼长度为 5.3，进行上述计算变成 5.5。

这就好比，测量鸢尾花获得原始数据时，用的是图 24 (a) 所示直尺。而我们在测量花萼长度、花萼宽度时，用的是如图 24 (b) 所示的直尺。直尺精度为 0.5 cm。而测量结果仅保留一位有效小数，这一位小数的数值可能是 0 或 5。

实际上鸢尾花原始数据本身也是“离散的”，因为原始数据仅仅保留一位有效小数位。只不过我们把数据看成是连续数据而已。从这个角度来看，在数据科学领域，电子数据离散、连续与否是相对的。

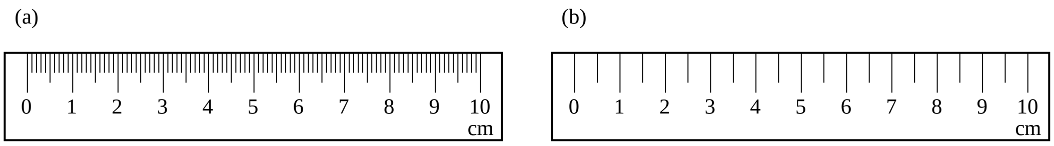


图 24. 两把直尺

“离散”的花萼长度、花萼宽度数据

图 25 所示为经过 round($2 \times d$)/2 运算得到的“离散”的花萼长度、花萼宽度数据散点图。

花萼长度 (X_1) 取值有 8 个，分别是 4.5、5.0、5.5、6.0、6.5、7.0、7.5、8.0。也就是说 X_1 的样本空间为 {4.5, 5.0, 5.5, 6.0, 6.5, 7.0, 7.5, 8.0}。

花萼宽度 (X_2) 取值有 6 个，分别是 2.0、2.5、3.0、3.5、4.0、4.5。X_2 的样本空间为 {2.0, 2.5, 3.0, 3.5, 4.0, 4.5}。

下一步，我们统计每个散点对应的频数，即散点图中网格线交点处样本数量。

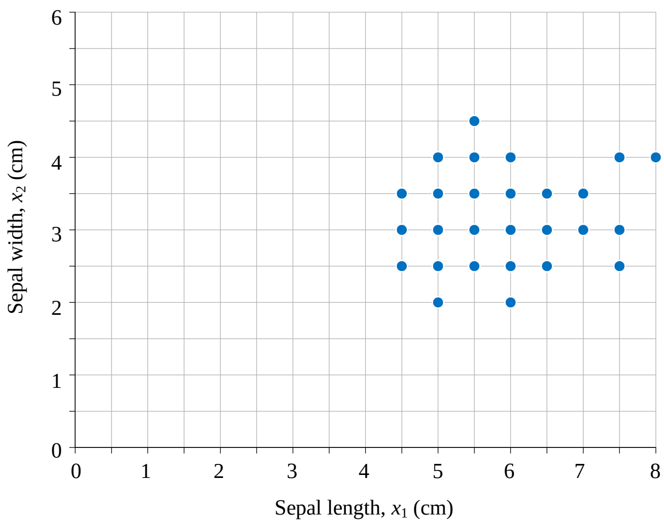


图 25. “离散”的鸢尾花花萼长度、花萼宽度散点图

频数 → 联合概率质量函数 $p_{X1,X2}(x_1,x_2)$

基于图 25 所示数据，我们可以得到图 26 所示频数和概率热图。为了区分频数和概率热图，两类热图采用不同色谱。

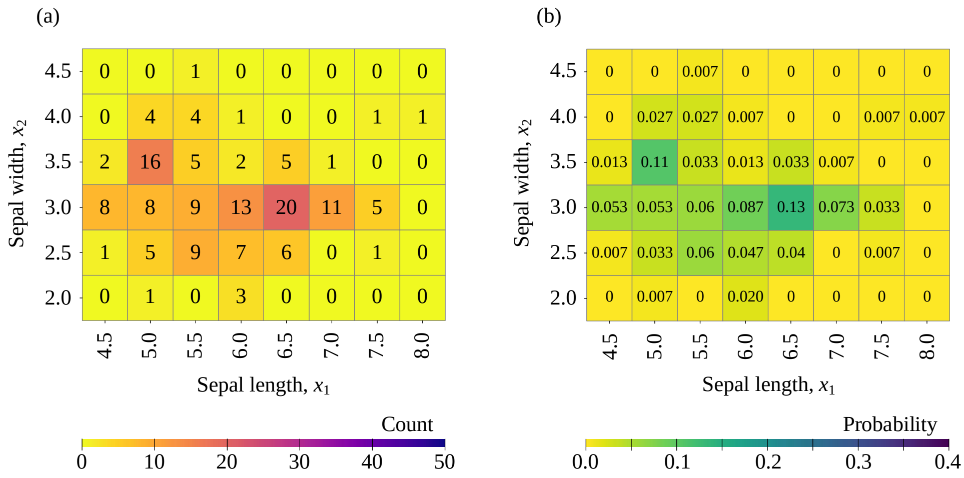


图 26. 频数和概率热图，全部样本点，不考虑分类

图 26 (a) 中频数之和为 150，即鸢尾花样本总数。从频数到概率的计算很简单，比如频数为 3，样本总数为 150，两者比值对应概率 0.02 = 3/150。

翻译成“概率语言”就是，花萼长度 (X_1) 为 6.5、花萼宽度 (X_2) 为 2.0 时，联合概率为 0.02：

$$p_{X1,X2}(6.5, 2.0) = 0.02 \tag{48}$$

采用穷举法，图 26 (b) 热图中所有取值之和为 1，即：

$$\sum_{x_1}\sum_{x_2} p_{X1,X2}(x_1, x_2) = 1 \tag{49}$$

用样本数来计算的话，上式相当于 150/150 = 1。也就是说，图 26 (b) 是对概率为 1 的某种特定的分割。

花萼长度边缘概率 $p_{X1}(x_1)$：偏求和

图 27 所示为求解花萼长度边缘概率的过程。

举个例子，当花萼长度 (X_1) 取值为 7.0 时，对应的边缘概率 $p_{X1}(7.0)$ 可以通过如下偏求和得到：

$$p_{X1}(7.0) = \sum_{x_2} p_{X1,X2}(7.0, x_2) = \underbrace{0}_{X_2=2.0} + \underbrace{0}_{X_2=2.5} + \underbrace{0.073}_{X_2=3.0} + \underbrace{0.007}_{X_2=3.5} + \underbrace{0}_{X_2=4.0} + \underbrace{0}_{X_2=4.5} = 0.08 \tag{50}$$

上式相当于，固定花萼长度 (X_1) 为 7.0，然后穷举花萼宽度 (X_2) 所有概率值，然后求和。

从频数角度来看，上式相当于：

$$p_{X1}(7.0) = \frac{\overbrace{0}^{X_2=2.0} + \overbrace{0}^{X_2=2.5} + \overbrace{11}^{X_2=3.0} + \overbrace{1}^{X_2=3.5} + \overbrace{0}^{X_2=4.0} + \overbrace{0}^{X_2=4.5}}{150} = \frac{12}{150} = 0.08 \tag{51}$$

