

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

## 2.3 期望



### 本节你将掌握的核心技能：

- ▶ 期望是用一个数来概括随机变量整体平均趋势的位置；
- ▶ 期望将所有可能取值按照概率进行加权；
- ▶ 常数的期望等于其自身，因为它没有任何随机性；
- ▶ 对随机变量整体平移不会改变其形状，只会让期望随之平移；
- ▶ 随机变量按比例缩放或翻转时，其期望也按相同比例变化；
- ▶ 两个随机变量相加后的期望等于它们期望的相加，与是否独立无关。

### 期望

在研究随机变量时，我们常常希望用一个简单的数来概括这个随机变量“整体上”会落在什么位置。**期望** (expectation)，也叫**期望值** (expected value)，就是用来完成这一任务的工具。

**期望** 把一个随机变量所有可能的取值“压缩”为一个能够代表其平均趋势的数。

假设离散随机变量  $X$  有  $n$  个取值  $\{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$ ，每个取值出现的概率由概率质量函数  $p_X(x)$  描述。

期望值  $E(X)$  的定义可以理解为：用“出现的概率”作为权重，对所有可能的取值求一个加权平均。形式化地写为：

$$E(X) = \mu_X = \underbrace{x^{(1)}}_{\text{Scalar}} p_X(x^{(1)}) + x^{(2)} p_X(x^{(2)}) + \dots + x^{(n)} p_X(x^{(n)}) = \sum_{i=1}^n x^{(i)} \cdot \underbrace{p_X(x^{(i)})}_{\text{Weight}} \quad (1)$$

为了方便，我们经常把 (1) 简写作：

$$E(X) = \sum_x x \cdot p_X(x) \quad (2)$$

这里的  $\sum_x (\cdot)$  表示对随机变量所有可能的取值进行遍历并求和，也可以理解成“穷举所有结果”。

由于所有概率的总和必须等于 1，PMF  $p_X(x)$  自然满足

$$\sum_x p_X(x) = 1 \quad (3)$$

因此，期望值确实是一个严格意义上的加权平均。

从更直观的角度来看，期望是对随机变量的一种“降维操作”。原本随机变量  $X$  可能有很多取值，但经过运算符  $E(\cdot)$  之后，这些复杂的可能性被折叠成一个代表其整体水平的数。

对于多元随机变量来说，期望向量往往也被称为分布的“质心”，因为它在几何上起到类似平衡点或中心位置的作用。。

## 质量均匀的颜色子

举个例子，当我们研究一个均匀色子的随机性时，最自然的问题之一是：掷一次色子，它“平均上”会落在什么位置？

为了回答这个问题，我们将掷色子的点数视为一个离散随机变量  $X$ 。对于一枚质量均匀、六个面概率相同的色子来说， $X$  的所有可能取值为 1、2、3、4、5、6，每个点出现的概率都是  $1/6$ 。

在这种最典型的均匀分布情形下， $X$  的期望值计算非常简单，就是用所有取值按相同的概率进行平均：

$$E(X) = \sum_x x \cdot \underbrace{p_X(x)}_{\text{Weight}} = \sum_x x \cdot \frac{1}{6} = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5 \quad (4)$$

从几何直觉上理解，期望值可以看作分布在数轴上位置的“质心”。

虽然色子只会出现整数点数，但它们的概率分布在数轴上形成了一个离散但均衡的结构，而期望值 3.5 正好落在这个结构的几何中心位置。

图 1 展示了六面均匀色子的 PMF 以及由此得到的期望值所在位置，从图中也可以直观看出，这个中心位于 3 与 4 的正中间。

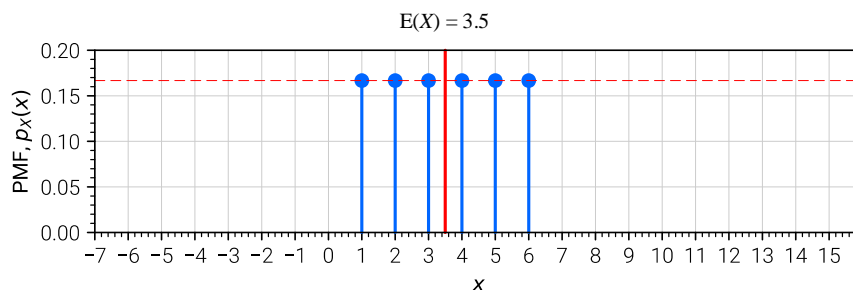


图 1. 掷均匀色子试验 PMF 和期望值位置

进一步，如果进行重复试验，我们可以观察到一个重要的统计现象。图 2 展示了随着掷色子次数的增加，试验结果的算术平均值如何变化。

起初，由于样本数量很少，平均值会出现较大的随机波动；但随着试验次数越来越多，这个平均值会逐渐稳定，并最终逼近理论值 3.5。这一现象体现了大数定律的思想：真实世界中的随机过程，在大量重复之后，其平均行为会与理论期望越来越接近。

