8 条件概率

Conditional Expectation and Variance

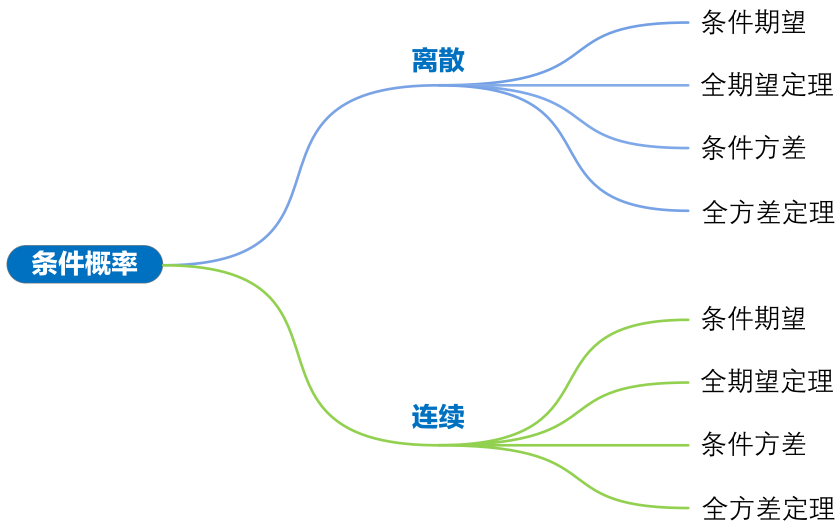
离散、连续随机变量的条件期望、条件方差

每一种科学，只要达到一定程度的成熟，就会自动成为数学的一部分。

Every kind of science, if it has only reached a certain degree of maturity, automatically becomes a part of mathematics.

—— 大卫·希尔伯特 (David Hilbert) | 德国数学家 | 1862 ~ 1943

- `matplotlib.pyplot.errorbar()` 绘制误差棒
- `matplotlib.pyplot.stem()` 绘制火柴梗图
- `numpy.mean()` 计算均值
- `numpy.sqrt()` 计算平方根
- `numpy.std()` 计算标准差，默认分母为 n，不是 n - 1
- `numpy.var()` 计算方差，默认分母为 n，不是 n - 1
- `seaborn.heatmap()` 绘制热图


条件概率
离散
条件期望
全期望定理
条件方差
全方差定理
连续
条件期望
全期望定理
条件方差
全方差定理

8.1 离散随机变量：条件期望

条件**期望** (conditional expectation 或 conditional expected value)，或条件**均值** (conditional mean)，是一个随机变量的相对于一个条件概率分布的**期望**。换句话说，这是给定的一个或多个其他随机变量值的条件下，某个特定随机变量的**期望**。

类似地，条件**方差** (conditional variance) 与一般**方差**的定义几乎一致。计算条件**方差**时，只不过将**期望**换成了条件**期望**，并将概率换成了条件概率而已。

条件**期望**和条件**方差**这两个概念在数据科学、机器学习算法中格外重要，本章分别讲解离散随机变量和随机变量的条件**期望**和条件**方差**。

本书第 12 章则专门介绍高斯条件概率。

大家应该已经看到，本章**期望**、**方差**交替出现，为了帮助大家阅读，我们用给**期望**、**方差**涂了不同颜色。

什么是条件期望?

条件**期望**其实很好理解。比如，一个笼子里 10 只动物，其中 6 只鸡 (60%)、4 只兔 (40%)。如图 1 所示，分别只考虑鸡，只考虑兔，这就是“条件”。

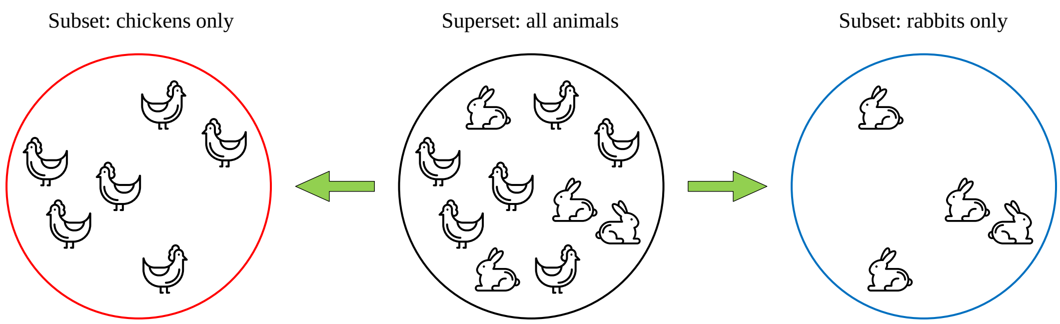


图 1. 解释条件

如图 2 所示，鸡的平均体重为 2 公斤，这个数值就是条件**期望**。再举个例子，兔子的平均体重为 4 公斤，这也是条件**期望**。

本书后续会用鸢尾花数据做例子给大家继续讲解条件**期望**。

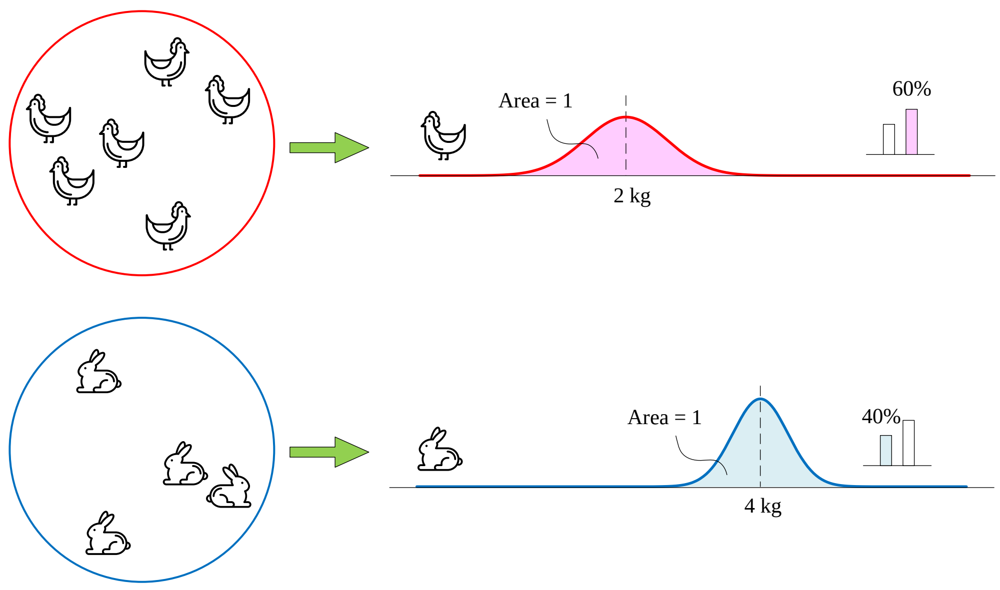


图 2. 解释条件**期望**

条件**期望** $\mathrm{E}(Y|X = x)$

如果 X 和 Y 均为离散随机变量，给定 $X = x$ 条件下，Y 的条件**期望** $\mathrm{E}(Y|X = x)$ (conditional mean of Y given $X = x$) 定义为：

$$
\begin{aligned}
\mathrm{E}\left(Y \,\middle|\, \underbrace{X = x}_{\text{Given}}\right) &= \overbrace{\sum_{y} y \cdot \underbrace{p_{Y|X}\left(y|x\right)}_{\text{Conditional}}}^{\text{Expectation}} \\
&= \sum_{y} y \cdot \frac{\overbrace{p_{X,Y}\left(x,y\right)}^{\text{Joint}}}{\underbrace{p_{X}\left(x\right)}_{\text{Marginal}}} = \frac{1}{\underbrace{p_{X}\left(x\right)}_{\text{Marginal}}} \sum_{y} y \cdot \overbrace{p_{X,Y}\left(x,y\right)}^{\text{Joint}}
\end{aligned}
\tag{1}
$$

(1) 相当于求加权平均数。

从几何角度来看，如图 3 所示，条件概率质量函数 $p_{Y|X}\left(y|x\right)$ 分别乘以对应 y 值 (黄色高亮)，然后求和，结果就是条件**期望** $\mathrm{E}(Y|X = x)$。

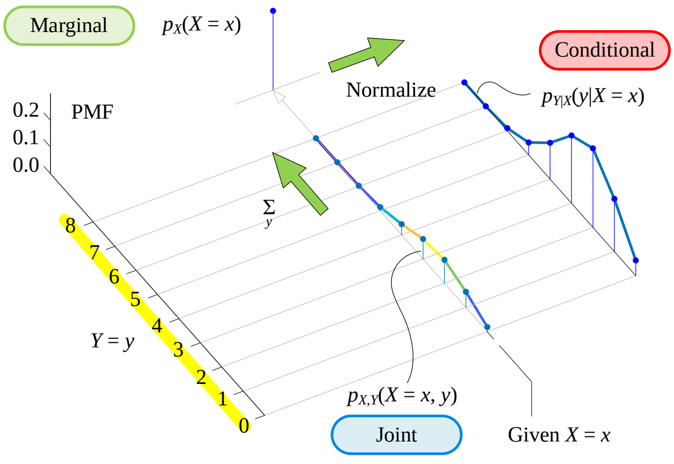


图 3. 条件概率 PMF $p_{Y|X}(y|x)$，X 和 Y 均为离散随机变量

解剖条件期望 $\mathrm{E}(Y|X = x)$

下面，我们进一步解剖 (1)。

给定 $X = x$ 条件下，也就是说离散随机变量 X 固定在 x，满足这个条件的样本构成了全新的“样本空间”。

$p_{Y|X}\left(y|x\right)$ 是给定 $X = x$ 条件下 Y 的概率质量函数，相当于 (1) 中加权平均数中的权重。

回忆本书第 4 章，利用贝叶斯定理，$p_X\left(x\right) > 0$，条件概率质量函数 $p_{Y|X}\left(y|x\right)$ 可以通过联合 PMF $p_{X,Y}\left(x, y\right)$ 和边缘 PMF $p_X\left(x\right)$ 相除得到：

$$p_{Y|X}\left(y|x\right) = \frac{p_{X,Y}\left(x,y\right)}{\underbrace{p_X\left(x\right)}_{\text{Normalize}}} \tag{2}$$

其中，分母中的边缘概率 $p_X\left(x\right)$ 起到归一化的效果。

(1) 中大西格玛求和 $\sum_y(\cdot)$ 代表“穷举”一切可能的 y 值，计算“$y \times$ 条件概率 $p_{Y|X}\left(y|x\right)$”之和，也就是“$y \times$ 权重”之和，即加权平均数。

比较期望 $\mathrm{E}(Y)$、条件期望 $\mathrm{E}(Y|X = x)$

对比离散随机变量 Y 的期望 $\mathrm{E}(Y)$、条件期望 $\mathrm{E}(Y|X = x)$：

$$\begin{aligned} \mathrm{E}\left(Y\right) &= \sum_y y \cdot \underbrace{p_Y\left(y\right)}_{\text{Weight}} \\ \mathrm{E}\left(Y \mid X = x\right) &= \sum_y y \cdot \underbrace{p_{Y|X}\left(y \mid x\right)}_{\text{Weight}} \end{aligned} \tag{3}$$

容易发现，我们不过是把求均值的权重从边缘 PMF $p_Y(y)$ 换成了条件 PMF $p_{Y|X}(y|x)$。$\sum_y y$ 都是遍历所有 y 的取值。

作为权重，$p_Y(y)$ 和 $p_{Y|X}(y|x)$ 的求和都为 1，即：

$$
\begin{aligned}
\sum_y \underbrace{p_Y(y)}_{\text{Marginal}} = 1 \\
\sum_y \underbrace{p_{Y|X}(y|x)}_{\text{Conditional}} = 1
\end{aligned}
\tag{4}
$$

上两式实际上都是本书第 3 章介绍的**全概率定理** (law of total probability) 的体现。

⚠ 注意，**期望** E(Y) 是一个标量值。而 $\mathrm{E}(Y|X = x)$ 在不同的 $X = x$ 条件下结果不同，即 $\mathrm{E}(Y|X)$ 代表一组数。也就是说，$\mathrm{E}(Y|X)$ 可以看做是个向量。本书前文提过，求期望 E() 运算相当于“归纳”，降维。也就是说 $\mathrm{E}(Y|X)$ 中“Y”已经被“压缩”成了一个数值，但是 X 还是可变的。

既然 $\mathrm{E}(Y|X)$ 代表一组数，我们立刻就会想到 $\mathrm{E}(Y|X)$ 肯定也有**期望**，即均值！

也就是说，笼子里的鸡的平均体重、兔子的平均体重，这两个均值还能再算一个均值，即笼子里所有动物的平均体重。

全期望定理

全**期望**定理 (law of total expectation)，又叫**双重期望**定理 (double expectation theorem)、重叠**期望**定理 (iterated total expectation)，具体指的是：

$$
\underbrace{\mathrm{E}(Y)}_{\text{Expectation}} = \overbrace{\mathrm{E}\left[\underbrace{\mathrm{E}(Y|X)}_{\text{Conditional expectation}}\right]}^{\text{Expectation of conditonal expectations}} = \sum_x \underbrace{\mathrm{E}(Y|X=x)}_{\text{Conditional expectation}} \cdot \underbrace{p_X(x)}_{\text{Marginal}}
\tag{5}
$$

推导过程如下，不要求大家记忆：

$$
\begin{aligned}
\mathrm{E}\left[\underbrace{\mathrm{E}(Y|X)}_{\text{Conditional expectation}}\right] &= \sum_x \underbrace{\mathrm{E}(Y|X=x)}_{\text{Conditional expectation}} \cdot \underbrace{p_X(x)}_{\text{Marginal}} = \sum_x \underbrace{\left\{\sum_y y \cdot \underbrace{p_{Y|X}(y|x)}_{\text{Conditional}}\right\}}_{\text{Conditional expectation}} \cdot \underbrace{p_X(x)}_{\text{Marginal}} \\
&= \sum_x \sum_y y \cdot \overbrace{\underbrace{p_{Y|X}(y|x)}_{\text{Conditional}} \cdot \underbrace{p_X(x)}_{\text{Marginal}}}^{\text{Use Bayes' Rule}} = \sum_x \sum_y y \cdot \overbrace{p_{X,Y}(x,y)}^{\text{Joint}} \\
&= \sum_x \sum_y y \cdot \overbrace{\underbrace{p_{X|Y}(x|y)}_{\text{Conditional}} \cdot \underbrace{p_Y(y)}_{\text{Marginal}}}^{\text{Use Bayes' Rule}} = \sum_y y \cdot \underbrace{p_Y(y)}_{\text{Marginal}} \cdot \underbrace{\overbrace{\sum_x p_{X|Y}(x|y)}^{=1}}_{\text{Law of total probability}} \\
&= \sum_y y \cdot \underbrace{p_Y(y)}_{\text{Marginal}} = \mathrm{E}(Y)
\end{aligned}
\tag{6}
$$

⚠ 注意，以上推导中，二重求和调换变量顺序，这是因为 x 和 y 构成的网格“方方正正”；否则，不能轻易调换求和顺序。这和调换二重积分变量顺序类似。

《数学要素》第 14 章探讨过这个问题，请大家回顾。

其实，全**期望**定理很好理解！

还是用本章前文的例子。前文提到，笼子里的鸡 (60%) 的平均体重 2 kg，兔子 (40%) 的平均体重为 4 kg。整个笼子里所有动物的平均体重就是 2 × 60% + 4 × 40% = 2.8 kg。