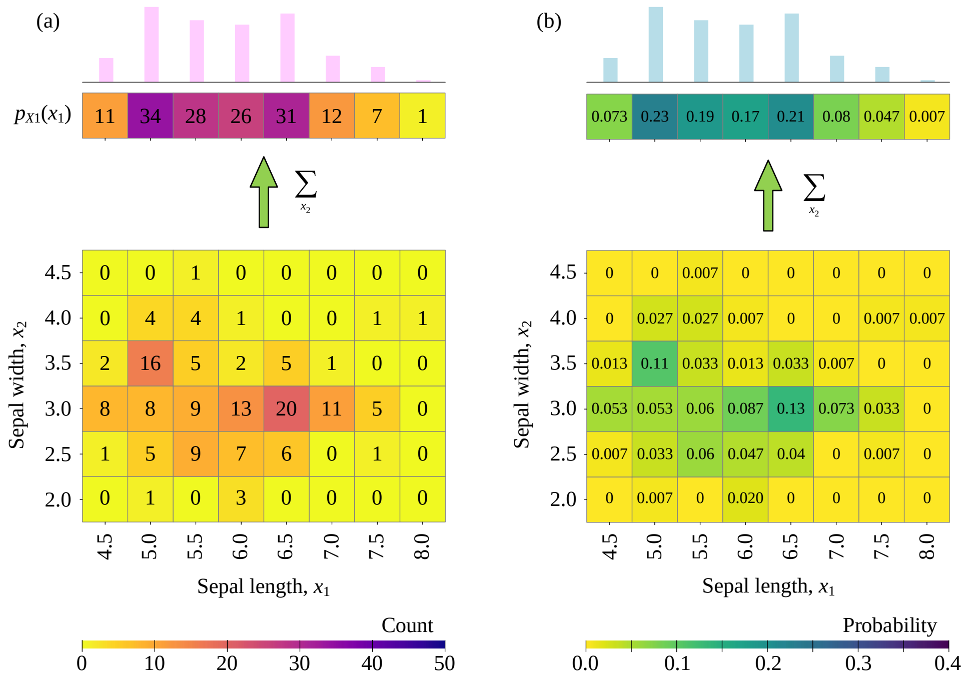


图 27. 花萼长度的边缘频数和概率热图，不考虑分类

花萼长度边缘概率 $p_{X2}(x_2)$：偏求和

图 28 所示为求解花萼宽度边缘概率的过程。

举个例子，当花萼宽度 (X_2) 取值为 2.0 时，对应的边缘概率 $p_{X2}(2.0)$ 可以通过如下偏求和得到：

$$p_{X2}(2.0) = \sum_{x_1} p_{X1,X2}(x_1, 2.0) = \underset{X_1=4.5}{0} + \underset{X_1=5.0}{0.007} + \underset{X_1=5.5}{0} + \underset{X_1=6.0}{0.02} + \underset{X_1=6.5}{0} + \underset{X_1=7.0}{0} + \underset{X_1=7.5}{0} + \underset{X_1=8.0}{0} = 0.027 \quad (52)$$

上式相当于，固定花萼长度 (X_1) 为 7.0，然后穷举花萼宽度 (X_2) 所有概率值，然后求和。

从频数角度来看，上式相当于：

$$p_{X1}(5.5) = \frac{\overset{X_2=2.0}{0} + \overset{X_2=2.5}{0} + \overset{X_2=3.0}{11} + \overset{X_2=3.5}{1} + \overset{X_2=4.0}{0} + \overset{X_2=4.5}{0}}{150} = \frac{12}{150} = 0.08 \quad (53)$$

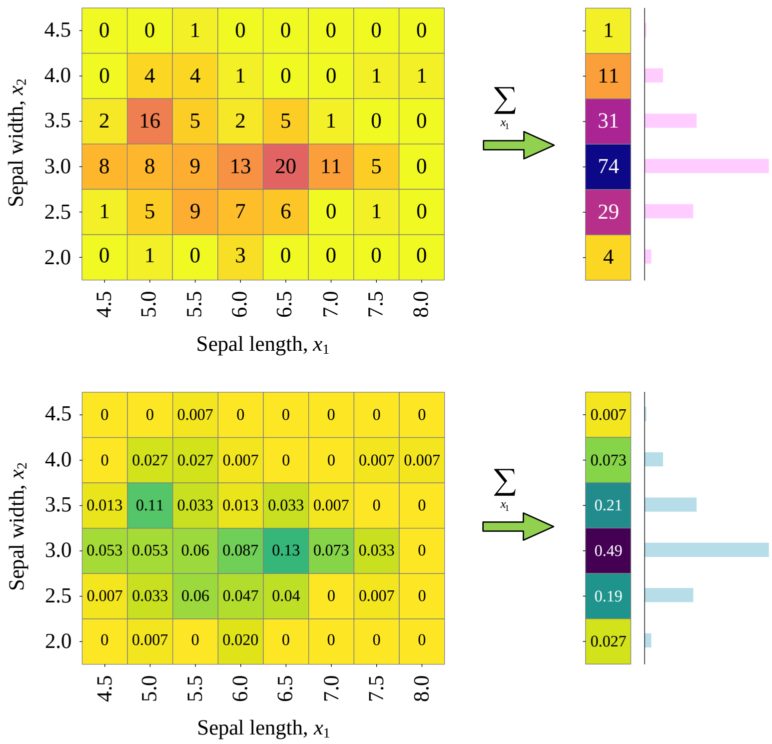


图 28. 花萼宽度的边缘频数和概率热图，不考虑分类

期望值、方差

花萼长度 X_1 的期望值：

$$
\begin{aligned}
\mathrm{E}(X_1) &= \sum_{x_1} x_1 \cdot p_{X1}(x_1) \\
&= \underset{\text{cm}}{4.5} \times 0.073 + \underset{\text{cm}}{5.0} \times 0.23 + \underset{\text{cm}}{5.5} \times 0.19 + \underset{\text{cm}}{6.0} \times 0.17 + \\
&\quad \underset{\text{cm}}{6.5} \times 0.21 + \underset{\text{cm}}{7.0} \times 0.08 + \underset{\text{cm}}{7.5} \times 0.047 + \underset{\text{cm}}{8.0} \times 0.007 \\
&= 5.836 \text{ cm}
\end{aligned} \tag{54}
$$

请大家自行写出上式对应的矩阵运算式。

然后，计算花萼长度 X_1 平方的期望值：

$$
\begin{aligned}
\mathrm{E}(X_1^2) &= \sum_{x_1} x_1^2 \cdot p_{X1}(x_1) \\
&= \underset{\text{cm}^2}{4.5^2} \times 0.073 + \underset{\text{cm}^2}{5.0^2} \times 0.23 + \underset{\text{cm}^2}{5.5^2} \times 0.19 + \underset{\text{cm}^2}{6.0^2} \times 0.17 + \\
&\quad \underset{\text{cm}^2}{6.5^2} \times 0.21 + \underset{\text{cm}^2}{7.0^2} \times 0.08 + \underset{\text{cm}^2}{7.5^2} \times 0.047 + \underset{\text{cm}^2}{8.0^2} \times 0.007 \\
&= 34.741 \text{ cm}^2
\end{aligned} \tag{55}
$$