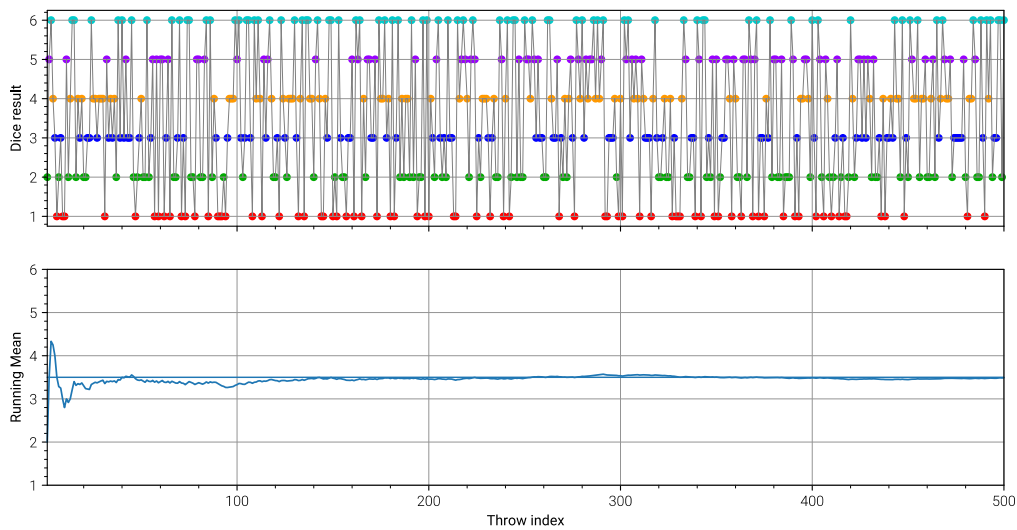


图 2. 投色子试验均值随试验次数变化

## 常数

在概率论中，期望值不仅适用于具有多个可能取值的随机变量，也可以应用于常数这种特殊情况。所谓常数，是指在每一次试验中取值都固定不变的量。例如，如果随机变量  $X$  在所有情况下都取同一个值  $c$ ，那么  $X$  就是一个常数。

常数的期望值非常直观：因为它无论如何变化，结果总是  $c$ ，所以它的平均值自然也是  $c$ 。用公式表示就是：

$$E(c) = c \quad (5)$$

可以从具体例子中更容易理解这个概念。如果一个随机试验，每次结果都是 3，无论我们进行多少次实验，算术平均值始终是 3，因此  $E(3) = 3$ 。

同理，一个随机试验，每次结果都是 -3，那么它的期望值也就是 -3，即  $E(-3) = -3$ 。

从直觉上看，常数的期望值与加权平均的概念完全一致。对于一般的随机变量，期望值是所有可能取值按概率加权后的平均；而常数的所有“可能取值”只有一个，并且概率为 1，因此加权平均自然等于这个唯一取值。

## 加常数

对随机变量加上一个常数，期望值也会相应地增加同样的常数。

换句话说，如果有随机变量  $X$  和一个常数  $c$ ，定义新的随机变量  $Y = X + c$ ，那么  $Y$  的期望值满足：

$$E(Y) = E(X + c) = E(X) + c \quad (6)$$

为了让这个性质更直观，可以通过掷均匀色子的例子理解。假设  $X$  表示掷一枚六面均匀色子的点数，已知  $E(X) = 3.5$ 。

如果我们定义一个新的随机变量  $Y = X + 3$ ，也就是每次掷出的点数都加 3，那么根据加常数的期望性质， $Y$  的期望值就是：

$$E(Y) = E(X + 3) = E(X) + 3 = 6.5 \quad (7)$$

和图 1 相比，图 3 展示了  $X$  加上常数 3 后的新随机变量  $Y$  的分布以及期望位置，可以直观地看到整个分布向右平移了 3 个单位，同时期望值（红色竖线）也随之平移。