前文提过，2 kg、4 kg 都是都是条件**期望**。

2.8 kg 就是“条件**期望**的**期望**”。笼子里的鸡占比较高，因此整个笼子里动物的平均体重稍微“偏向”鸡体重的“条件**期望**”。

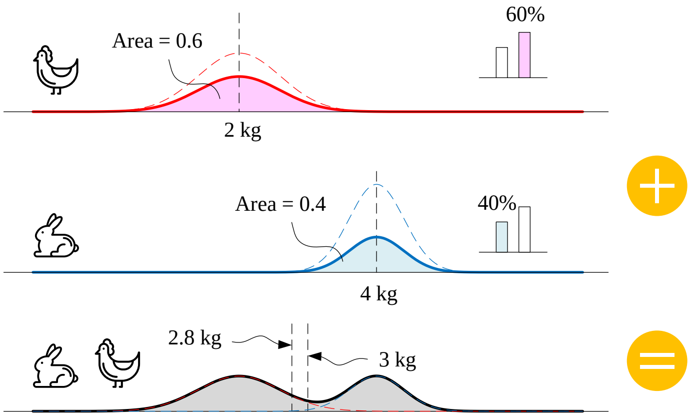


图 4. 解释全**期望**定理

大家如果要问，为什么求“条件**期望**的**期望**”要用加权平均？而不是用 (2 + 4) / 2 = 3 kg？

为了回答这个问题，我们举个极端例子来解释。除了所有鸡之外，如果整个笼子里只有一只兔子，它的体重为 8 kg，也就是说“所有”兔子的平均体重也是 8 kg。假设所有鸡的平均体重还是 2 kg。大家自己思考，如果用 2 kg 和 8 kg 的平均值 5 kg 代表整个笼子里所有动物的平均体重，这是否合理？

条件**期望** $\mathrm{E}(X|Y = y)$

同理，如图 5 所示，给定 $Y = y$ 这个条件下，$p_Y(y) > 0$，X 的条件**期望** $\mathrm{E}(X|Y = y)$ 定义为：

$$\begin{aligned}\mathrm{E}\left(x \,\Big|\, \underbrace{Y=y}_{\text{Given}}\right) &= \overbrace{\sum_{x} x \cdot \underbrace{p_{X|Y}(x|y)}_{\text{Conditional}}}^{\text{Expectation}} \\ &= \sum_{x} x \cdot \frac{\overbrace{p_{X,Y}(x,y)}^{\text{Joint}}}{\underbrace{p_Y(y)}_{\text{Marginal}}} = \frac{1}{\underbrace{p_Y(y)}_{\text{Marginal}}} \sum_{x} x \cdot \overbrace{p_{X,Y}(x,y)}^{\text{Joint}}\end{aligned} \tag{7}$$

请大家自行分析上式，并比较 $\mathrm{E}(X)$ 和 $\mathrm{E}(X|Y=y)$：

$$\begin{aligned}\mathrm{E}(X) &= \sum_{x} x \cdot \underbrace{p_X(x)}_{\text{Weight}} \\ \mathrm{E}(X \mid Y=y) &= \sum_{x} x \cdot \underbrace{p_{X|Y}(x \mid y)}_{\text{Weight}}\end{aligned} \tag{8}$$

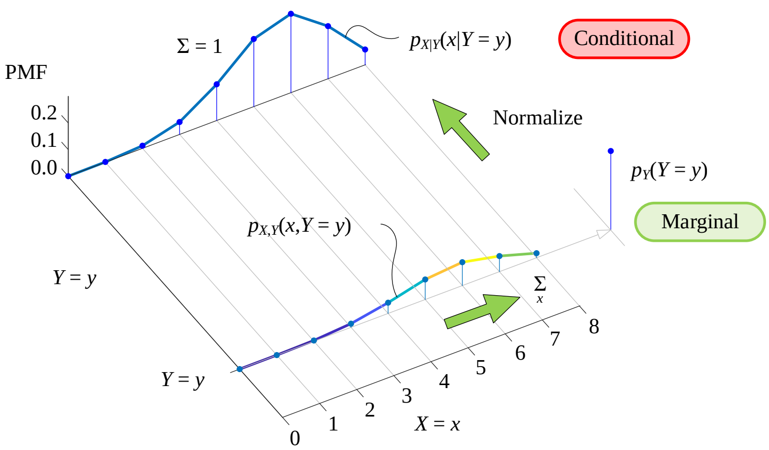


图 5. 条件概率 PMF $p_{X|Y}(x|y)$，X 和 Y 均为离散随机变量

对于条件**期望** $\mathrm{E}(X|Y)$，全**期望**定理为：

$$\underbrace{\mathrm{E}(X)}_{\text{Expectation}} = \mathrm{E}\left[\underbrace{\mathrm{E}(X \mid Y)}_{\text{Conditional expectation}}\right] \tag{9}$$

基于事件的条件**期望**

给定事件 C 发生的条件下 ($\Pr(C) > 0$)，随机变量 X 的条件**期望**为：

$$\begin{aligned}\mathrm{E}\left(X \mid C\right) &= \sum_x x \cdot \underbrace{p_{X \mid C}\left(x \mid C\right)}_{\text{Conditional}} \\ &= \sum_x x \cdot \frac{\overbrace{p_{X,C}\left(x, C\right)}^{\text{Joint}}}{\Pr\left(C\right)}\end{aligned} \tag{10}$$

举个例子，事件 C 可以是鸢尾花数据中指定的标签。

这个式子类似前文的两个随机变量的条件**期望**，大家会在本章后续看到上式的用途。

独立

特别地，如果 X 和 Y 独立，则：

$$\begin{aligned}\mathrm{E}\left(Y \mid X = x\right) &= \mathrm{E}\left(Y\right) \\ \mathrm{E}\left(X \mid Y = y\right) &= \mathrm{E}\left(X\right)\end{aligned} \tag{11}$$

8.2 离散随机变量：条件方差

在上一节的基础上，本节介绍离散随机变量的条件**方差**。

条件方差 var(Y|X = x)

给定 $X = x$ 条件下，Y 的条件**方差** var(Y|X = x) (conditional variance of Y given $X = x$) 定义为：

$$\begin{aligned}\mathrm{var}\left(Y \mid X = x\right) &= \overbrace{\sum_y \left(\underbrace{y - \overbrace{\mathrm{E}\left(Y \mid X = x\right)}^{\text{Expectation}}}_{\text{Deviation}}\right)^2 \cdot \underbrace{p_{Y \mid X}\left(y \mid x\right)}_{\text{Conditional}}}^{\text{Expectation}} \\ &= \sum_y \left(y - \mathrm{E}\left(Y \mid X = x\right)\right)^2 \cdot \frac{\overbrace{p_{X,Y}\left(x, y\right)}^{\text{Joint}}}{\underbrace{p_X\left(x\right)}_{\text{Marginal}}} \\ &= \frac{1}{\underbrace{p_X\left(x\right)}_{\text{Marginal}}} \sum_y \left(\underbrace{y - \mathrm{E}\left(Y \mid X = x\right)}_{\text{Deviation}}\right)^2 \cdot \overbrace{p_{X,Y}\left(x, y\right)}^{\text{Joint}}\end{aligned} \tag{12}$$

下面解剖上式。

$\mathrm{E}\left(Y \mid X = x\right)$ 是 (1) 中求得的条件**期望**，也就是计算偏差的基准。

$y-\mathrm{E}(Y|X=x)$代表偏差，即每个 y 和 $\mathrm{E}(Y|X=x)$ 之间的偏离。$y-\mathrm{E}(Y|X=x)$平方后，再以 $p_{Y|X}(y|x)$ 为权重，求平均值，结果就是条件**方差**。

对比离散随机变量 Y 的**方差**和条件**方差**：

$$
\begin{aligned}
\operatorname{var}(Y) &= \sum_{y}\Big(\underbrace{y-\mathrm{E}(Y)}_{\text{Deviation}}\Big)^2 \cdot \underbrace{p_Y(y)}_{\text{Weight}} \\
\operatorname{var}(Y) &= \sum_{y}\Big(\underbrace{y-\mathrm{E}(Y\mid X=x)}_{\text{Deviation}}\Big)^2 \cdot \underbrace{p_{Y|X}(y\mid x)}_{\text{Weight}}
\end{aligned}
\tag{13}
$$

可以发现两处变差异，度量偏差的基准从 $\mathrm{E}(Y)$ 变成 $\mathrm{E}(Y\mid X=x)$。加权平均的权重从 $p_Y(y)$ 变成 $p_{Y|X}(y\mid x)$。

类似**方差**的简便计算技巧，条件**方差** $\operatorname{var}(Y|X=x)$ 也有如下计算技巧：

$$
\begin{aligned}
\operatorname{var}(Y) &= \mathrm{E}(Y^2)-\mathrm{E}(Y)^2 \\
\operatorname{var}(Y|X=x) &= \mathrm{E}(Y^2|X=x)-\mathrm{E}(Y|X=x)^2
\end{aligned}
\tag{14}
$$

全方差定理

全方差定理 (law of total variance)，又叫**重叠期望定理** (law of iterated variance)，指的是：

$$
\operatorname{var}(Y) = \underbrace{\mathrm{E}\big(\operatorname{var}(Y\mid X)\big)}_{\text{Expectation of conditional variance}} + \underbrace{\operatorname{var}\big(\mathrm{E}(Y\mid X)\big)}_{\text{Variance of conditional expectation}}
\tag{15}
$$

$\mathrm{E}\big(\operatorname{var}(Y\mid X)\big)$ 是条件**方差**的**期望** (加权平均数)：

$$
\underbrace{\mathrm{E}\big(\operatorname{var}(Y\mid X)\big)}_{\text{Expectation of conditional variance}} = \sum_{x}\operatorname{var}(Y\mid X=x)\cdot p_X(x)
\tag{16}
$$

条件**方差**的**期望** $\mathrm{E}\big(\operatorname{var}(Y\mid X)\big)$ 还不够解释整体的**方差**。缺少的成分是条件**期望**的**方差** $\operatorname{var}\big(\mathrm{E}(Y\mid X)\big)$：

$$
\underbrace{\operatorname{var}\big(\mathrm{E}(Y\mid X)\big)}_{\text{Variance of conditional expectation}} = \sum_{x}\big(\mathrm{E}(Y\mid X=x)-\mathrm{E}(Y)\big)^2\cdot p_X(x)
\tag{17}
$$

根据全**期望**定理，$\mathrm{E}(Y\mid X=x)$的**期望**为 $\mathrm{E}(Y)$。

换个方向思考，(15) 相当于对 $\operatorname{var}(Y)$ 的分解：

$$\begin{aligned}\operatorname{var}(Y) &= \underbrace{\mathrm{E}\left(\operatorname{var}(Y \mid X)\right)}_{\text{Expectation of conditional variance}} + \underbrace{\operatorname{var}\left(\mathrm{E}(Y \mid X)\right)}_{\text{Variance of conditional expectation}} \\ &= \sum_x \underbrace{\operatorname{var}(Y \mid X = x)}_{\text{Deviation within a subset}} \cdot \underbrace{p_X(x)}_{\text{Weight}} + \underbrace{\sum_x \Big(\overbrace{\mathrm{E}(Y \mid X = x) - \mathrm{E}(Y)}^{\text{Deviation of a subset from superset}}\Big)^2 \cdot \overbrace{p_X(x)}^{\text{Weight}}}_{\text{Deviation among all subsets}}\end{aligned} \tag{18}$$

这方便我们理解哪些成分 (子集内部、子集之间) 以多大的比例贡献了整体的**方差**。

如图 6 所示，条件**方差**的**期望**解释的是子集 (鸡子集、兔子集) 各自内部差异。

条件**期望**的**方差**解释的是子集 (鸡子集、兔子集) 和母集 (所有动物) 之间的差异。

而代表鸡子集、兔子集就是鸡、兔各自的平均体重 (条件**期望**)，代表母集就是笼子里所有动物的平均体重 (总体**期望**)。

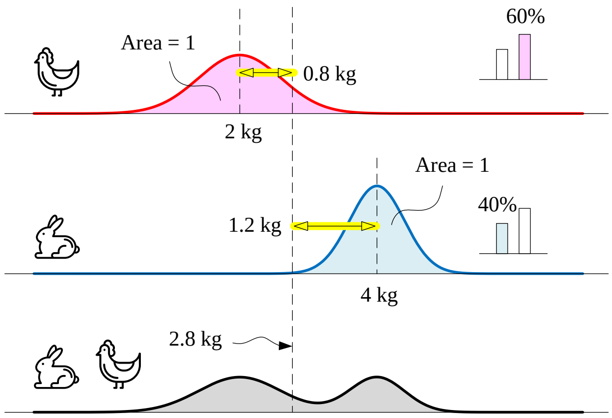


图 6. 解释全**方差**定理

比较图 6 和图 7，条件**方差**的**期望**不变，但是条件**期望**的**方差**增大。如图 7 所示，子集内部差异 (**方差**) 不变，如果增大子集之间的差异，也就是增大了子集和母集的差异，这会导致整体的**方差**增大。

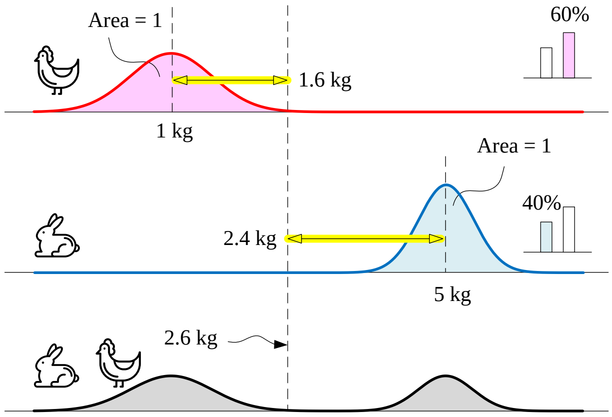


图 7. 解释全方差定理，增大子集之间差异，整体方差增大

类似全方差定理，也存在如下**全协方差定理** (law of total covariance)：

$$\text{cov}(X_1, X_2) = \text{E}(\text{cov}(X_1, X_2 | Y)) + \text{cov}(\text{E}(X_1 | Y), \text{E}(X_2 | Y)) \tag{19}$$

本章不展开分析全协方差定理。

条件方差 var(X|Y = y)

给定 $Y = y$ 条件下，X 的条件方差 $\text{E}(X|Y = y)$ (conditional variance of X given $Y = y$) 定义为：

$$\begin{aligned}
\text{var}(X|Y=y) &= \overbrace{\sum_x \Big(\underbrace{x - \overbrace{\text{E}(X|Y=y)}^{\text{Expectation}}}_{\text{Deviation}}\Big)^2 \cdot \underbrace{p_{X|Y}(x|y)}_{\text{Conditional}}}^{\text{Expectation}} \\
&= \sum_x \big(x - \text{E}(X|Y=y)\big)^2 \cdot \frac{\overbrace{p_{X,Y}(x,y)}^{\text{Joint}}}{\underbrace{p_Y(y)}_{\text{Marginal}}} \\
&= \frac{1}{\underbrace{p_Y(y)}_{\text{Marginal}}} \sum_x \Big(\underbrace{x - \text{E}(X|Y=y)}_{\text{Deviation}}\Big)^2 \cdot \overbrace{p_{X,Y}(x,y)}^{\text{Joint}}
\end{aligned} \tag{20}$$