由此可以求得花萼长度 X_1 的方差：

$$\operatorname{var}(X_1) = \underbrace{\mathrm{E}(X_1^2)}_{\text{Expectaton of } X^2} - \underbrace{\mathrm{E}(X_1)^2}_{\text{Square of } \mathrm{E}(X_1)} = 0.6749 \tag{56}$$

上式的平方根便是 X_1 的均方差：

$$\sigma_{X1} = \sqrt{\operatorname{var}(X_1)} = 0.821 \text{ cm} \tag{57}$$

请大家自行计算：花萼宽度 X_2 的期望值、X_2 平方期望值。由此，可以求得花萼宽度 X_2 的方差，然后计算 X_2 的标准差。

独立

前文提过，如果假设 X_1 和 X_2 独立，则联合概率可通过下式计算得到：

$$p_{X1,X2}(x_1, x_2) = p_{X1}(x_1) \cdot p_{X2}(x_2) \tag{58}$$

图 29 所示为，假设 X_1 和 X_2 独立，联合概率的热图。

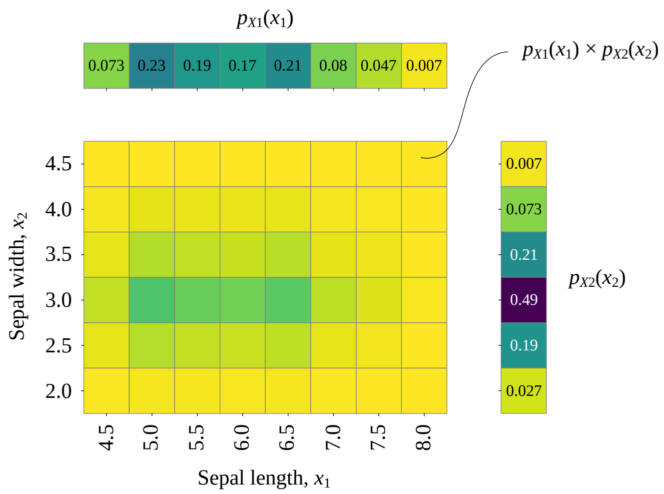


图 29. 联合概率，假设独立

这实际上就是《矩阵力量》介绍的向量张量积，也相当于如图 30 所示的矩阵乘法。

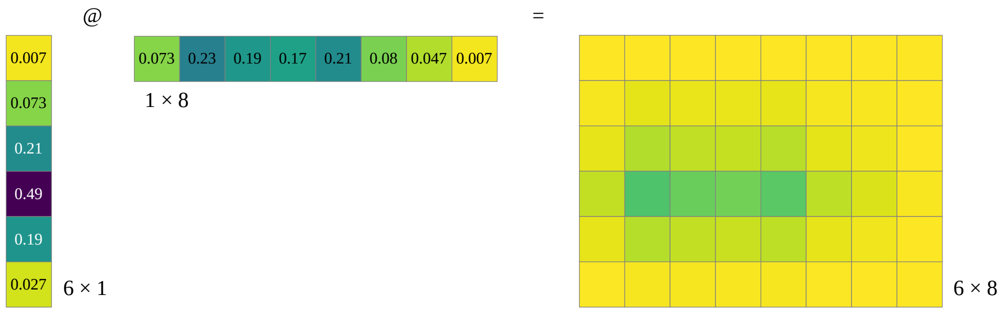


图 30. X_1 和 X_2 条件独立，矩阵乘法

图 29 中矩阵所有元素之和为 1。追根溯源，这体现的是乘法的分配律：

$$\underbrace{\sum_{x_1} p_{X1}(x_1)}_{=1} \cdot \underbrace{\sum_{x_2} p_{X2}(x_2)}_{=1} = 1 \tag{59}$$

为了配合热图形式，用如下方式展开上式：

$$\underbrace{\left\{p_{X2}(4.5)+p_{X2}(4.0)+\cdots+p_{X2}(2.0)\right\}}_{=1} \cdot \underbrace{\left\{p_{X1}(4.5)+p_{X1}(5.0)+\cdots+p_{X1}(8.0)\right\}}_{=1} = 1 \tag{60}$$

展开的每一个元素对应热图矩阵的每个元素：

$$\begin{gathered} p_{X2}(4.5)\cdot p_{X1}(4.5)+p_{X2}(4.5)\cdot p_{X1}(5.0)+\cdots+p_{X2}(4.5)\cdot p_{X1}(8.0)+ \\ p_{X2}(4.0)\cdot p_{X1}(4.5)+p_{X2}(4.0)\cdot p_{X1}(5.0)+\cdots+p_{X2}(4.0)\cdot p_{X1}(8.0)+ \\ \cdots+ \\ p_{X2}(2.0)\cdot p_{X1}(4.5)+p_{X2}(2.0)\cdot p_{X1}(5.0)+\cdots+p_{X2}(2.0)\cdot p_{X1}(8.0)=1 \end{gathered} \tag{61}$$

给定花萼长度，花萼宽度的条件概率 $p_{X2 \mid X1}(x_2 \mid x_1)$

如图 31 所示，给定花萼长度 $X_1 = 5.0$ 作为条件，这相当于在整个样本空间中，单独划出一个区域。这个区域将是新的“条件概率样本空间”，对应图 31 中的浅蓝色背景区域。

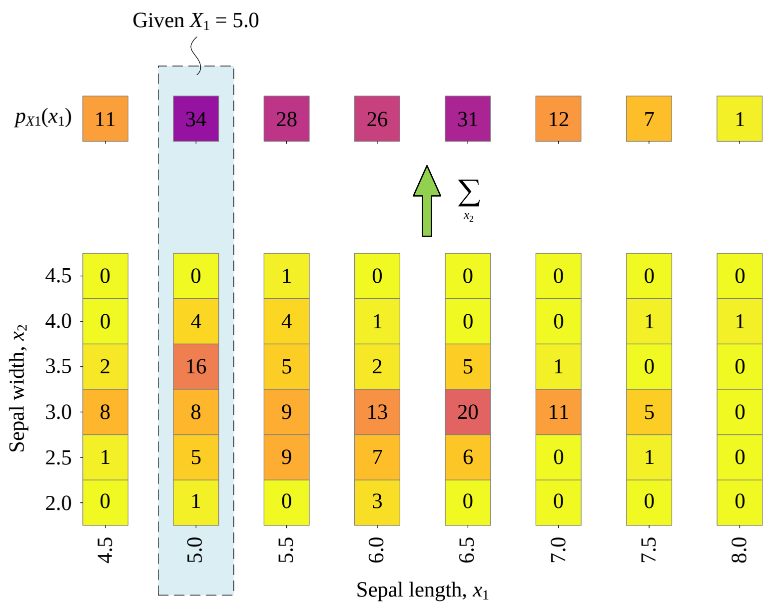


图 31. 频数视角，给定花萼长度，如何计算花萼宽度的条件概率

采用穷举法，这个区域中的条件概率有如下几个：

$$
\begin{aligned}
p_{X2|X1}(x_2=4.5 \mid x_1=5.0) &= \frac{0}{34} = 0 \\
p_{X2|X1}(x_2=4.0 \mid x_1=5.0) &= \frac{4}{34} \approx 0.12 \\
p_{X2|X1}(x_2=3.5 \mid x_1=5.0) &= \frac{16}{34} \approx 0.47 \\
p_{X2|X1}(x_2=3.0 \mid x_1=5.0) &= \frac{8}{34} \approx 0.24 \\
p_{X2|X1}(x_2=2.5 \mid x_1=5.0) &= \frac{5}{34} \approx 0.15 \\
p_{X2|X1}(x_2=2.0 \mid x_1=5.0) &= \frac{1}{34} \approx 0.029
\end{aligned} \tag{62}
$$