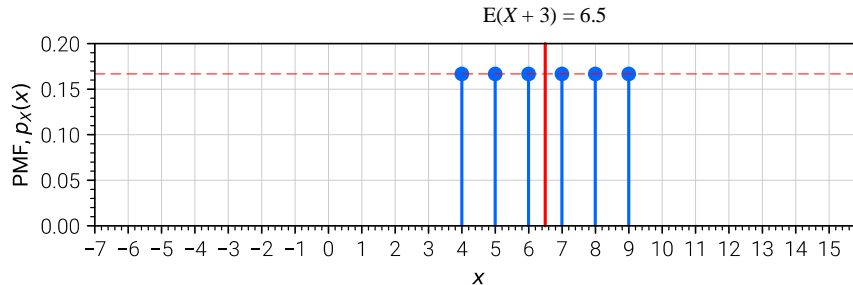


图 3. 掷均匀色子试验色子点数为  $X$ ,  $X+3$  的期望

同样地，如果我们定义  $Y = X - 3$ ，也就是每次掷出的点数减少 3，那么新的期望值就是：

$$E(Y) = E(X - 3) = E(X) - 3 = 0.5 \quad (8)$$

和图 1 相比，图 4 显示了这个减法操作后的结果，分布整体向左平移了 3 个单位，期望值（红色竖线）也相应降低。

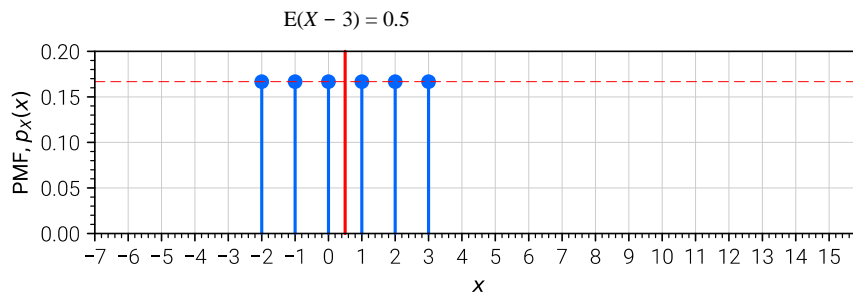


图 4. 掷均匀色子试验色子点数为  $X$ ,  $X-3$  的期望

直观地理解，加常数的操作不会改变随机变量的波动范围或分布形状，只会整体平移分布位置，期望值正是描述分布中心位置的数，因此会随常数的加减而平移。

## 倍数

设有随机变量  $X$  和一个常数  $a$ ，定义新的随机变量  $Y = aX$ ，那么  $Y$  的期望值满足线性关系：

$$E(aX) = aE(X) \quad (9)$$

这一性质在直观上很容易理解：如果随机变量的每一个可能取值都按相同比例放大或缩小，那么它们的平均位置也会按相同比例放大或缩小。

以掷均匀色子为例，假设  $X$  表示色子的点数，已知  $E(X) = 3.5$ 。如果定义  $Y = 2X$ ，也就是将每次掷出的点数乘以 2，那么新的期望值为：

$$E(Y) = E(2X) = 2E(X) = 7 \quad (10)$$

图5展示了掷色子实验中  $X$  的分布，以及每个点数放大两倍后的分布和期望位置。和图1相比，可以直观地看到，整个分布沿数轴正方向拉伸了一倍，期望值也随之翻倍。