条件方差 $\text{var}(X|Y = y)$ 也有如下计算技巧：

$$\text{var}(X|Y=y) = \text{E}(X^2|Y=y) - \text{E}(X|Y=y)^2 \tag{21}$$

对于随机变量 X，它的全方差定理为：

$$\text{var}(X) = \underbrace{\text{E}(\text{var}(X|Y))}_{\text{Expectation of conditional variance}} + \underbrace{\text{var}(\text{E}(X|Y))}_{\text{Variance of conditional expectation}} \tag{22}$$

8.3 离散随机变量条件期望、条件方差：以鸢尾花为例

给定花萼长度，条件期望 $\mathrm{E}(X_2 \mid X_1 = x_1)$

大家已经在本书第 4 章见过图 8 中左图。这幅图给出的是条件概率 $p_{X2 \mid X1}(x_2 \mid x_1)$。提醒大家回忆，图中 $p_{X2 \mid X1}(x_2 \mid x_1)$ 每列 PMF (即概率) 和为 1，即满足 (4)。

下面，我们试着利用图 8 中左图计算花萼长度 $X_1 = 6.5$ 为条件下，条件期望 $\mathrm{E}(X_2 \mid X_1 = 6.5)$:

$$
\begin{aligned}
\mathrm{E}\left(X_2 \middle| X_1 = 6.5\right) &= \sum_{x_2} x_2 \cdot p_{X2|X1}\left(x_2 \mid 6.5\right) \\
&= \underset{\text{cm}}{2.0} \times 0 + \underset{\text{cm}}{2.5} \times 0.19 + \underset{\text{cm}}{3.0} \times 0.65 + \underset{\text{cm}}{3.5} \times 0.16 + \underset{\text{cm}}{4.0} \times 0 + \underset{\text{cm}}{4.5} \times 0 \\
&\approx 2.984 \ \text{cm}
\end{aligned}
\tag{23}
$$

注意，上式中条件概率的结果还是 cm。建议大家手算剩余所有 $\mathrm{E}(X_2 \mid X_1 = x_1)$。

图 8 中右上图给出的是热图 $x_2 \cdot p_{X2|X1}\left(x_2 \mid x_1\right)$，相当于一个二元函数。

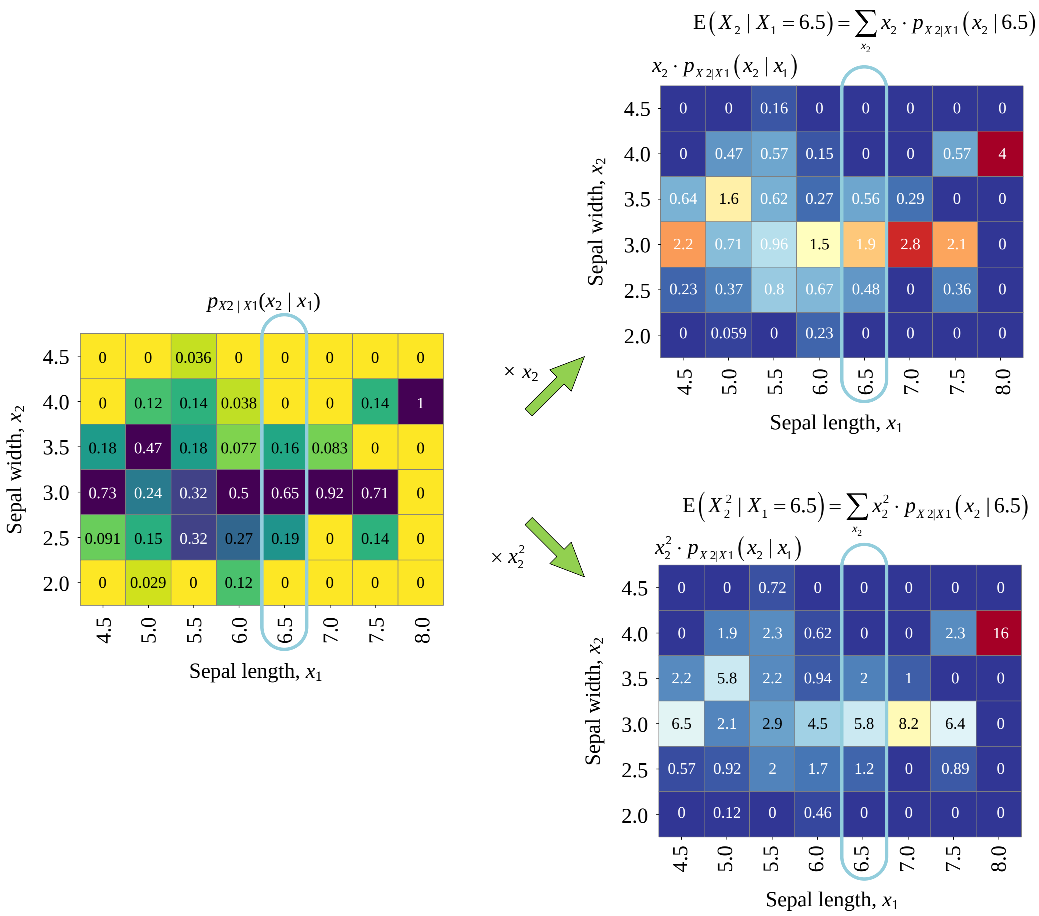


图 8. 给定花萼长度 X_1，花萼宽度 X_2 的条件概率 $p_{X2 \mid X1}(x_2 \mid x_1)$ 热图，$x_2 \times p_{X2 \mid X1}(x_2 \mid x_1)$ 热图，$x_2^2 \cdot p_{X2|X1}\left(x_2 \mid x_1\right)$ 热图

图 9 所示为从矩阵乘法视角看条件**期望** $\mathrm{E}(X_2 \mid X_1 = x_1)$ 运算。

图 10 所示为条件**期望** $\mathrm{E}(X_2 \mid X_1 = x_1)$ 的火柴梗图。图 10 中还给出了鸢尾花花萼长度 X_1 的边缘 PMF $p_{X1}(x_1)$。

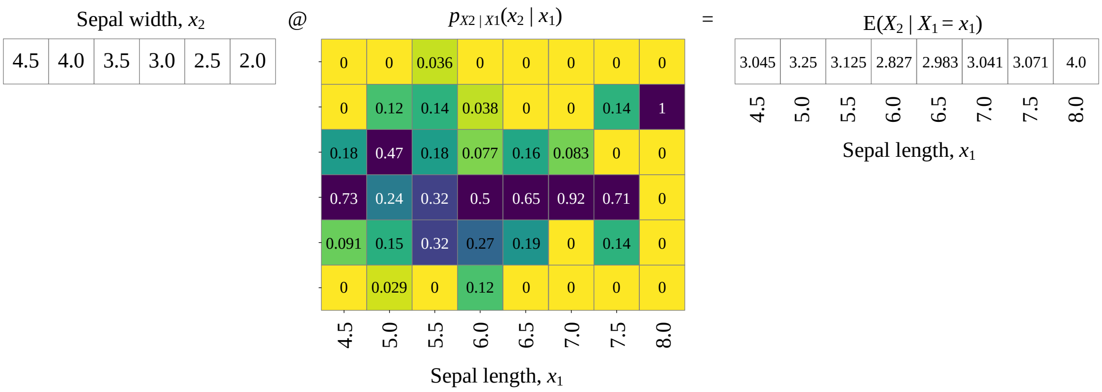


图 9. 矩阵乘法视角看条件**期望** $\mathrm{E}(X_2 \mid X_1 = x_1)$

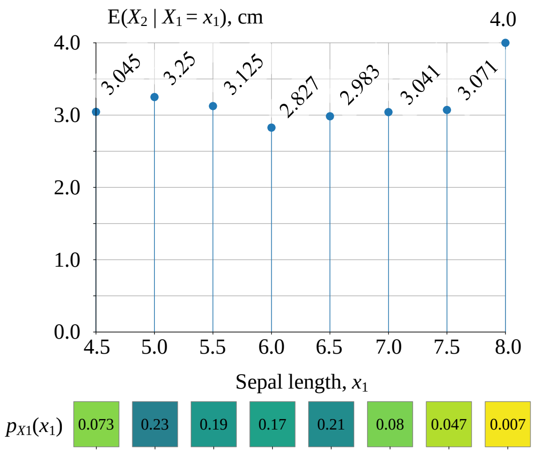


图 10. 给定花萼长度 X_1，花萼宽度 X_2 的条件**期望** $\mathrm{E}(X_2 \mid X_1 = x_1)$，和边缘 PMF $p_{X1}(x_1)$

根据 (5) 的全**期望**定理，我们可以利用条件**期望** $\mathrm{E}(X_2 \mid X_1 = x_1)$ 和边缘 PMF $p_{X1}(x_1)$ 计算**期望** $\mathrm{E}(X_2)$：

$$
\begin{aligned}
\mathrm{E}(X_2) &= \sum_{x_1} \mathrm{E}(X_2 \mid X_1 = x_1) \cdot p_{X1}(x_1) \\
&= \underbrace{3.045}_{\text{cm}} \times 0.073 + \underbrace{3.25}_{\text{cm}} \times 0.23 + \underbrace{3.125}_{\text{cm}} \times 0.19 + \underbrace{2.827}_{\text{cm}} \times 0.17 + \\
&\quad \underbrace{2.983}_{\text{cm}} \times 0.21 + \underbrace{3.041}_{\text{cm}} \times 0.08 + \underbrace{3.071}_{\text{cm}} \times 0.047 + \underbrace{4}_{\text{cm}} \times 0.007 \\
&\approx 3.063 \ \text{cm}
\end{aligned}
\tag{24}
$$

给定花萼长度，条件**方差** $\mathrm{var}(X_2 \mid X_1 = x_1)$

利用 (12) 计算花萼长度 $X_1 = 6.5$ 为条件下，条件**方差** $\mathrm{var}(X_2 \mid X_1 = 6.5)$：

$$\begin{aligned}\mathrm{var}\left(X_2 \middle| X_1 = 6.5\right) &= \sum_{x_2}\left(x_2 - \mathrm{E}\left(X_2 \middle| X_1 = 6.5\right)\right)\cdot p_{X2|X1}\left(x_2 \mid 6.5\right)\\ &= \underbrace{\left(2.0-2.985\right)^2}_{\mathrm{cm}^2}\times 0+\underbrace{\left(2.5-2.985\right)^2}_{\mathrm{cm}^2}\times 0.19+\underbrace{\left(3.0-2.985\right)^2}_{\mathrm{cm}^2}\times 0.65+\\ &\quad\underbrace{\left(3.5-2.985\right)^2}_{\mathrm{cm}^2}\times 0.16+\underbrace{\left(4.0-2.985\right)^2}_{\mathrm{cm}^2}\times 0+\underbrace{\left(4.0-2.985\right)^2}_{\mathrm{cm}^2}\times 0\\ &\approx 0.088\ \mathrm{cm}^2\end{aligned} \tag{25}$$

条件**方差** $\mathrm{var}(X_2 \mid X_1 = 6.5)$ 的单位为 cm^2。同样建议大家手算剩余条件**方差** $\mathrm{var}(X_2 \mid X_1 = x_1)$。

采用技巧计算，计算条件**期望**。首先计算花萼长度 $X_1 = 6.5$ 为条件下，花萼宽度平方的**期望**：

$$\begin{aligned}\mathrm{E}\left(X_1^2 \middle| X_1 = 6.5\right) &= \sum_{x_2} x_1^2 \cdot p_{X2|X1}\left(x_2 \mid 6.5\right)\\ &= \underbrace{2.0^2}_{\mathrm{cm}^2}\times 0+\underbrace{2.5^2}_{\mathrm{cm}^2}\times 0.19+\underbrace{3.0^2}_{\mathrm{cm}^2}\times 0.65+\underbrace{3.5^2}_{\mathrm{cm}^2}\times 0.16+\underbrace{4.0^2}_{\mathrm{cm}^2}\times 0+\underbrace{4.5^2}_{\mathrm{cm}^2}\times 0\\ &\approx 9\ \mathrm{cm}^2\end{aligned} \tag{26}$$

图 11 所示为花萼宽度平方值 X_2^2 的条件**期望** $\mathrm{E}\left(X_2^2 \mid X_1 = x_1\right)$ 的火柴梗图。

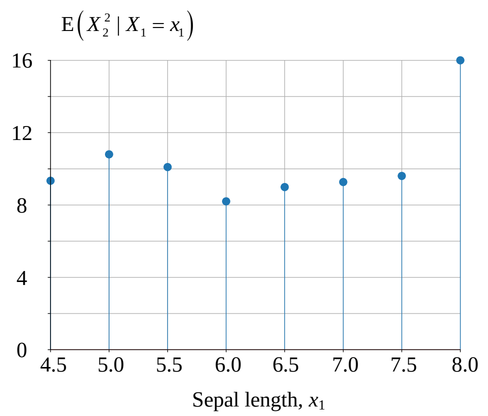


图 11. 给定花萼长度 X_1，花萼宽度平方值 X_2^2 的条件**期望** $\mathrm{E}\left(X_2^2 \mid X_1 = x_1\right)$

然后计算条件**方差**：

$$\mathrm{var}\left(X_2 \middle| X_1 = 6.5\right) = \mathrm{E}\left(X_1^2 \middle| X_1 = 6.5\right) - \mathrm{E}\left(X_2 \middle| X_1 = 6.5\right)^2 = 9 - 2.984^2 \approx 0.088 \tag{27}$$

图 12 所示为花萼长度取不同值时条件**方差** $\text{var}(X_2 \mid X_1 = x_1)$ 的火柴梗图，请大家用作检查自己手算结果。

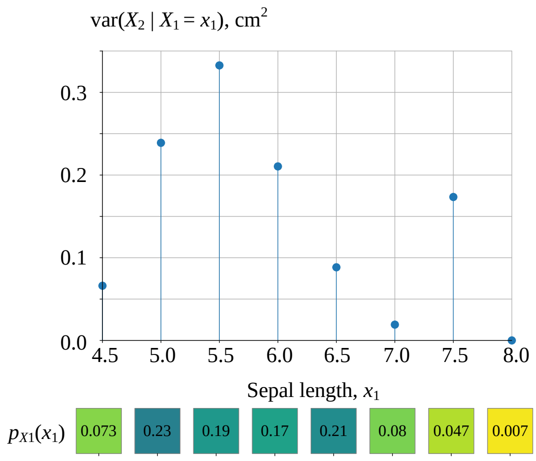


图 12. 给定花萼长度 X_1，花萼宽度的条件**方差** $\text{var}(X_2 \mid X_1 = x_1)$

大家肯定早就发现，条件**期望** $\text{E}(X_2 \mid X_1 = x_1)$、条件**方差** $\text{var}(X_2 \mid X_1 = x_1)$ 都消去了 x_2 这个变量，两者仅仅随着 $X_1 = x_1$ 取值变化。这也不难理解，**期望**和**方差**代表“汇总”，本质上就是“降维”。某个维度上的信息细节不再重要，我们把这个“压扁”。