如图 32 所示，利用贝叶斯定理，(62) 中条件概率可以通过下式计算：

$$
\begin{aligned}
p_{X2|X1}(x_2=4.5 \mid x_1=5.0) &= \frac{p_{X1,X2}(x_1=5.0, x_2=4.5)}{p_{X1}(x_1=5.0)} \approx \frac{0}{0.23} = 0 \\
p_{X2|X1}(x_2=4.0 \mid x_1=5.0) &= \frac{p_{X1,X2}(x_1=5.0, x_2=4.0)}{p_{X1}(x_1=5.0)} \approx \frac{0.027}{0.23} \approx 0.12 \\
p_{X2|X1}(x_2=3.5 \mid x_1=5.0) &= \frac{p_{X1,X2}(x_1=5.0, x_2=3.5)}{p_{X1}(x_1=5.0)} \approx \frac{0.11}{0.23} \approx 0.47 \\
p_{X2|X1}(x_2=3.0 \mid x_1=5.0) &= \frac{p_{X1,X2}(x_1=5.0, x_2=3.0)}{p_{X1}(x_1=5.0)} \approx \frac{0.053}{0.23} \approx 0.24 \\
p_{X2|X1}(x_2=2.5 \mid x_1=5.0) &= \frac{p_{X1,X2}(x_1=5.0, x_2=2.5)}{p_{X1}(x_1=5.0)} \approx \frac{0.033}{0.23} \approx 0.15 \\
p_{X2|X1}(x_2=2.0 \mid x_1=5.0) &= \frac{p_{X1,X2}(x_1=5.0, x_2=2.0)}{p_{X1}(x_1=5.0)} \approx \frac{0.007}{0.23} \approx 0.029
\end{aligned} \tag{63}
$$

其中，

$$
\begin{aligned}
p_{X1}(x_1=5.0) = {} & p_{X1,X2}(x_1=5.0, x_2=4.5) + p_{X1,X2}(x_1=5.0, x_2=4.0) + \\
& p_{X1,X2}(x_1=5.0, x_2=3.5) + p_{X1,X2}(x_1=5.0, x_2=3.0) + \\
& p_{X1,X2}(x_1=5.0, x_2=2.5) + p_{X1,X2}(x_1=5.0, x_2=2.0) \\
& \approx 0 + 0.027 + 0.11 + 0.053 + 0.033 + 0.007 \approx 0.23
\end{aligned} \tag{64}
$$

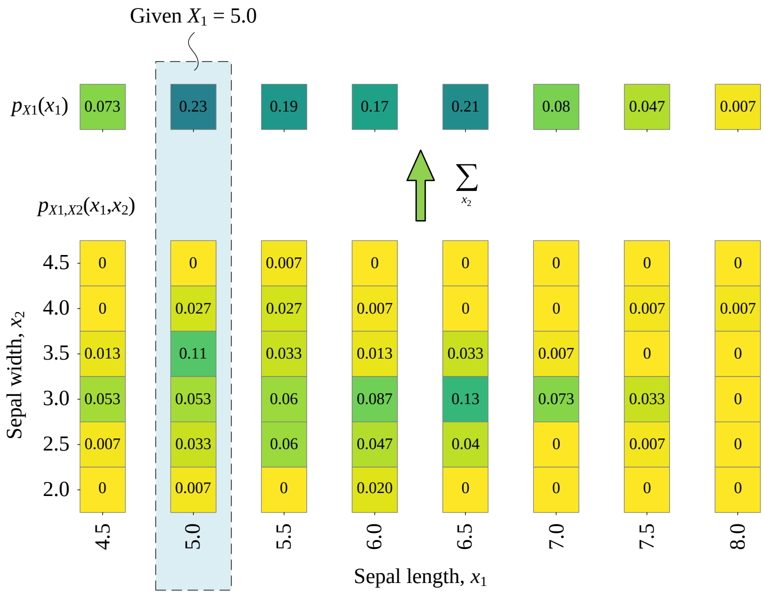


图 32. 概率视角，给定花萼长度，如何计算花萼宽度的条件概率

本章前文提过，从函数角度来看，$p_{X2\mid X1}(x_2 \mid x_1)$ 本质上也是个二元离散函数，具体如图 36 所示。

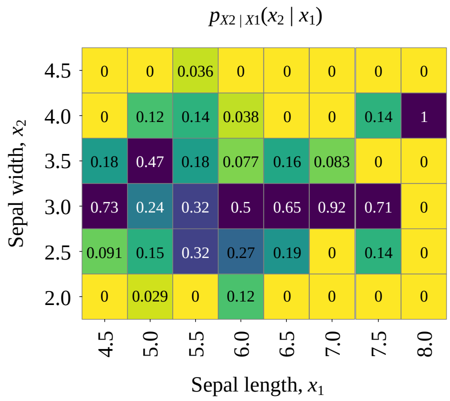


图 33. 给定花萼长度，花萼宽度的条件概率 $p_{X2\mid X1}(x_2 \mid x_1)$

如图 35 所示，每一列条件概率求和为 1：

$$\sum_{x_2} p_{X2|X1}\left(x_2 \mid x_1\right) = 1 \tag{65}$$

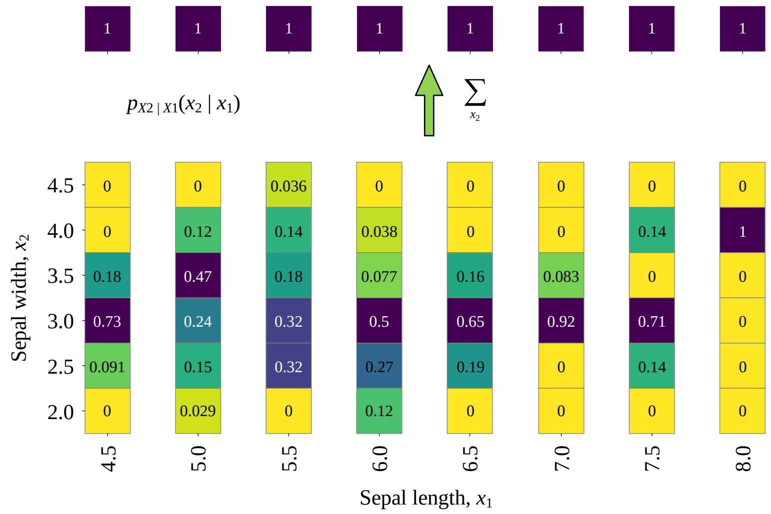


图 34. 给定花萼长度，花萼宽度的条件概率，每一列条件概率求和为 1

给定花萼宽度，花萼长度的条件概率 $p_{X1 \mid X2}(x_1 \mid x_2)$

请大家自行计算，给定花萼宽度为 3.0，每个条件概率的具体值。

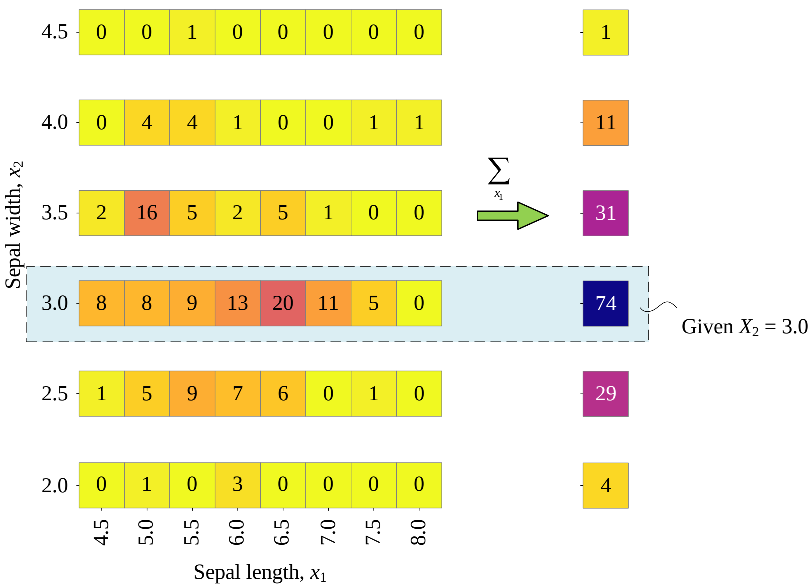


图 35. 频数视角，给定花萼宽度，如何计算花萼长度的条件概率

从函数角度来看，$p_{X1 \mid X2}(x_1 \mid x_2)$ 也是个二元离散函数，具体如图 36 所示。

大家是否立刻想到，既然我们可以求得花萼长度的期望值，我们是否可以求得给定花萼宽度条件下的花萼长度的期望、方差？

答案是肯定的！本书第 8 章将介绍条件期望 (conditional expectation)、条件方差 (conditional variance)。

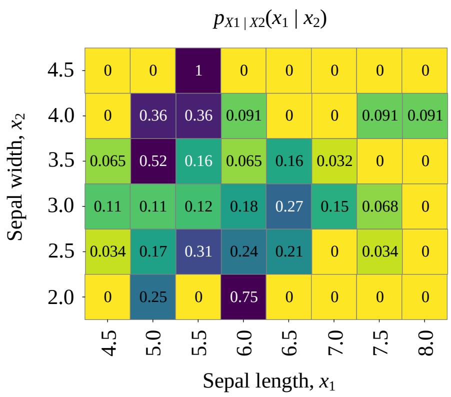


图 36. 给定花萼宽度，花萼长度的条件概率 $p_{X1 | X2}(x_1 | x_2)$

如图 37 所示，每一行条件概率求和为 1：

$$\sum_{x_1} p_{X1|X2}\left(x_1 | x_2\right) = 1 \tag{66}$$

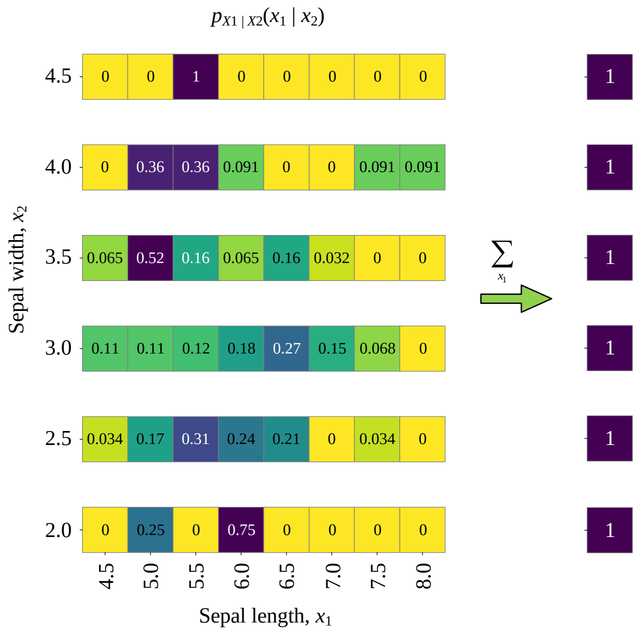


图 37. 给定花萼宽度，花萼长度的条件概率每一行条件概率求和为 1

4.10 以鸢尾花数据为例：考虑分类标签

本节讨论在考虑分类标签条件下，如何计算鸢尾花数据的条件概率。

给定分类标签 $Y = C_1$ (setosa)

图 38 (a) 所示为给定分类标签 $Y = C_1$ (setosa) 条件下，鸢尾花数据集中 50 个样本数据的频数热图。图 38 中频数除以 50 便得到图 38 (b) 所示条件概率 $p_{X1,X2 \mid Y}(x_1, x_2 \mid y = C_1)$ 热图。

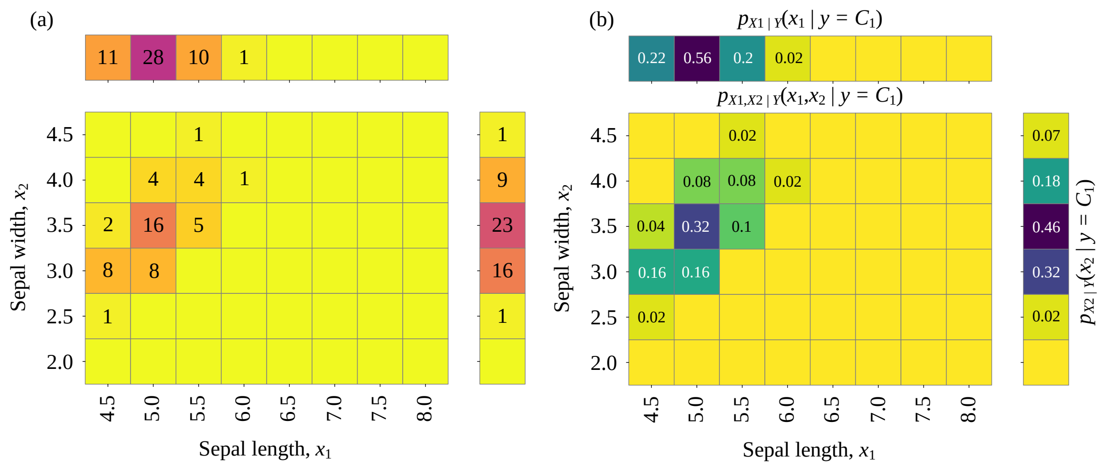


图 38. 频数和条件概率 $p_{X1,X2 \mid Y}(x_1, x_2 \mid y = C_1)$ 热图，给定分类标签 $Y = C_1$ (setosa)