压扁过程中，不同的聚合方式得到不同的统计量，比如**期望**、**方差**等等。

全方差定理：还原方差 $\text{var}(X_2)$

根据 (17) 中给出的全**方差**定理，下面我们利用条件**方差** $\text{var}(X_2 \mid X_1)$ 和条件**期望** $\text{E}(X_2 \mid X_1)$ 计算花萼宽度的**方差** $\text{var}(X_2)$。$\text{var}(X_2)$ 可以写成两部分之和：

$$\text{var}(X_2) = \underbrace{\text{E}\left(\text{var}\left(X_2 \mid X_1\right)\right)}_{\text{Expectation of conditional variance}} + \underbrace{\text{var}\left(\text{E}\left(X_2 \mid X_1\right)\right)}_{\text{Variance of conditional expectation}} \tag{28}$$

第一部分是条件**方差**的**期望** $\text{E}\left(\text{var}\left(X_2 \mid X_1\right)\right)$：

$$\underbrace{\text{E}\left(\text{var}\left(X_2 \mid X_1\right)\right)}_{\text{Expectation of conditional variance}} = \sum_{x_1} \text{var}\left(X_2 \mid X_1 = x_1\right) \cdot p_{X1}\left(x_1\right) \tag{29}$$

代入具体数值，我们可以计算得到 $\text{E}\left(\text{var}\left(X_2 \mid X_1\right)\right)$：

$$
\begin{aligned}
\underbrace{\mathrm{E}\left(\operatorname{var}\left(X_2 \mid X_1\right)\right)}_{\text{Expectation of conditional variance}} &= \sum_{x_1} \operatorname{var}\left(X_2 \mid X_1 = x_1\right) \cdot p_{X1}\left(x_1\right) \\
&\approx \underbrace{0.066}_{\text{cm}^2} \times 0.073 + \underbrace{0.238}_{\text{cm}^2} \times 0.226 + \underbrace{0.332}_{\text{cm}^2} \times 0.186 + \underbrace{0.210}_{\text{cm}^2} \times 0.173 + \\
&\quad \underbrace{0.088}_{\text{cm}^2} \times 0.206 + \underbrace{0.019}_{\text{cm}^2} \times 0.08 + \underbrace{0.173}_{\text{cm}^2} \times 0.046 + \underbrace{0}_{\text{cm}^2} \times 0.006 \\
&\approx \underbrace{0.0048}_{X_1=4.5} + \underbrace{0.0541}_{X_1=5.0} + \underbrace{0.0620}_{X_1=5.5} + \underbrace{0.0364}_{X_1=6.0} + \underbrace{0.0182}_{X_1=6.5} + \underbrace{0.0015}_{X_1=7.0} + \underbrace{0.0080}_{X_1=7.5} + \underbrace{0}_{X_1=8.0} \\
&\approx 0.185 \text{ cm}^2
\end{aligned}
\tag{30}
$$

第二部分是条件**期望**的**方差** $\operatorname{var}\left(\mathrm{E}\left(X_2 \mid X_1\right)\right)$。代入具体值计算得到：

$$
\begin{aligned}
\underbrace{\operatorname{var}\left(\mathrm{E}\left(X_2 \mid X_1\right)\right)}_{\text{Variance of conditional expectation}} &= \sum_{x_1} \left(\mathrm{E}\left(X_2 \mid X_1 = x_1\right) - \mathrm{E}\left(X_2\right)\right)^2 \cdot p_{X_1}\left(x_1\right) \\
&\approx 0.025 \text{ cm}^2
\end{aligned}
\tag{31}
$$

如果大家看到这还会犯糊涂，不理解为什么 $\sum_{x_1}$ 求和遍历的是 x_1？我告诉大家一个小技巧，因为 X_2 已经被“折叠”！不管是条件期望 $\mathrm{E}\left(X_2 \mid X = x_1\right)$、还是期望 $\mathrm{E}\left(X_2\right)$，都已经将 X_2 折叠成一个具体的数值，因此无法遍历。

这样 X_2 的方差约为：

$$
\begin{aligned}
\operatorname{var}\left(X_2\right) &= \underbrace{\mathrm{E}\left(\operatorname{var}\left(X_2 \mid X_1\right)\right)}_{\text{Expectation of conditional variance}} + \underbrace{\operatorname{var}\left(\mathrm{E}\left(X_2 \mid X_1\right)\right)}_{\text{Variance of conditional expectation}} \\
&\approx 0.185 + 0.025 = 0.211 \text{cm}^2
\end{aligned}
\tag{32}
$$

在 $\operatorname{var}(X_2)$ 中，第一部分 $\mathrm{E}\left(\operatorname{var}\left(X_2 \mid X_1\right)\right)$ 贡献超过 85%。而 $\mathrm{E}\left(\operatorname{var}\left(X_2 \mid X_1\right)\right)$ 可以进一步展开，图 13 所示为各个不同成分对花萼宽度 X_2 的方差 $\operatorname{var}(X_2)$ 的贡献，这也可以叫做**钻取** (drill down)。

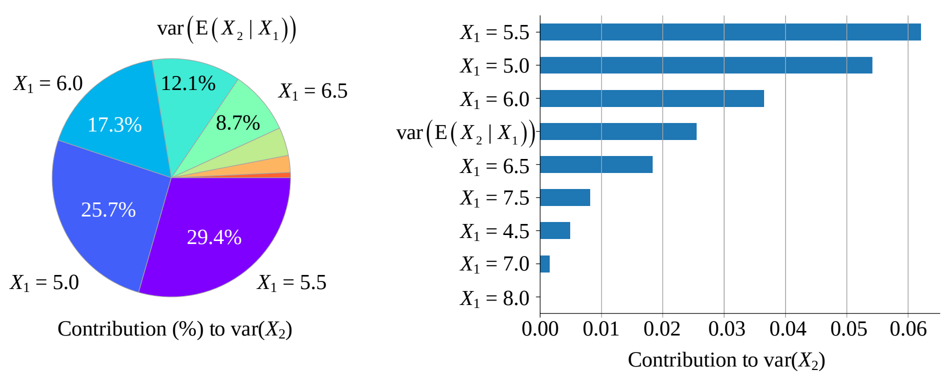


图 13. 各个不同成分对花萼宽度 X_2 的方差 $\operatorname{var}(X_2)$ 的贡献

给定花萼长度，条件**标准差** $\operatorname{std}(X_2 \mid X_1 = x_1)$

(25) 开方便获得条件**标准差** $\mathrm{std}(X_2 \mid X_1 = 6.5)$：

$$\sigma_{X_2|X_1=6.5} = \mathrm{std}\left(X_2 \middle| X_1 = 6.5\right) = 0.295 \text{ cm} \tag{33}$$

上式的单位和鸢尾花宽度单位一致，我们便可以把条件**标准差**和图 10 画在一起，得到图 14。这幅图给出的是 $\mathrm{E}(X_2 \mid X_1 = x_1) \pm \mathrm{std}(X_2 \mid X_1 = x_1)$。

圆点 ● 展示的是 $\mathrm{E}(X_2 \mid X_1 = x_1)$，即条件**期望**，代表给定 $X_1 = x_1$ 条件下，鸢尾花数据在花萼宽度上的一种“预测”！这和我们讲过的回归思想本质上相同。$\mathrm{E}(X_2 \mid X_1 = x_1)$ 代表当 $X_1 = x_1$ 时鸢尾花花萼宽度最合适的“预测”。也就是说，回归可以看成是条件概率！本书后续还会沿着这个思路展开讨论。

而我们用**误差棒** (error bar) 展示$\pm\ \mathrm{std}(X_2 \mid X_1 = x_1)$，代表给定 $X_1 = x_1$ 条件下，鸢尾花数据在花萼宽上的“波动”。误差棒的宽度越大，说明波动越大；反之，则说明波动越小。

特别地，当花萼长度 X_1 为 8.0 cm 时，条件均**方差** $\mathrm{std}(X_2 \mid X_1 = 8.0)$ 为 0。这是因为，这一处只有一个样本点。

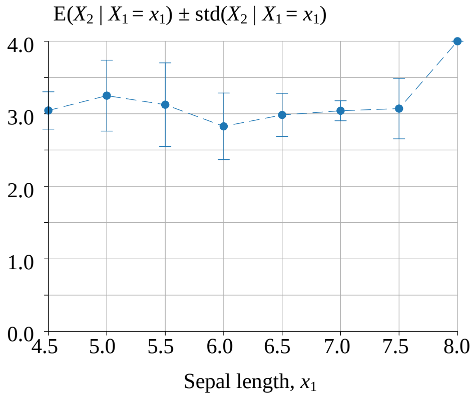


图 14. 给定花萼长度 X_1，花萼宽度 X_2 的条件**期望** $\mathrm{E}(X_2 \mid X_1 = x_1) \pm \mathrm{std}(X_2 \mid X_1 = x_1)$

给定花萼宽度，条件期望 $\mathrm{E}(X_1 \mid X_2 = x_2)$

图 15 给出的是条件概率 $p_{X1 \mid X2}(x_1 \mid x_2)$。同样提醒大家注意图中 $p_{X1 \mid X2}(x_1 \mid x_2)$ 每行 PMF (即概率) 和为 1。

利用 图 15 计算花萼宽度 $X_2 = 2.0$ 为条件下，条件**期望** $\mathrm{E}(X_1 \mid X_2 = 2.0)$：

$$\begin{aligned}\mathrm{E}\left(X_1 \middle| X_2 = 2.0\right) &= \sum_{x_1} x_1 \cdot p_{X1|X2}\left(x_1 \mid 2.0\right) \\ &= \underset{\text{cm}}{4.5} \times 0 + \underset{\text{cm}}{5.0} \times 0.25 + \underset{\text{cm}}{5.5} \times 0 + \underset{\text{cm}}{6.0} \times 0.75 + \\ &\quad \underset{\text{cm}}{6.5} \times 0 + \underset{\text{cm}}{7.0} \times 0 + \underset{\text{cm}}{7.5} \times 0 + \underset{\text{cm}}{8.0} \times 0 \\ &\approx 5.7 \text{ cm}\end{aligned} \tag{34}$$

条件概率的结果还是 cm。同样建议大家手算剩余所有 $\mathrm{E}(X_1 \mid X_2 = x_2)$。

此外，请大家也根据全**期望**定理，利用 $\mathrm{E}(X_1 \mid X_2 = x_2)$ 计算 $\mathrm{E}(X_1)$。并用条件**方差** $\mathrm{var}(X_1 \mid X_2)$ 和条件**期望** $\mathrm{E}(X_1 \mid X_2)$ 计算花萼长度的**方差** $\mathrm{var}(X_1)$。

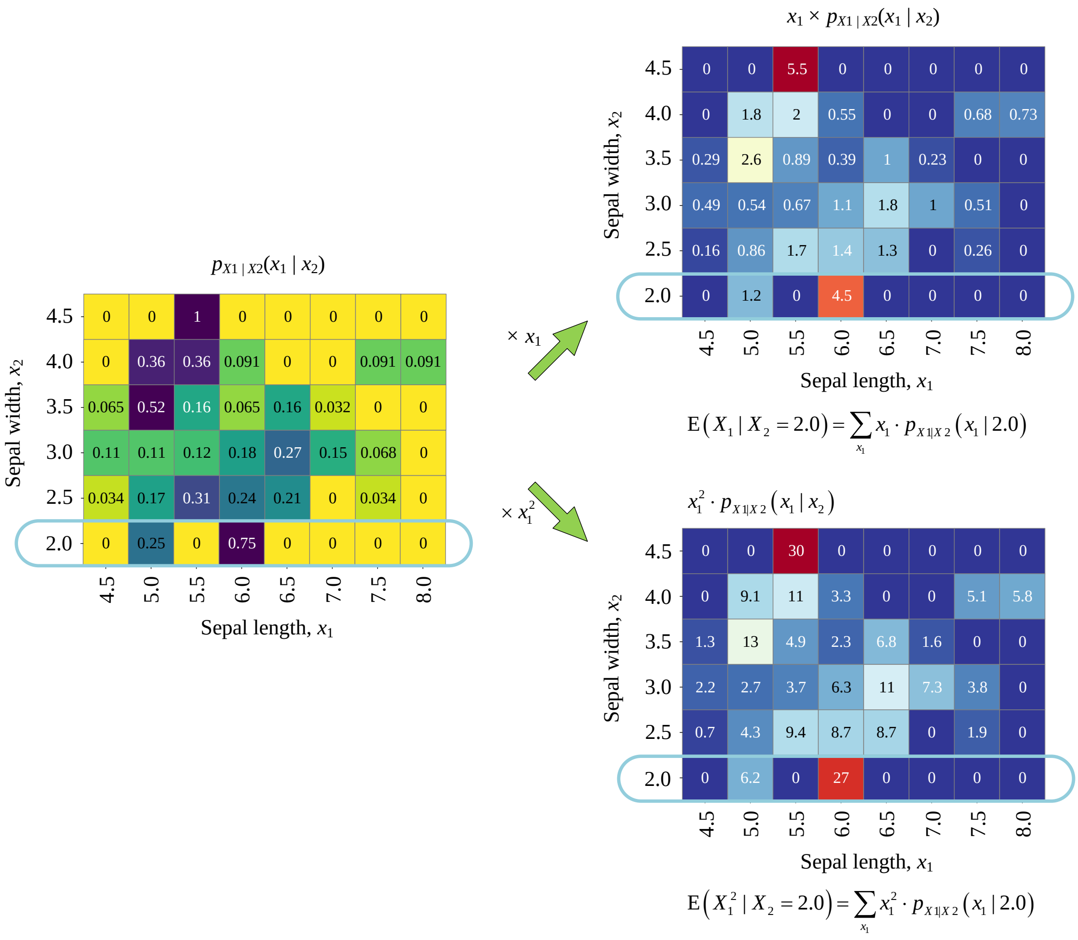


图 15. 给定花萼宽度，花萼长度的条件概率 $p_{X1 \mid X2}(x_1 \mid x_2)$

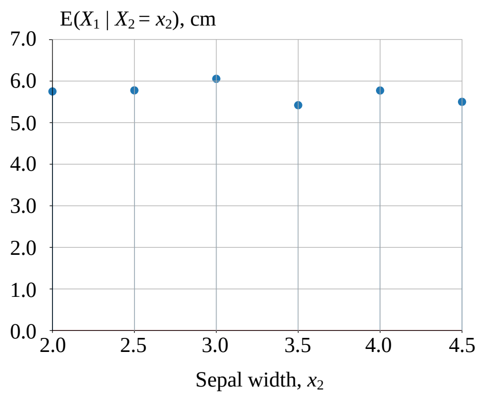


图 16. 给定花萼宽度 X_2，花萼宽度的条件**期望** $E(X_1 \mid X_2 = x_2)$

条件**方差** $\mathrm{var}(X_1 \mid X_2 = x_2)$

在花萼宽度 $X_2 = 2.0$ 为条件下，条件**方差** $\mathrm{var}(X_1 \mid X_2 = 2.0)$：

$$
\begin{aligned}
\mathrm{var}\left(X_1 \middle| X_2 = 2.0\right) &= \sum_{x_1} \left(x_1 - \mathrm{E}\left(X_1 \middle| X_2 = 2.0\right)\right)^2 \cdot p_{X1|X2}\left(x_1 \mid 2.0\right) \\
&= \underbrace{\left(4.5-5.75\right)^2}_{\mathrm{cm}^2} \times 0 + \underbrace{\left(5.0-5.75\right)^2}_{\mathrm{cm}^2} \times 0.25 + \underbrace{\left(5.5-5.75\right)^2}_{\mathrm{cm}^2} \times 0 + \underbrace{\left(6.0-5.75\right)^2}_{\mathrm{cm}^2} \times 0.75 + \\
&\quad \underbrace{\left(6.5-5.75\right)^2}_{\mathrm{cm}^2} \times 0 + \underbrace{\left(7.0-5.75\right)^2}_{\mathrm{cm}^2} \times 0 + \underbrace{\left(7.5-5.75\right)^2}_{\mathrm{cm}^2} \times 0 + \underbrace{\left(8.0-5.75\right)^2}_{\mathrm{cm}^2} \times 0 \\
&= 0.1875 \ \mathrm{cm}^2
\end{aligned}
\tag{35}
$$