此外，请大家根据频数热图，自行计算两个条件概率：$p_{X1 \mid X2, Y}(x_1 = 5.0 \mid x_2 = 3.0, y = C_1)$ 和 $p_{X2 \mid X1, Y}(x_2 = 3.0 \mid x_1 = 5.0, y = C_1)$。

给定分类标签 $Y = C_2$ (versicolor)

图 39 (a) 所示为给定分类标签 $Y = C_2$ (versicolor) 条件下，鸢尾花数据集中 50 个样本数据的频数热图。图 39 中频数除以 50 便得到图 39 (b) 所示条件概率 $p_{X1,X2 \mid Y}(x_1, x_2 \mid y = C_2)$ 热图。

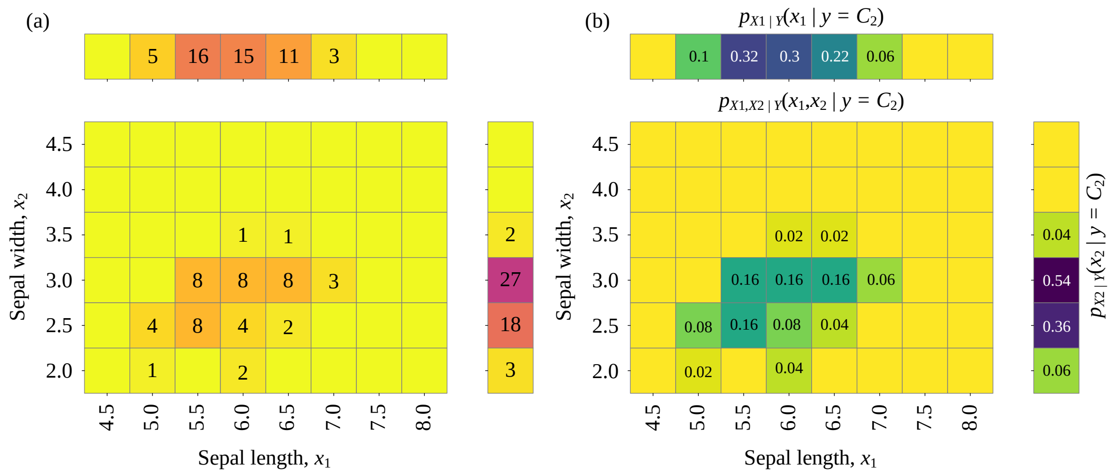


图 39. 频数和条件概率 $p_{X1,X2 \mid Y}(x_1, x_2 \mid y = C_2)$ 热图，给定分类标签 $Y = C_2$ (versicolor)

给定分类标签 $Y = C_3$ (virginica)

请大家自行分析图 40。

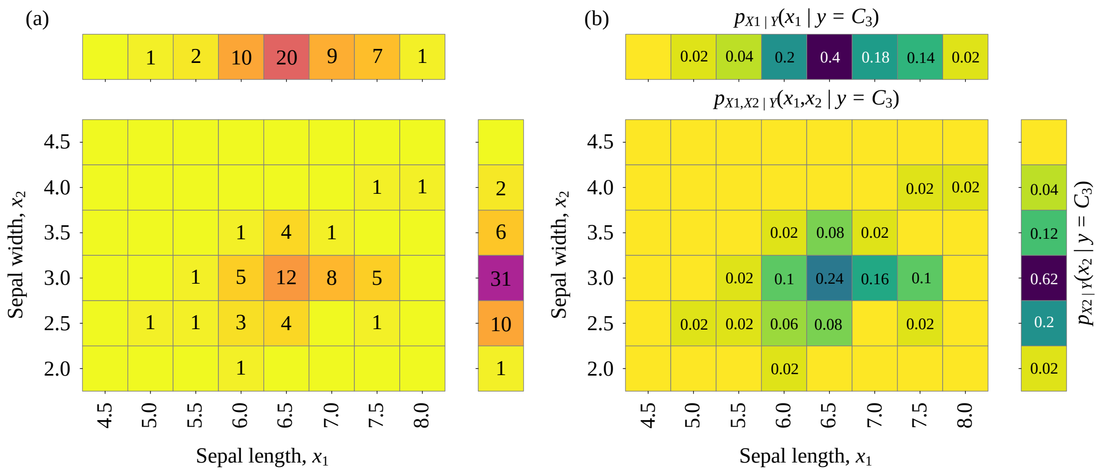


图 40. 频数和条件概率 $p_{X1,X2\,|\,Y}(x_1, x_2 \,|\, y = C_3)$ 热图，给定分类标签 $Y = C_3$ (virginica)

全概率

如图 41 所示，利用全概率定理，我们可以通过下式计算 $p_{X1,X2}(x_1, x_2)$:

$$
\begin{aligned}
p_{X1,X2}\left(x_1,x_2\right) &= \sum_y \underbrace{p_{X1,X2,Y}\left(x_1,x_2,y\right)}_{\text{Joint}} \\
&= \sum_y \underbrace{p_{X1,X2|Y}\left(x_1,x_2 \mid y\right)}_{\text{Conditional}} \cdot \underbrace{p_Y\left(y\right)}_{\text{Marginal}} \\
&= p_{X1,X2|Y}\left(x_1,x_2 \mid C_1\right) \cdot p_Y\left(C_1\right) + \\
&\quad p_{X1,X2|Y}\left(x_1,x_2 \mid C_2\right) \cdot p_Y\left(C_2\right) + \\
&\quad p_{X1,X2|Y}\left(x_1,x_2 \mid C_3\right) \cdot p_Y\left(C_3\right)
\end{aligned}
\tag{67}
$$

从几何角度来看，联合概率质量函数 $p_{X1,X2}(x_1, x_2, y)$ 相当于一个“立方体”。上式相当于，将立方体在 Y 方向上压扁成 $p_{X1,X2}(x_1, x_2)$ 平面。本章最后将继续这一话题。

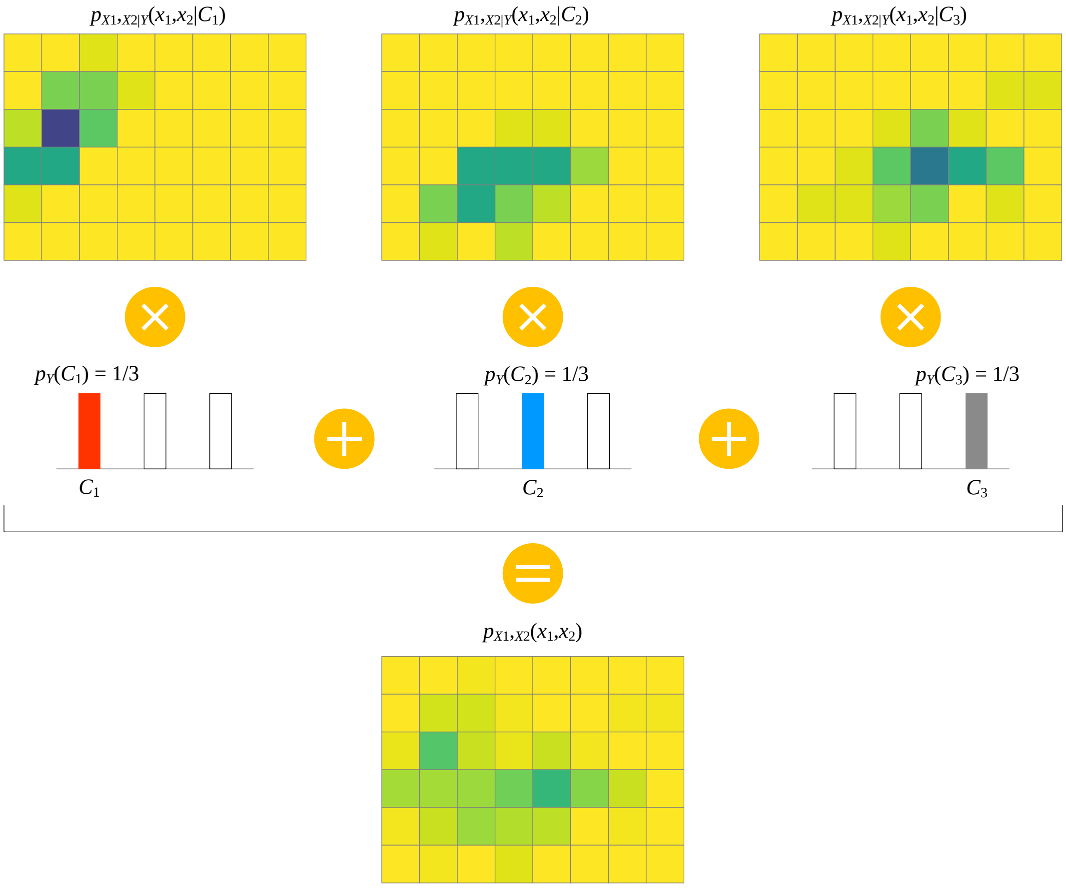


图 41. 利用全概率定理，计算 $p_{X1,X2}(x_1, x_2)$

条件独立

图 42 所示为给定 $Y = C_1$ 条件下，假设 X_1 和 X_2 条件独立，利用 $p_{X1\mid Y}(x_1 \mid y = C_1)$、$p_{X2\mid Y}(x_2 \mid y = C_1)$ 估算 $p_{X1,X2\mid Y}(x_1,x_2 \mid y = C_1)$、请大家自行分析图 43、图 44。将这些条件概率质量函数代入 (67)，我们也可以计算得到另外一个 $p_{X1,X2}(x_1, x_2)$。这实际上是估算 $p_{X1,X2}(x_1, x_2)$ 的一种方法。本书后续还会介绍这种方法及其应用。

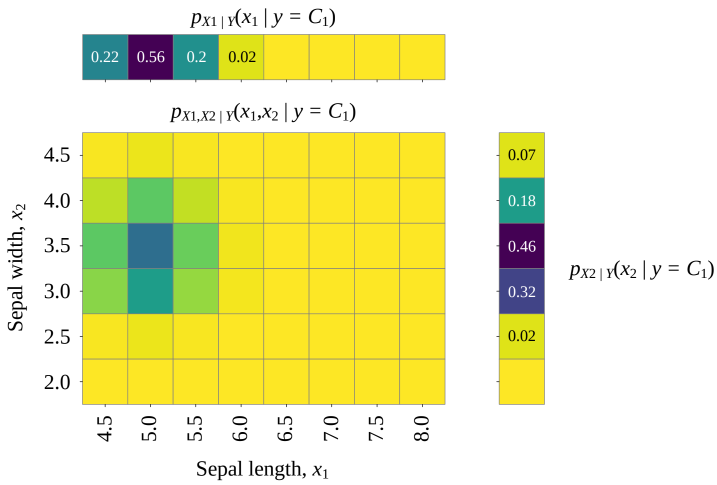


图 42. 给定 $Y = C_1$，假设 X_1 和 X_2 条件独立，计算 $p_{X1,X2|Y}(x_1,x_2 | y = C_1)$

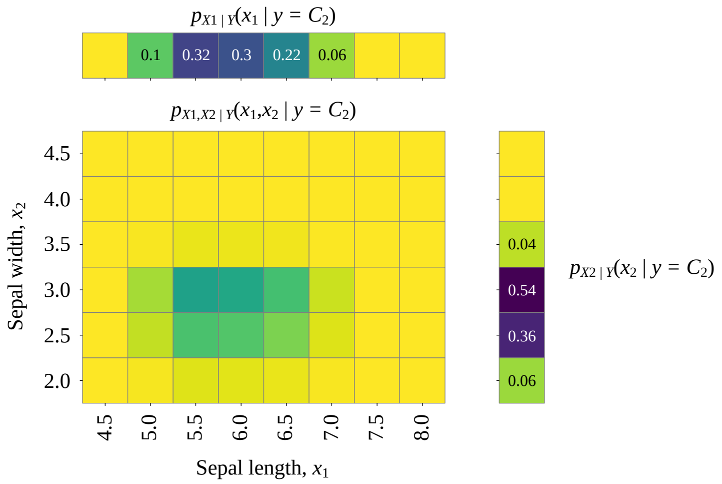


图 43. 给定 $Y = C_2$，假设 X_1 和 X_2 条件独立，计算 $p_{X1,X2|Y}(x_1,x_2 | y = C_2)$

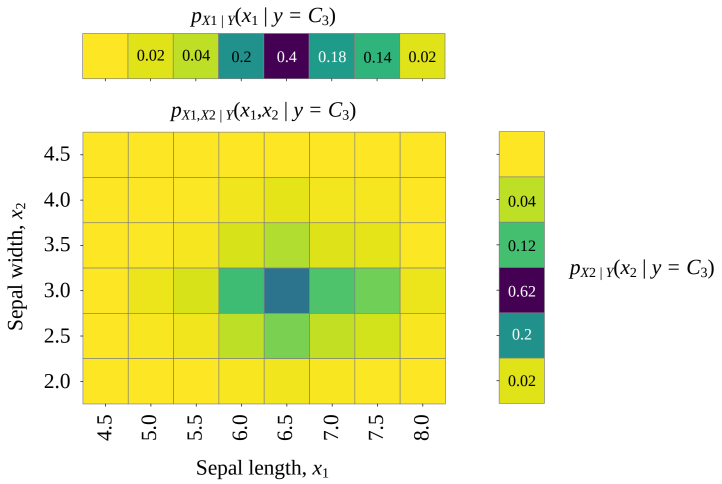


图 44. 给定 $Y = C_3$，假设 X_1 和 X_2 条件独立，计算 $p_{X1,X2|Y}(x_1,x_2 | y = C_3)$

代码 Bk5_Ch04_02.py 绘制前两节大部分图像。

4.10 再谈概率 1：展开、折叠

偏求和：压扁

本章前文提到，几何上，$p_{X1,X2,X3}(x_1, x_2, x_3)$ 代表一个三维立方体。而偏求和是个降维过程，看上去把立方体在不同维度上压扁。

如图 45 所示，$p_{X1,X2,X3}(x_1, x_2, x_3)$ 在 x_1 上偏求和，压扁得到 $p_{X2,X3}(x_2, x_3)$：

$$p_{X2,X3}(x_2, x_3) = \sum_{x_1} p_{X1,X2,X3}(x_1, x_2, x_3) \tag{68}$$

如图 45 所示，$p_{X2,X3}(x_2, x_3)$ 代表一个二维平面，相当于一个矩阵。而 $p_{X2,X3}(x_2, x_3)$ 进一步沿着 x_2 折叠便得到边缘概率质量函数 $p_{X3}(x_3)$：