条件**方差** $\mathrm{var}(X_1 \mid X_2 = 2.0)$ 的单位为 cm^2。同样建议大家手算剩余条件**方差** $\mathrm{var}(X_1 \mid X_2 = x_2)$。

利用条件**方差**计算技巧，首先计算花萼宽度 $X_2 = 2.0$ 为条件下，花萼长度平方的**期望**：

$$
\begin{aligned}
\mathrm{E}\left(X_1^2 \middle| X_2 = 2.0\right) &= \sum_{x_1} x_1^2 \cdot p_{X1|X2}\left(x_2 \mid 2.0\right) \\
&= \underbrace{4.5^2}_{\mathrm{cm}^2} \times 0 + \underbrace{5.0^2}_{\mathrm{cm}^2} \times 0.25 + \underbrace{5.5^2}_{\mathrm{cm}^2} \times 0 + \underbrace{6.0^2}_{\mathrm{cm}^2} \times 0.75 + \\
&\quad \underbrace{6.5^2}_{\mathrm{cm}^2} \times 0 + \underbrace{7.0^2}_{\mathrm{cm}} \times 0 + \underbrace{7.5^2}_{\mathrm{cm}^2} \times 0 + \underbrace{8.0^2}_{\mathrm{cm}^2} \times 0 \\
&= 33.25 \ \mathrm{cm}^2
\end{aligned}
\tag{36}
$$

图 17 所示为给定花萼长度 X_2，花萼宽度平方值 X_1^2 的条件**期望** $\mathrm{E}\left(X_1^2 \mid X_2 = x_2\right)$。请大家自行代入计算条件**方差** $\mathrm{var}(X_1 \mid X_2 = 2.0)$。

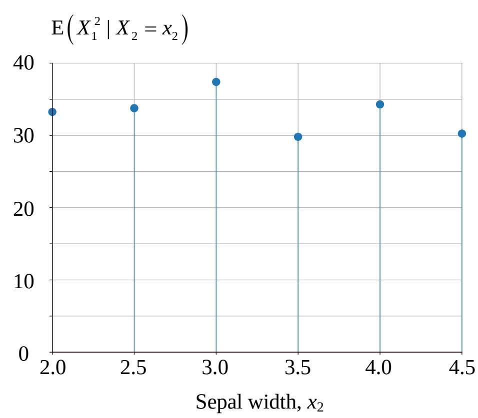


图 17. 给定花萼长度 X_2，花萼宽度平方值 X_1^2 的条件**期望** $\mathrm{E}\left(X_1^2 \mid X_2 = x_2\right)$

图 18 所示为条件**方差** $\mathrm{var}(X_1 \mid X_2 = x_2)$ 的火柴梗图。同样，条件**期望** $\mathrm{E}(X_1 \mid X_2 = x_2)$、条件**方差** $\mathrm{var}(X_1 \mid X_2 = x_2)$ 都“折叠”了 x_1 这个维度，两者仅仅随着 $X_2 = x_2$ 取值变化。

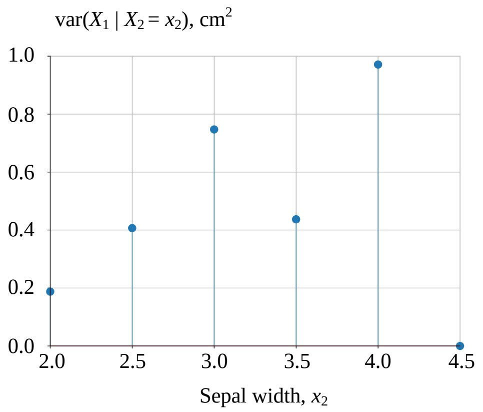


图 18. 给定花萼宽度 X_2，花萼宽度的条件**方差** $\mathrm{var}(X_1 \mid X_2 = x_2)$

类似图 14，我们也绘制给定花萼宽度 X_2，花萼长度 X_1 的条件**期望** $\mathrm{E}(X_1 \mid X_2 = x_2) \pm \mathrm{std}(X_1 \mid X_2 = x_2)$。请大家自行分析这幅图像。

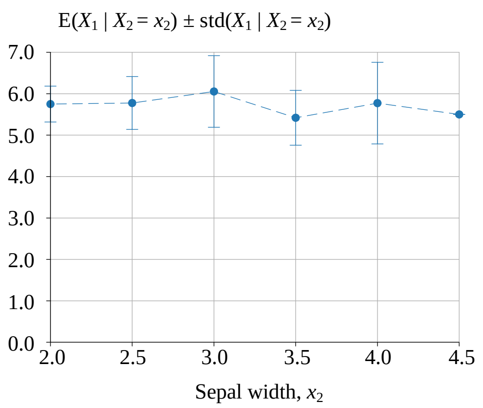


图 19. 给定花萼宽度 X_2，花萼长度 X_1 的条件**期望** $\mathrm{E}(X_1 \mid X_2 = x_2) \pm \mathrm{std}(X_1 \mid X_2 = x_2)$

考虑标签：花萼长度

给定鸢尾花分类标签 $Y = C_1$，花萼长度 X_1 的条件**期望**：

$$
\begin{aligned}
\mathrm{E}\left(X_1 \middle| Y=C_1\right) &= \sum_{x_1} x_1 \cdot p_{X1|Y}\left(x_1 \mid C_1\right) \\
&= \underset{\text{cm}}{4.5} \times 0.22 + \underset{\text{cm}}{5.0} \times 0.56 + \underset{\text{cm}}{5.5} \times 0.2 + \underset{\text{cm}}{6.0} \times 0.02 + \\
&\quad \underset{\text{cm}}{6.5} \times 0 + \underset{\text{cm}}{7.0} \times 0 + \underset{\text{cm}}{7.5} \times 0 + \underset{\text{cm}}{8.0} \times 0 \\
&= 5.01 \text{ cm}
\end{aligned}
\tag{37}
$$

给定鸢尾花分类标签 $Y = C_1$，花萼长度 X_1 平方**期望**：

$$
\begin{aligned}
\mathrm{E}\left(X_1^2 \middle| Y=C_1\right) &= \sum_{x_1} x_1^2 \cdot p_{X1|Y}\left(x_1 \mid C_1\right) \\
&= \underset{\text{cm}^2}{4.5^2} \times 0.22 + \underset{\text{cm}^2}{5.0^2} \times 0.56 + \underset{\text{cm}^2}{5.5^2} \times 0.2 + \underset{\text{cm}^2}{6.0^2} \times 0.02 + \\
&\quad \underset{\text{cm}^2}{6.5^2} \times 0 + \underset{\text{cm}^2}{7.0^2} \times 0 + \underset{\text{cm}^2}{7.5^2} \times 0 + \underset{\text{cm}^2}{8.0^2} \times 0 \\
&= 25.225 \text{ cm}^2
\end{aligned}
\tag{38}
$$

给定鸢尾花分类标签 $Y = C_1$，花萼长度 X_1 条件**方差**：

$$
\begin{aligned}
\operatorname{var}\left(X_1 \middle| Y=C_1\right) &= \mathrm{E}\left(X_1^2 \middle| Y=C_1\right) - \mathrm{E}\left(X_1 \middle| Y=C_1\right)^2 \\
&= 25.225 - 5.01^2 \\
&= 0.1249 \text{ cm}^2
\end{aligned}
\tag{39}
$$

给定鸢尾花分类标签 $Y = C_1$，花萼长度 X_1 条件**标准差**：

$$
\sigma_{X_1|Y=C_1} = \sqrt{\operatorname{var}\left(X_1 \middle| Y=C_1\right)} = \sqrt{0.1249} = 0.353 \text{ cm}
\tag{40}
$$

请大家自行计算剩余两种情况 ($Y = C_2, C_3$)。并利用全**期望**定理，计算 $\mathrm{E}(X_1)$。

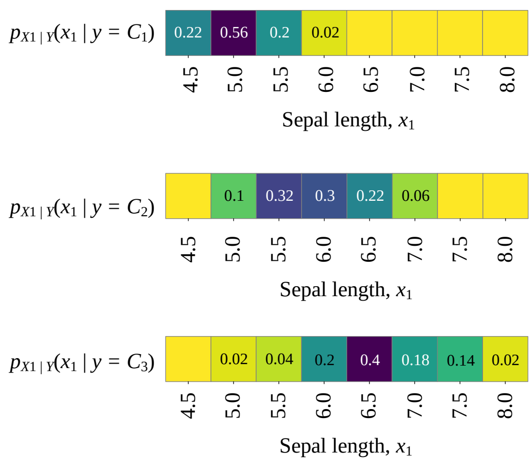


图 20. 给定鸢尾花标签 Y，花萼长度的条件 PMF

考虑标签：花萼宽度

给定鸢尾花分类标签 $Y = C_1$，花萼宽度 X_2 的条件**期望**：

$$\begin{aligned}\mathrm{E}\left(X_2 | Y = C_1\right) &= \sum_{x_2} x_2 \cdot p_{X2|Y}\left(x_2 | C_1\right) \\ &= \underset{\text{cm}}{4.5} \times 0.07 + \underset{\text{cm}}{4.0} \times 0.18 + \underset{\text{cm}}{3.5} \times 0.46 + \underset{\text{cm}}{3.0} \times 0.32 + \underset{\text{cm}}{2.5} \times 0.02 + \underset{\text{cm}}{2.0} \times 0 \\ &= 3.43 \text{ cm}\end{aligned} \tag{41}$$

给定鸢尾花分类标签 $Y = C_1$，花萼宽度 X_2 平方**期望**：

$$\begin{aligned}\mathrm{E}\left(X_2^2 | Y = C_1\right) &= \sum_{x_2} x_2^2 \cdot p_{X2|Y}\left(x_2 | C_1\right) \\ &= \underset{\text{cm}^2}{4.5^2} \times 0.07 + \underset{\text{cm}^2}{4.0^2} \times 0.18 + \underset{\text{cm}^2}{3.5^2} \times 0.46 + \underset{\text{cm}^2}{3.0^2} \times 0.32 + \underset{\text{cm}^2}{2.5^2} \times 0.02 + \underset{\text{cm}^2}{2.0^2} \times 0 \\ &= 11.925 \text{ cm}^2\end{aligned} \tag{42}$$

给定鸢尾花分类标签 $Y = C_1$，花萼宽度 X_2 条件**方差**：

$$\begin{aligned}\operatorname{var}\left(X_2 | Y = C_1\right) &= \mathrm{E}\left(X_2^2 | Y = C_1\right) - \mathrm{E}\left(X_2 | Y = C_1\right)^2 \\ &= 11.925 - 3.43^2 \\ &= 0.1601 \text{ cm}^2\end{aligned} \tag{43}$$

给定鸢尾花分类标签 $Y = C_1$，花萼宽度 X_2 条件**标准差**：

$$\sigma_{X_2|Y=C_1} = \sqrt{\operatorname{var}\left(X_2 | Y = C_1\right)} = \sqrt{0.1601} \approx 0.4 \text{ cm} \tag{44}$$

请大家自行计算鸢尾花其他标签条件下花萼长度、花萼宽度的条件**期望**、条件**方差**、条件**标准差**。

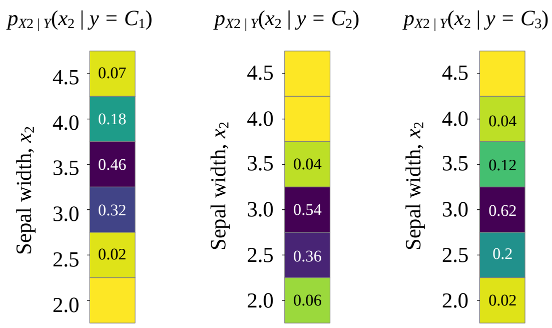


图 21. 给定鸢尾花标签 Y，花萼宽度的条件 PMF

Bk5_Ch08_01.py 代码绘制本节大部分图像。代码中用到了矩阵乘法和广播原则，请大家注意区分。

8.4 连续随机变量：条件期望

本节介绍如何计算连续随机变量的条件期望。

条件期望 E(Y|X = x)

如果 X 和 Y 均为连续随机变量，如图 22 所示，在给定 $X = x$ 条件下，条件期望 $\mathrm{E}(Y|X = x)$ 定义为：

$$
\begin{aligned}
\mathrm{E}\left(Y \middle| \underbrace{X=x}_{\text{Given}}\right) &= \overbrace{\int_{-\infty}^{+\infty} y \cdot \underbrace{f_{Y|X}\left(y|x\right)}_{\text{Conditional}} \mathrm{d}y}^{\text{Expectation}} \\
&= \int_{-\infty}^{+\infty} y \cdot \frac{\overbrace{f_{X,Y}\left(x,y\right)}^{\text{Joint}}}{\underbrace{f_X\left(x\right)}_{\text{Marginal}}} \mathrm{d}y = \frac{1}{\underbrace{f_X\left(x\right)}_{\text{Marginal}}} \int_{-\infty}^{+\infty} y \cdot \overbrace{f_{X,Y}\left(x,y\right)}^{\text{Joint}} \mathrm{d}y
\end{aligned}
\tag{45}
$$

上式中，边缘概率 $f_X(x)$ 可以通过下式得到：

$$
f_X\left(x\right) = \int_{-\infty}^{+\infty} f_{X,Y}\left(x,y\right)\mathrm{d}y \tag{46}
$$

(46) 代入 (45) 得到：

$$
\mathrm{E}\left(Y|X=x\right) = \frac{1}{\int_{-\infty}^{+\infty} f_{X,Y}\left(x,y\right)\mathrm{d}y} \int_{-\infty}^{+\infty} y \cdot f_{X,Y}\left(x,y\right)\mathrm{d}y \tag{47}
$$

上式，相当于消去了 y，这和本章前文提到的“降维”、“折叠”本质上没有任何区别。对于离散随机变量，折叠用的数学工具为求和符号 Σ；连续随机变量则用积分符号 $\int$。

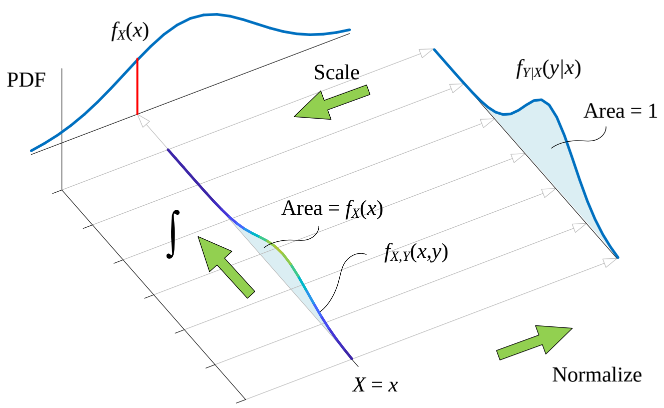


图 22. 联合概率 PDF $f_{X,Y}(x,y)$ 和条件概率 PDF $f_{Y|X}(y|x)$ 的关系，X 和 Y 均为连续随机变量

条件期望 $\mathrm{E}(X|Y=y)$

同理，如图 23 所示，条件期望 $\mathrm{E}(X|Y=y)$ 定义为：

$$\mathrm{E}\left(X\,|Y=y\right)=\frac{1}{\int_{-\infty}^{+\infty} f_{X,Y}\left(x,y\right)\mathrm{d}\,x}\int_{-\infty}^{+\infty} x\cdot f_{X,Y}\left(x,y\right)\mathrm{d}\,x \tag{48}$$

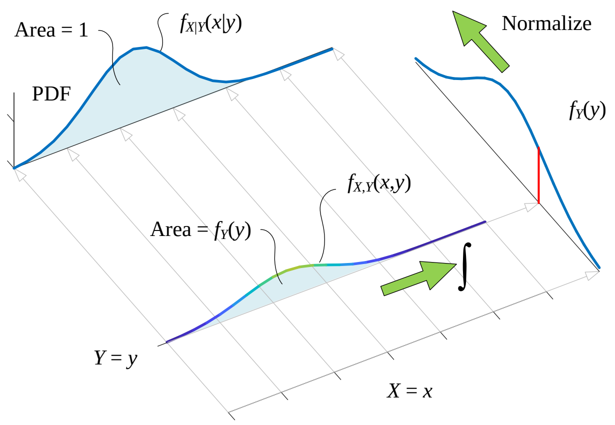


图 23. 联合概率 PDF $f_{X,Y}(x,y)$ 和条件概率 PDF $f_{X|Y}(x|y)$ 的关系，X 和 Y 均为连续随机变量

8.5 连续随机变量：条件方差

本节介绍如何求连续随机变量的条件方差。

条件方差 $\mathrm{var}(Y|X=x)$

在给定 $X=x$ 条件下，条件方差 $\mathrm{var}(Y|X=x)$ (conditional variance of Y given $X=x$) 定义为：

$$\begin{aligned}\mathrm{var}\left(Y\,|X=x\right)&=\mathrm{E}\left\{\left(Y-\mathrm{E}\left(Y\,|X=x\right)\right)^2|x\right\}\\&=\int_y\left(y-\mathrm{E}\left(Y\,|X=x\right)\right)^2\cdot f_{Y|X}\left(y|x\right)\mathrm{d}\,y\end{aligned} \tag{49}$$

对于连续随机变量，求条件方差也可以用 (14) 这个技巧。

条件方差 $\mathrm{var}(X|Y=y)$

条件方差 $\mathrm{var}(X|Y=y)$ 定义为：

$$\begin{aligned}\operatorname{var}(X|Y=y) &= \mathrm{E}\left\{\left(X-\mathrm{E}(X|Y=y)\right)^2|y\right\} \\ &= \int_x \left(X-\mathrm{E}(X|Y=y)\right)^2 \cdot f_{X|Y}(x|y)\,\mathrm{d}x\end{aligned} \tag{50}$$

有了以上理论基础，本书第 12 章将以二元高斯分布为例，继续深入讲解条件**期望**和条件**方差**。

8.6 连续随机变量：以鸢尾花为例

以鸢尾花为例：条件期望 $\mathrm{E}(X_2 \mid X_1 = x_1)$、条件方差 $\operatorname{var}(X_2 \mid X_1 = x_1)$

图 24 (a) 所示为条件概率 PDF $f_{X2 \mid X1}(x_2 \mid x_1)$ 随花萼长度、花萼宽度变化曲面。本书前文提过 $f_{X2 \mid X1}(x_2 \mid x_1)$ 也是一个二元函数。这个二元函数的重要特点有两个：

$$\begin{gathered} f_{X2|X1}(x_2 \mid x_1) \ge 0 \\ \int_{x_2} f_{X2|X1}(x_2 \mid x_1)\,\mathrm{d}x_2 = 1 \end{gathered} \tag{51}$$

正如图 24 (a) 所示，阴影区域的面积为 1。

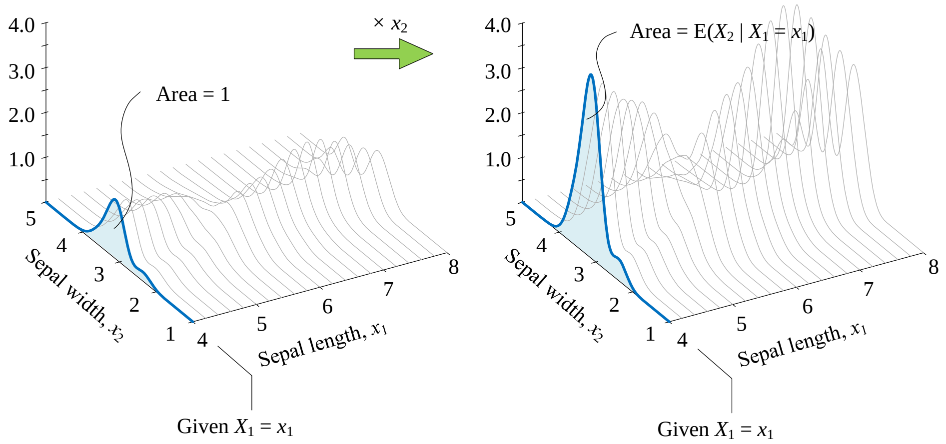


图 24. $f_{X2 \mid X1}(x_2 \mid x_1)$ 条件概率密度三维等高线和平面等高线，不考虑分类

根据 (47)，为了计算条件**期望** $\mathrm{E}(X_2 \mid X_1 = x_1)$，我们需要计算 $x_2 \cdot f_{X2 \mid X1}(x_2 \mid x_1)$ 和 x_2 围成图像的面积，即图 24 (b) 阴影部分面积：

$$\mathrm{E}\left(X_2 \mid X_1 = x_1\right) = \int_{x_2} x_2 \cdot \underbrace{f_{X2|X1}\left(x_2 \middle| x_1\right)}_{\text{Conditional}} \mathrm{d}\, x_2 \tag{52}$$

然后，我们可以计算鸢尾花宽度平方的条件**期望** $\mathrm{E}\left(X_2^2 \mid X_1 = x_1\right)$：

$$\mathrm{E}\left(X_2^2 \mid X_1 = x_1\right) = \int_{x_2} x_2^2 \cdot \underbrace{f_{X2|X1}\left(x_2 \middle| x_1\right)}_{\text{Conditional}} \mathrm{d}\, x_2 \tag{53}$$

然后，可以利用技巧求得条件**方差** $\mathrm{var}(X_2 \mid X_1 = x_1)$：

$$\mathrm{var}\left(X_2 \mid X_1 = x_1\right) = \mathrm{E}\left(X_2^2 \mid X_1 = x_1\right) - \mathrm{E}\left(X_2 \mid X_1 = x_1\right)^2 \tag{54}$$

上式开平方得到，条件均**方差** $\mathrm{std}(X_2 \mid X_1 = x_1)$。

我们知道条件**期望** $\mathrm{E}(X_2 \mid X_1 = x_1)$、条件均**方差** $\mathrm{std}(X_2 \mid X_1 = x_1)$ 都随着 $X_1 = x_1$ 取值变化，而且它们两个单位都是 cm。我们想办法把它们画在一幅图上，具体如图 25 所示。

条件**期望** $\mathrm{E}(X_2 \mid X_1 = x_1)$ 实际上就是“回归”，给定输入条件 $X_1 = x_1$，求 X_2 的输出值。图 25 中黑色实线相当于“回归曲线”。

图 25 还有两条**带宽** (bandwidth)，它们分别代表 $\mu_{X_2|X_1=x_1} \pm \sigma_{X_2|X_1=x_1}$、$\mu_{X_2|X_1=x_1} \pm 2\sigma_{X_2|X_1=x_1}$。带宽随着 $X_1 = x_1$ 移动，条件均**方差**越大，带宽就越宽。

比较图 25、图 26，给定 $X_1 = x_1$ 条件下，X_2 上散点越集中，条件均**方差** $\mathrm{std}(X_2 \mid X_1 = x_1)$ 越小，比如 $X_1 = 7$ cm；相反，X_2 上散点越分散，条件均**方差** $\mathrm{std}(X_2 \mid X_1 = x_1)$ 越大，比如 $X_1 = 5.5$ cm。

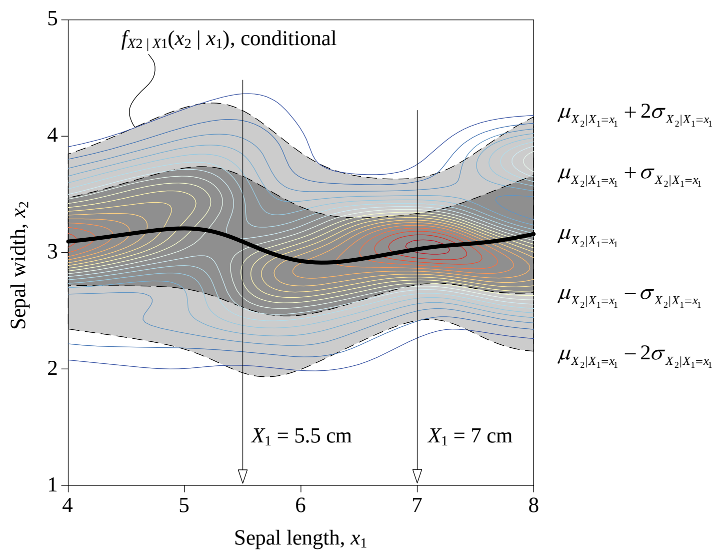


图 25. 条件**期望** $\mathrm{E}(X_2 \mid X_1 = x_1)$、条件均方差 $\mathrm{std}(X_2 \mid X_1 = x_1)$ 之间的关系

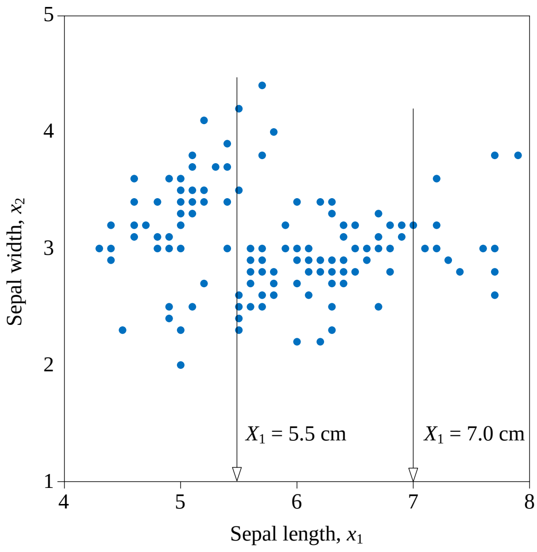


图 26. 鸢尾花数据花萼长度、花萼宽度散点图，不考虑分类

以鸢尾花为例：条件期望 $\mathrm{E}(X_1 \mid X_2 = x_2)$、条件方差 $\mathrm{var}(X_1 \mid X_2 = x_2)$

为了计算条件期望 $\mathrm{E}(X_1 \mid X_2 = x_2)$，我们需要计算 $x_1 \cdot f_{X1 \mid X2}(x_1 \mid x_2)$ 和 x_1 围成图像的面积，即图 27 (b) 阴影部分面积：

$$\mathrm{E}\left(X_1 \mid X_2 = x_2\right) = \int_{-\infty}^{+\infty} x_1 \cdot \underbrace{f_{X1|X2}\left(x_1 \middle| x_2\right)}_{\text{Conditional}} \mathrm{d}\, x_1 \tag{55}$$

然后，我们可以计算鸢尾长度平方的条件期望 $\mathrm{E}\left(X_1^2 \mid X_2 = x_2\right)$：

$$\mathrm{E}\left(X_1^2 \mid X_2 = x_2\right) = \int_{-\infty}^{+\infty} x_1^2 \cdot \underbrace{f_{X1|X2}\left(x_1 \middle| x_2\right)}_{\text{Conditional}} \mathrm{d}\, x_1 \tag{56}$$

然后，可以利用技巧求得条件方差 $\mathrm{var}(X_1 \mid X_2 = x_2)$：

$$\mathrm{var}\left(X_1 \mid X_2 = x_2\right) = \mathrm{E}\left(X_1^2 \mid X_2 = x_2\right) - \mathrm{E}\left(X_1 \mid X_2 = x_2\right)^2 \tag{57}$$

上式开平方得到条件均方差 $\mathrm{std}(X_1 \mid X_2 = x_2)$。

我们知道条件期望 $\mathrm{E}(X_1 \mid X_2 = x_2)$、条件标准差 $\mathrm{std}(X_1 \mid X_2 = x_2)$ 都随着 $X_2 = x_2$ 取值变化，而且它们两个单位都是 cm。我们想办法把它们画在一幅图上，具体如图 28 所示。请大家自己从“回归”角度自行分析图 28。