$$\begin{aligned} p_{X3}(x_3) &= \sum_{x_2} p_{X2,X3}(x_2, x_3) \\ &= \sum_{x_2}\sum_{x_1} p_{X1,X2,X3}(x_1, x_2, x_3) \end{aligned} \tag{69}$$

而 $p_{X3}(x_3)$ 相当于一个向量。沿着哪个方向求和，就相当于完成了这个维度上数据的合并。这个维度因此便消失。

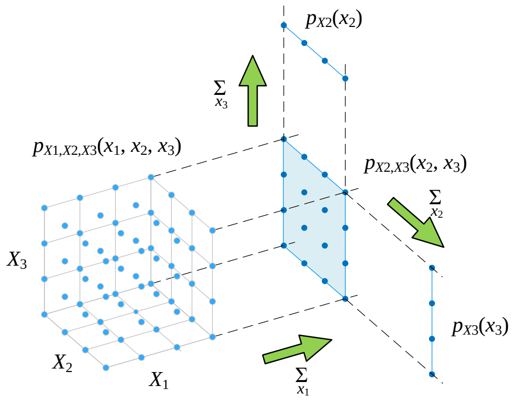


图 45. 先沿 X_1 方向压扁

换个方向，$p_{X2,X3}(x_2, x_3)$ 沿着 x_3 折叠便得到边缘概率质量函数 $p_{X2}(x_2)$：

$$\begin{aligned} p_{X2}(x_2) &= \sum_{x_3} p_{X2,X3}(x_2, x_3) \\ &= \sum_{x_3}\sum_{x_1} p_{X1,X2,X3}(x_1, x_2, x_3) \end{aligned} \tag{70}$$

而 $p_{X3}(x_3)$ 和 $p_{X2}(x_2)$ 进一步折叠，便获得概率 1：

$$1 = \sum_{x_3}\sum_{x_2}\sum_{x_1} p_{X1,X2,X3}(x_1, x_2, x_3) = \sum_{x_2}\sum_{x_3}\sum_{x_1} p_{X1,X2,X3}(x_1, x_2, x_3) \tag{71}$$

经过上述不同顺序的三重求和后，三个维度全部消失。

请大家沿着上述思路自行分析图 46 两幅图，并写出求和公式。请大家自己思考，如果 X_1、X_2、X_3 独立，如何计算 $p_{X1,X2,X3}(x_1, x_2, x_3)$？

此外，本节 X_1、X_2、X_3 均为离散随机变量，因此图 45 中每个点均代表概率值。如果 X_1、X_2、X_3 均为连续随机变量，图 45 这个立方体该是怎样的形式？要是 X_1、X_2 为连续随机变量，但是 X_3 为离散随机变量，图 45 这个立方体又是怎样的形式？

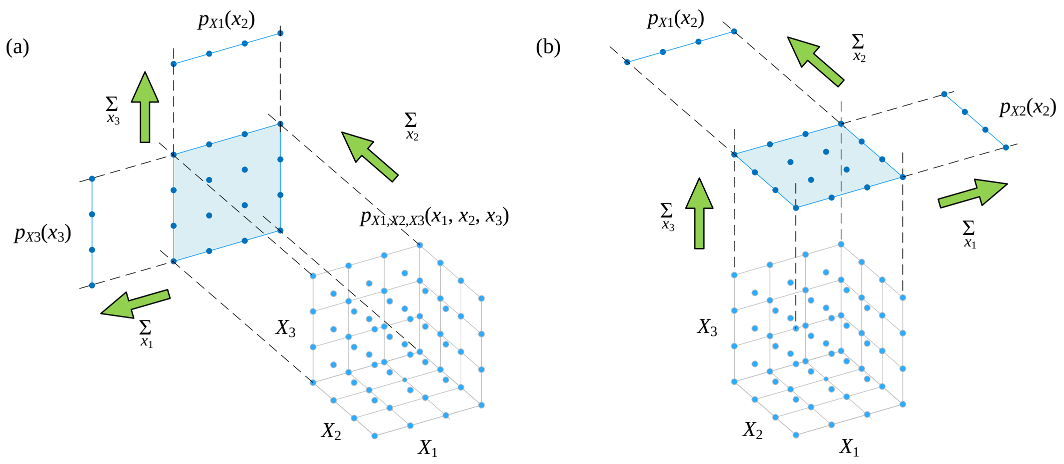


图 46. 分别先沿 X_2、X_3 方向压扁

条件概率：切片

如图 47 所示，条件概率 $p_{X1,X2\mid X3}(x_1, x_2\mid c)$ 相当于在 $X_3 = c$ 处切了一片，只考虑切片上的概率分布情况，而不考虑整个立方体的概率分布。也就是说 $X_3 = c$ 对应的切片是条件概率 $p_{X1,X2\mid X3}(x_1, x_2\mid c)$ 的样本空间。

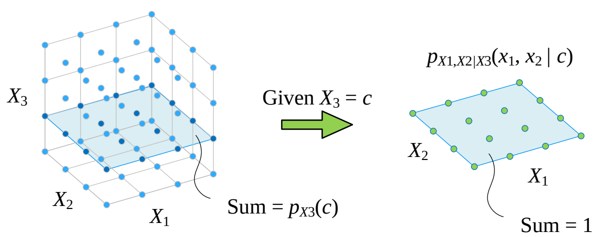


图 47. 给定 $X_3 = c$ 条件概率

首先将切片上的联合概率求和得到 $p_{X3}(c)$：

$$p_{X3}(c)=\sum_{x_2}\sum_{x_1} p_{X1,X2,X3}(x_1,x_2,c) \tag{72}$$

然后，用联合概率除以 $p_{X3}(c)$ 得到条件概率 $p_{X1,X2\mid X3}(x_1, x_2\mid c)$：

$$p_{X1,X2|X3}(x_1,x_2\mid c)=\frac{p_{X1,X2,X3}(x_1,x_2,c)}{p_{X3}(c)} \tag{73}$$

大家自己思考，如果给定 $X_3 = c$ 条件下，X_1 和 X_2 条件独立，意味着什么？