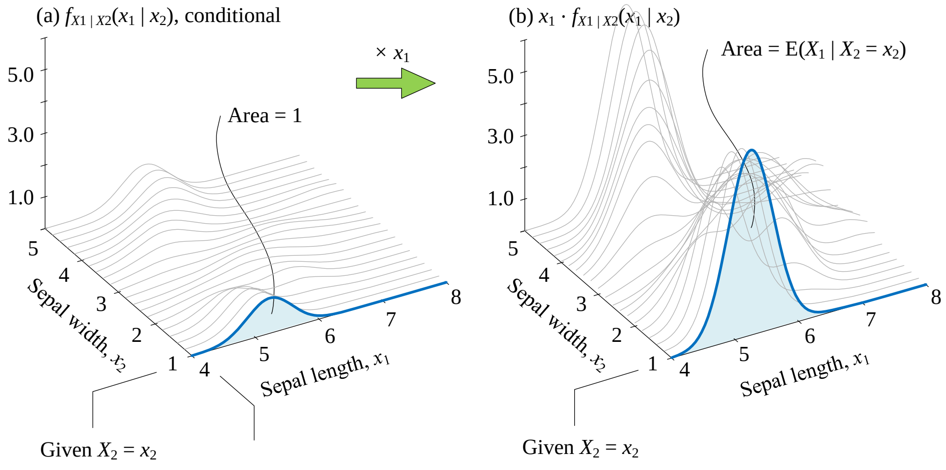


图 27. $f_{X1 | X2}(x_1 | x_2)$ 条件概率密度三维等高线和平面等高线

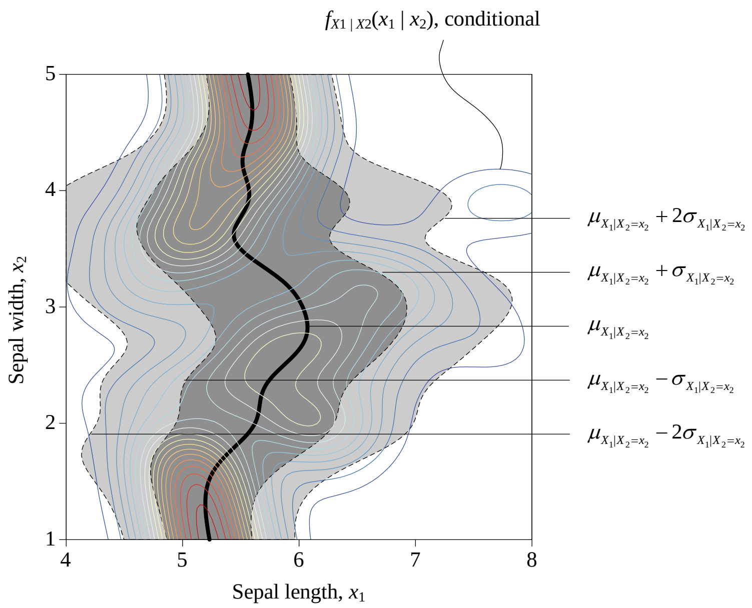


图 28. 条件**期望** E(X_1 | X_2 = x_2)、条件**标准差** std(X_1 | X_2 = x_2) 之间的关系

以鸢尾花为例，考虑标签

同理，我们可以计算给定标签条件下，鸢尾花花萼长度 (图 29)、花萼宽度 (图 30) 的条件**期望**、条件**方差**等。请大家自己完成这几个数值计算。

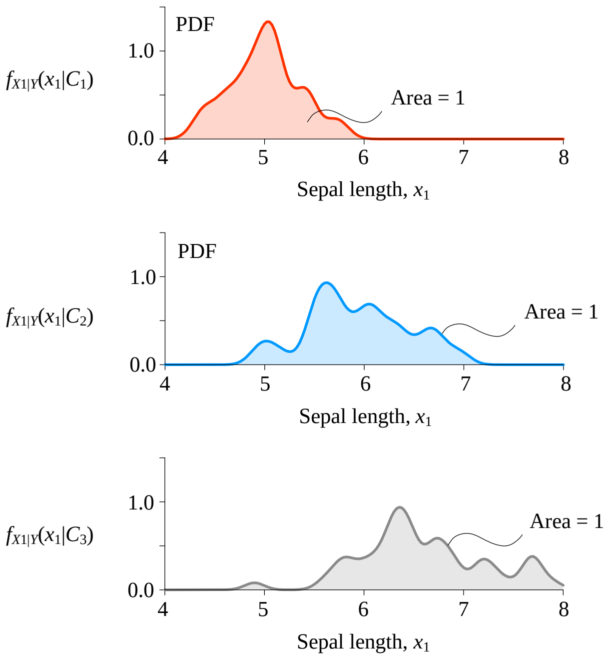


图 29. 给定鸢尾花标签 Y，花萼长度的条件概率密度，连续随机变量

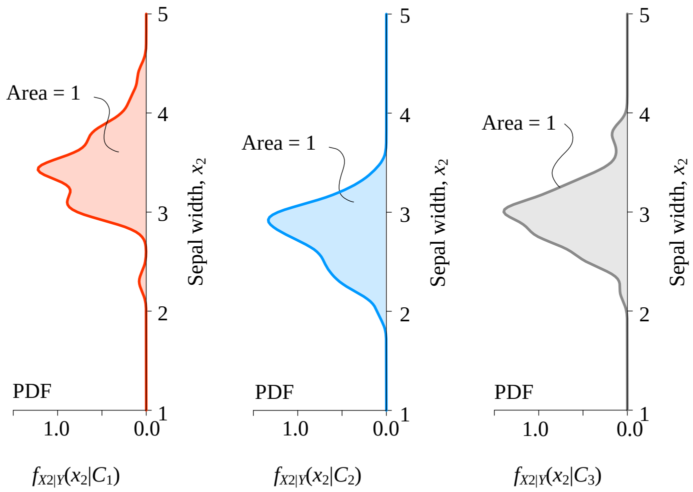


图 30. 给定鸢尾花标签 Y，花萼宽度的条件概率密度，连续随机变量

8.7 再谈如何分割 1

本书前文介绍过，概率分布无非就是各种方式将样本空间概率值 1 进行“切片、切块”、“切丝、切条”。本节从这个视角总结本书这个话题讲解的主要内容。

一元

一元随机变量在一个维度上切割“1”。如果随机变量 X 离散，如图 31 (a) 所示，概率值 1 被分割成若干份，每一份还是“概率”。也就是说一元离散随机变量概率质量函数 PMF $p_X(x)$ 对应概率值。$p_X(x)$ 对应的数学运算是 Σ。图 31 (a) 中所有概率值之和为 1：

$$\sum_{x} p_X(x) = 1 \tag{58}$$

如果随机变量 X 连续，如图 31 (b) 所示，X 则对应概率密度函数 PDF $f_X(x)$。$f_X(x)$ 积分结果才是概率值，因此 $f_X(x)$ 对应的数学运算符为 ∫。$f_X(x)$ 和横轴围成的面积为 1，对应样本空间概率值“1”：

$$\int_{x} f_X(x) = 1 \tag{59}$$

图 31 (b) 中连续随机变量 X 的取值范围是实数轴的一个区间。图 31 (c) 中连续随机变量 X 的取值范围是整个实数轴。图 31 (c) 中，$f_X(x)$ 和整个横轴围成的面积为 1。

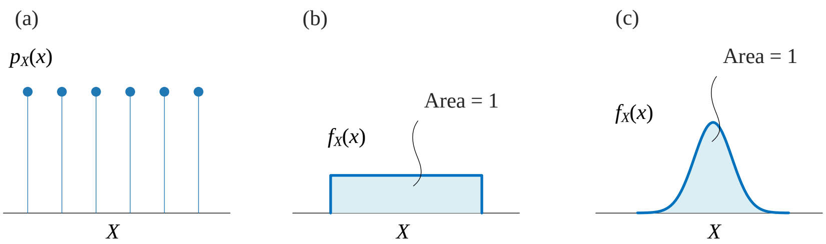


图 31. 一元随机变量

二元

二元随机变量 (X_1, X_2) 在两个维度上对样本空间进行分割。

如图 32 (a) 所示，如果 X_1 和 X_2 都是离散随机变量，概率质量函数 $p_{X1,X2}(x_1, x_2)$ 本身还是概率值。$p_{X1,X2}(x_1, x_2)$ 二重求和的结果为 1：

$$\sum_{x_1}\sum_{x_2} p_{X1,X2}(x_1, x_2) = 1 \tag{60}$$

大家试图调换求和顺序时，要格外小心。并不是所有的多重求和都可以任意调换求和先后顺序。

而 $p_{X1,X2}(x_1, x_2)$ 偏求和便得到边缘概率质量函数 $p_{X1}(x_1)$、$p_{X2}(x_2)$：

$$\begin{aligned}\sum_{x_2} p_{X1,X2}(x_1, x_2) &= p_{X1}(x_1) \\ \sum_{x_1} p_{X1,X2}(x_1, x_2) &= p_{X2}(x_2)\end{aligned} \tag{61}$$

如图 33 所示，二元随机变量偏求和将某个变量“消去”，这相当于折叠。

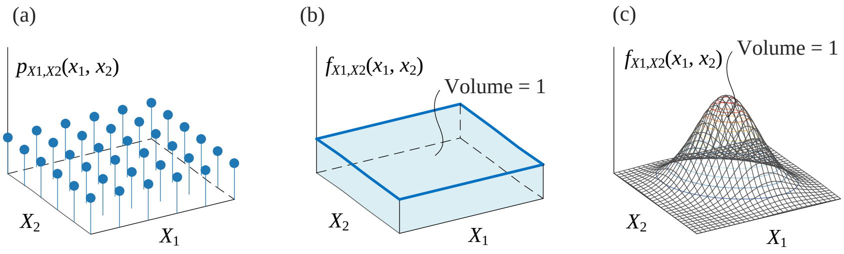


图 32. 二元随机变量

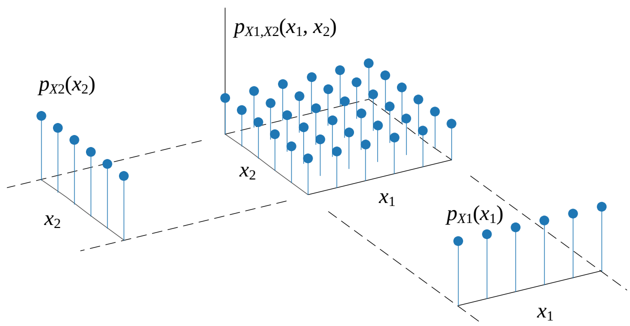


图 33. 二元随机变量偏求和，折叠某一变量

如图 32 (b) 所示，如果 X_1 和 X_2 都是连续随机变量，概率密度函数 $f_{X1,X2}(x_1, x_2)$ 如下二重积分的结果为 1：

$$\int_{x_2}\int_{x_1} f_{X1,X2}\left(x_1, x_2\right) \mathrm{d}\, x_1 \,\mathrm{d}\, x_2 = 1 \tag{62}$$

这相当于图 32 (b) 中几何体和水平面围成的几何图形的体积为 1。如图 32 (c) 所示，X_1 和 X_2 的取值范围也可以是整个水平面，即 $\mathbb{R}^2$ 。

$f_{X1,X2}(x_1, x_2)$ 偏积分边缘概率密度函数 $f_{X1}(x_1)$、$f_{X2}(x_2)$：

$$\begin{aligned} \int_{x_2} f_{X1,X2}\left(x_1, x_2\right) &= f_{X1}\left(x_1\right) \\ \int_{x_1} f_{X1,X2}\left(x_1, x_2\right) &= p_{X2}\left(x_2\right) \end{aligned} \tag{63}$$

三元

图 34 (a) 中 (X_1, X_2, X_3) 三个随机变量都是离散随机变量，每个点 (x_1, x_2, x_3) 处都有一个概率值，这些概率值可以写成概率质量函数 $p_{X1,X2,X3}(x_1, x_2, x_3)$ 这种形式。

请大家自己写出如何根据 $p_{X1,X2,X3}(x_1, x_2, x_3)$ 计算 $p_{X1,X2}(x_1, x_2)$、$p_{X1}(x_1)$。

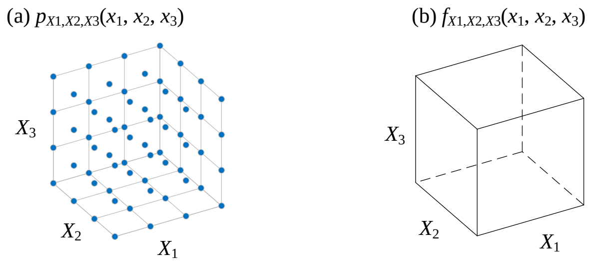


图 34. 三元随机变量

图 34 (b) 中 (X_1, X_2, X_3) 三个随机变量都是连续随机变量，整个 $\mathbb{R}^3$ 空间中的每一点 (x_1, x_2, x_3) 处都有一个概率密度值 $f_{X1,X2,X3}(x_1, x_2, x_3)$。这就是本书前文提到的"体密度"。也请大家自己写出如何根据 $f_{X1,X2,X3}(x_1, x_2, x_3)$ 计算 $f_{X1,X2}(x_1, x_2)$、$f_{X1}(x_1)$。

图 35 所示为在 X_3 取不同值时 $X_3 = c$，概率密度值 $f_{X1,X2,X3}(x_1, x_2, c)$ "切片"。强调一下，图 35 中 X_3 还是连续随机变量。

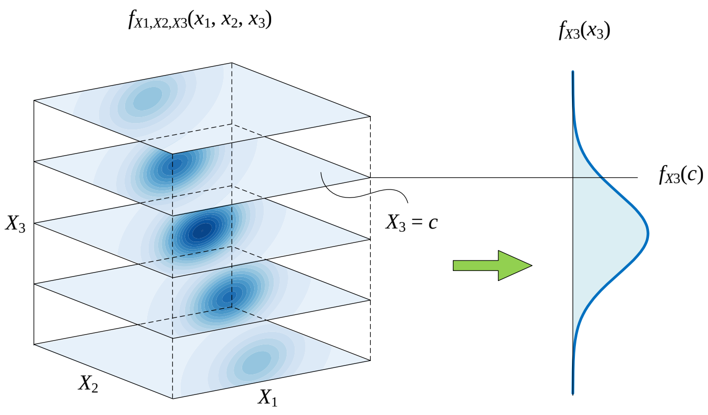


图 35. 三个随机变量都是连续随机变量

$f_{X1,X2,X3}(x_1, x_2, c)$ 这个"切片"对 x_1 和 x_2 二重积分得到的是边缘概率密度 $f_{X3}(c)$：

$$\int_{x_2}\int_{x_1} f_{X1,X2,X3}(x_1, x_2, c)\,\mathrm{d}x_1\,\mathrm{d}x_2 = f_{X3}(c) \tag{64}$$

上式相当于，我们不再关心图 35 中这些切片的具体等高线，而是将其归纳为一个数值。

混合

此外，多元随机变量还可以是离散和随机变量的混合形式。一个最简单的例子就是鸢尾花数据。如图 36 所示，分类标签将鸢尾花数据分成了三层，对应 C_1、C_2、C_3 三个标签。图 36 左侧的数据构成了样本空间 Ω。显然 C_1、C_2、C_3 互不相容，形成对样本空间 Ω 的分割。这体现的就是本书第 3 章讲过的全概率定理。

花萼长度 X_1、花萼宽度 X_2 都是连续随机变量，但是标签 Y 为离散随机变量。

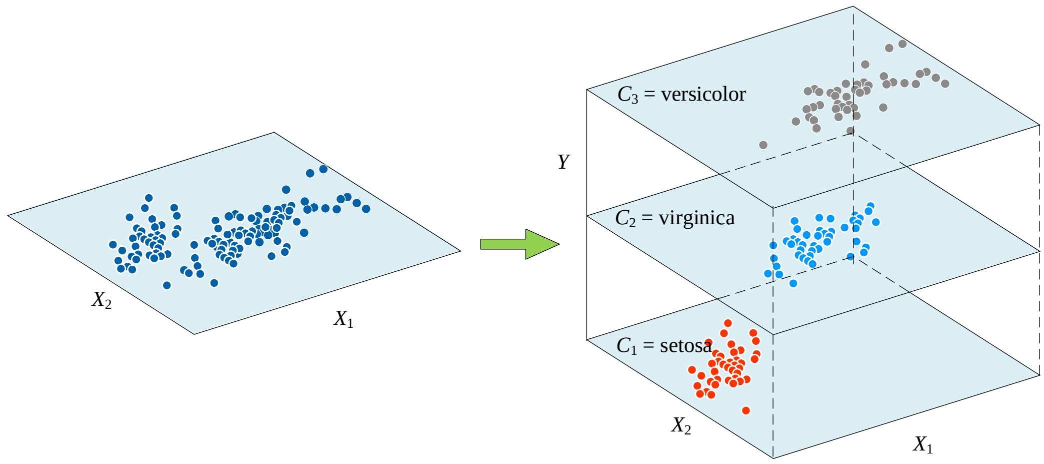


图 36. 分类标签将鸢尾花数据分层

如图 37 所示，每一类不同标签的样本数据都有其联合概率密度分布 $f_{X1,X2,Y}(x_1, x_2, C_1)$、$f_{X1,X2,Y}(x_1, x_2, C_2)$、$f_{X1,X2,Y}(x_1, x_2, C_3)$。

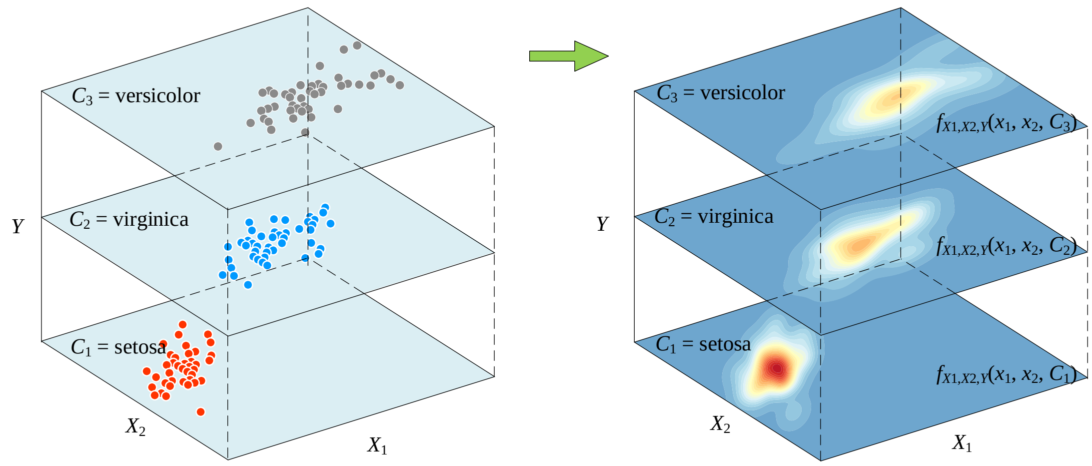


图 37. 鸢尾花数据，花萼长度 X_1、花萼宽度 X_2、标签 Y

图 38 所示为两个不同方向压扁 $f_{X1,X2,Y}(x_1, x_2, y)$。

$f_{X1,X2,Y}(x_1, x_2, C_1)$、$f_{X1,X2,Y}(x_1, x_2, C_2)$、$f_{X1,X2,Y}(x_1, x_2, C_3)$ 这三个平面分别二重积分得到 Y 的边缘概率密度：

$$\begin{gathered}\int_{x_2}\int_{x_1} f_{X1,X2,Y}\left(x_1,x_2,C_1\right)\mathrm{d}x_1\,\mathrm{d}x_2 = p_Y\left(C_1\right)\\ \int_{x_2}\int_{x_1} f_{X1,X2,Y}\left(x_1,x_2,C_2\right)\mathrm{d}x_1\,\mathrm{d}x_2 = p_Y\left(C_2\right)\\ \int_{x_2}\int_{x_1} f_{X1,X2,Y}\left(x_1,x_2,C_3\right)\mathrm{d}x_1\,\mathrm{d}x_2 = p_Y\left(C_3\right)\end{gathered} \tag{65}$$

显然，$p_Y(C_1)$、$p_Y(C_2)$、$p_Y(C_3)$之和为 1。

沿着 Y 方向将 $f_{X1,X2,Y}(x_1, x_2, y)$ 压扁得到 $f_{X1,X2,Y}(x_1, x_2)$：

$$f_{X1,X2}(x_1,x_2)=f_{X1,X2,Y}(x_1,x_2,C_1)+f_{X1,X2,Y}(x_1,x_2,C_2)+f_{X1,X2,Y}(x_1,x_2,C_3) \tag{66}$$

而 $f_{X1,X2,Y}(x_1, x_2)$ 和水平面构成的几何形体的体积为 1，即：

$$\int_{x_2}\int_{x_1} f_{X1,X2}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2=1 \tag{67}$$

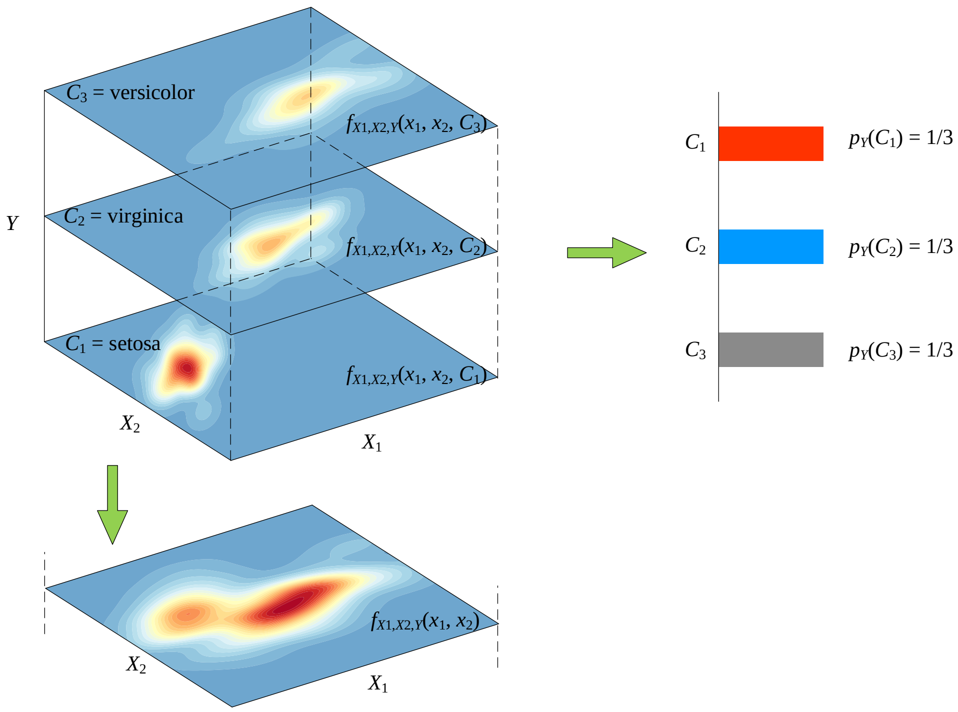


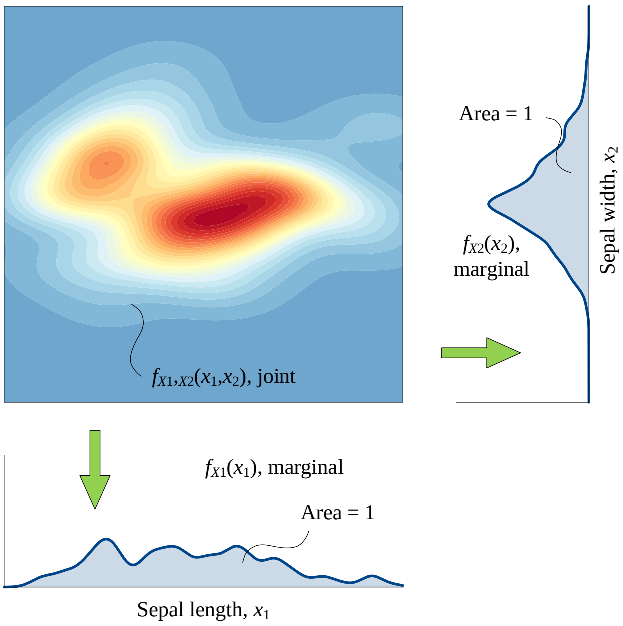
图 38. 两个不同方向压扁 $f_{X1,X2,Y}(x_1, x_2, y)$

此外，$f_{X1,X2}(x_1, x_2)$ 可以沿着不同方向进一步"压扁"得到边缘概率 $f_{X1}(x_1)$、$f_{X2}(x_2)$：

$$\begin{aligned}\int_{x_2} f_{X1,X2}(x_1,x_2)\,\mathrm{d}x_2 &= f_{X1}(x_1)\\ \int_{x_1} f_{X1,X2}(x_1,x_2)\,\mathrm{d}x_1 &= f_{X2}(x_2)\end{aligned} \tag{68}$$

$f_{X1}(x_1)$、$f_{X2}(x_2)$ 和 x_1、x_2 轴围成的面积也都是 1：

$$\begin{aligned}\int_{x_1} f_{X1}(x_1)\,\mathrm{d}x_1 &= 1\\ \int_{x_2} f_{X2}(x_2)\,\mathrm{d}x_2 &= 1\end{aligned} \tag{69}$$



总结来说，以上几种情况无非就是对 1 的“切片、切块”、“切丝、切条”。

此时，希望大家闭上眼睛想 $f_{X1,X2,Y}(x_1, x_2, C_1)$、$f_{X1,X2}(x_1, x_2)$ 的时候看到的是等高线，想 $f_{X1}(x_1)$ 看到曲线，想 $p_Y(C_1)$ 的时候看到一个数值 (1/3)。

不同的混合形式

图 39 所示为二元随机变量的不同离散、连续混合形式。图 39 (a) 两个随机变量都是连续。图 39 (b) 中 X_1 为离散随机变量，X_2 为连续随机变量；图 39 (c) 反之。图 39 (d) 中，两个随机变量都是离散随机变量。图 40 所示为三元随机变量的不同离散、连续混合形式，请大家自己分析其中子图。这实际上回答了本书第 4 章提出的问题。

此外，在本书贝叶斯推断中，大家会发现我们不再区分 PDF、PMF，概率分布函数全部统一为 $f()$。

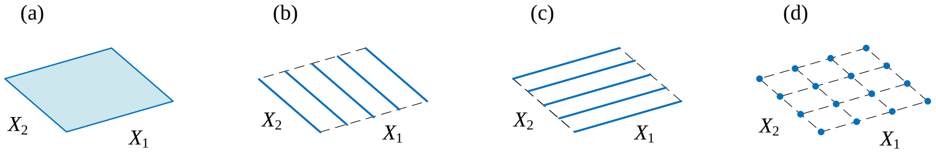


图 39. 二元元随机变量，混合

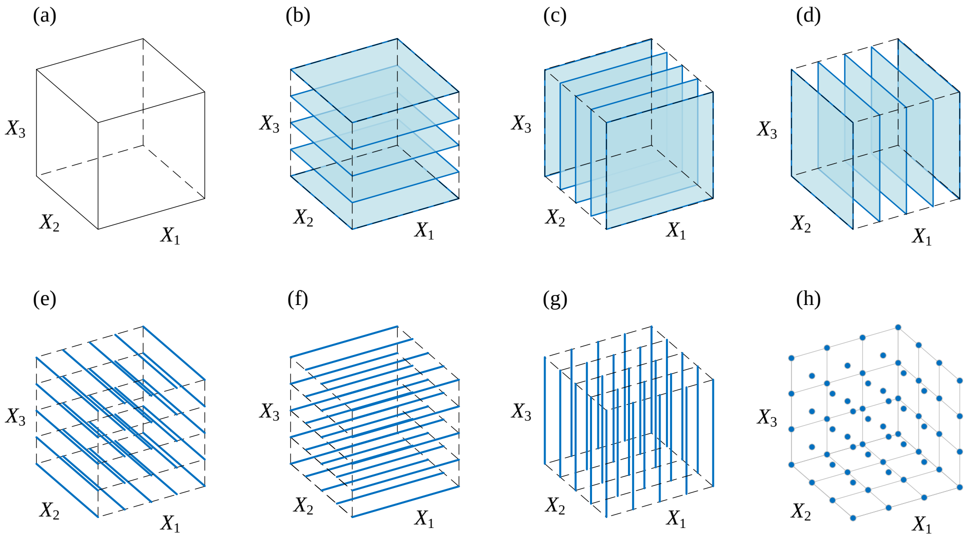


图 40. 三元随机变量，混合

条件概率：重新定义 1

条件概率其实很好理解，条件概率的"条件"就是"新的样本空间"，对应概率值为 1。也就是把从原始样本空间中切出来的"一片、一块、一丝、一条"作为新的样本空间。

如图 41 所示，给定标签为 $Y = C_2$ 条件下，利用贝叶斯定理，条件概率可以通过下式求得：

$$f_{X1,X2|Y}\left(x_1,x_2\,|\,C_2\right)=\frac{f_{X1,X2,Y}\left(x_1,x_2,C_2\right)}{p_Y\left(C_2\right)} \tag{70}$$

分母中的 $p_Y(C_2)$ 起到归一化的作用。也就是说，$f_{X1,X2|Y}\left(x_1,x_2\,|\,C_2\right)$ 二重积分的结果为 1：

$$\int_{x_2}\int_{x_1} f_{X1,X2|Y}\left(x_1,x_2\,|\,C_2\right)\mathrm{d}\,x_1\,\mathrm{d}\,x_2=1 \tag{71}$$

上式中这个"1"对应条件概率 $f_{X1,X2|Y}\left(x_1,x_2\,|\,C_2\right)$ 的条件——$Y = C_2$。$Y = C_2$ 就是这个条件概率的"新样本空间"。

本书第 6 章还介绍过，以鸢尾花花萼长度或宽度为条件的条件概率，请大家回顾。

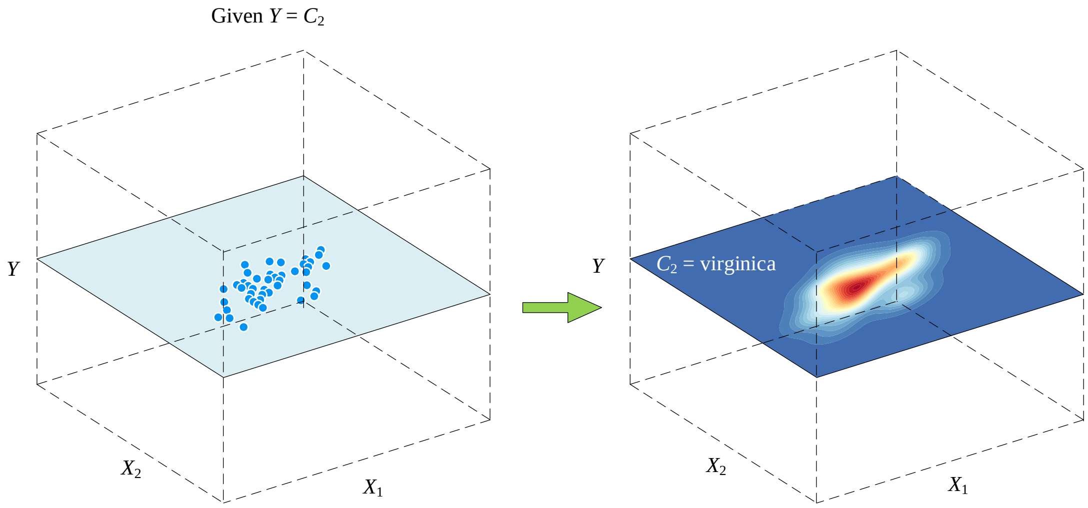


图 41. 条件概率，给定标签为 $Y = C_2$

条件期望是指在已知一些条件下，一个随机变量的期望值。同理，条件方差是指在给定某些条件下，随机变量的方差。它俩表示给定某些信息或事件之后，对随机变量的期望、方差的预测或估计。其实生活中条件期望、方差无处不在，大家多多留意。条件期望、方差在概率论、统计学和经济学等领域有广泛的应用，例如在回归分析、决策树、贝叶斯推断等中。