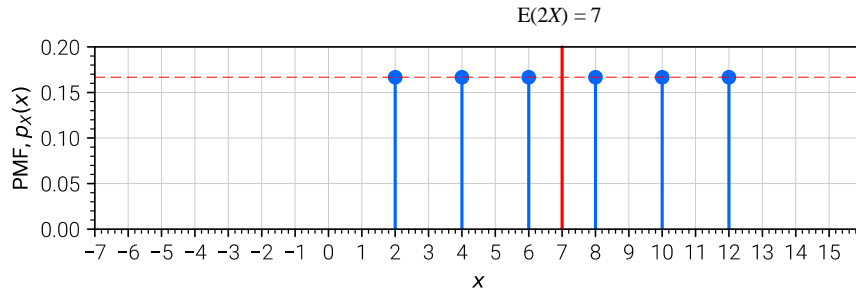


图5. 掷均匀色子试验色子点数为  $X$ ， $2X$  的期望

同样地，如果定义  $Y = -X$ ，也就是将点数取负，那么期望值会随之取相反数：

$$E(Y) = E(-X) = -E(X) = 3.5 \quad (11)$$

和图1相比，图6展示了这个操作后的效果，分布关于零点对称翻转，期望值也随之翻转。

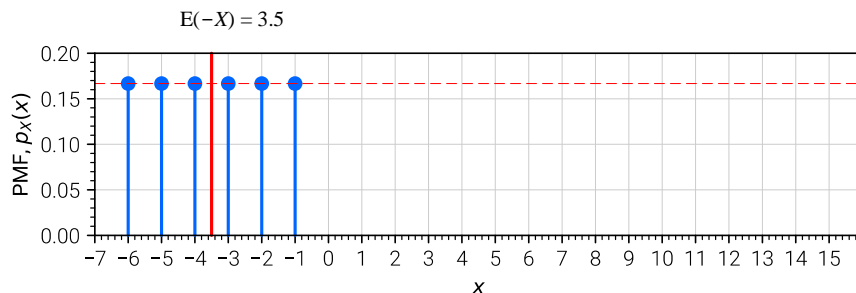


图6. 掷均匀色子试验色子点数为  $X$ ， $-X$  的期望

这一性质在统计学和机器学习中非常重要。比如，在特征缩放、线性回归模型的系数调整，或者数据标准化中，理解期望随倍数变换的线性规律，有助于我们快速预测变换后数据的中心位置而无需重新计算每个样本的加权平均值。

## 倍数 + 常数

前文提过，期望值具有线性性质，这意味着对随机变量同时进行乘法和加法操作时，期望值也服从一样的运算规律

$$E(aX + c) = aE(X) + c \quad (12)$$

这一公式可以理解为，先将随机变量  $X$  的平均位置按倍数  $a$  拉伸或翻转，然后整体平移  $c$  个单位。无论随机变量本身如何分布，这条规律始终成立。

以掷均匀色子为例，设  $X$  表示色子点数，已知  $E(X) = 3.5$ 。如果定义新的随机变量

$$Y = 2X + 3 \quad (13)$$

那么  $Y$  的期望值为

$$E(Y) = 2E(X) + 3 = 10 \quad (14)$$

图 7 展示了对  $X$  进行线性变换  $2X + 3$  后的新分布和期望位置。从图中可以直观地看到，原始分布被拉伸为原来的两倍，并整体向右平移了 3 个单位，期望值也随之移动到 10。

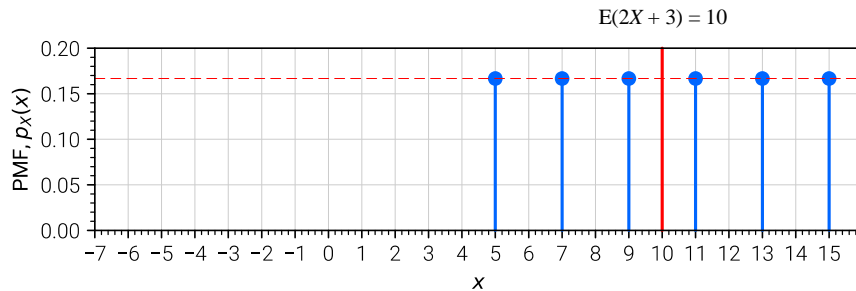


图 7. 掷均匀色子试验色子点数为  $X$ ,  $2X + 3$  的期望

## 求和

当我们研究两个随机变量  $X$  和  $Y$  时，一个非常常见的任务是理解它们的和  $X + Y$  的行为。在许多场景中，这种“求和”操作具有明确的物理或统计含义。例如，两个独立误差的叠加、或两次试验结果的累加，都是对随机变量求和的典型情况。

期望值的一个核心性质告诉我们：求和后的期望值等于两个期望值的相加，即

$$E(X + Y) = E(X) + E(Y) \quad (15)$$

**⚠ 注意**，这一性质与随机变量是否独立无关，因此非常普遍，也非常重要。

为了让读者理解其合理性，可以给出一个简单直观的证明思路。按照期望的定义，对所有可能的取值求加权平均：

$$\begin{aligned} E(X + Y) &= \sum_x \sum_y (x + y) f(x, y) \\ &= \sum_x \sum_y x f(x, y) + \sum_x \sum_y y f(x, y) \\ &= \sum_x x \sum_y f(x, y) + \sum_y y \sum_x f(x, y) \\ &= \sum_x x f_X(x) + \sum_y y f_Y(y) \\ &= E(X) + E(Y) \end{aligned} \quad (16)$$

第一个求和式  $\sum_x \sum_y x f(x, y)$  是对  $X$  按联合概率求加权平均，第二个求和式  $\sum_x \sum_y y f(x, y)$  是对  $Y$  按联合概率求加权平均。

经过整理之后，我们发现这正是  $E(X)$  和  $E(Y)$  的定义，因此结论成立。这个分解过程揭示了求和期望的本质：平均值的叠加可以逐项拆分。

为了让这一性质更具直观性，可以通过掷两颗均匀色子的例子进行说明。

设  $X$  和  $Y$  分别表示第一颗和第二颗色子的点数。因为每颗色子的期望都是 3.5，所以根据求和性质，我们立即得到：

$$E(X+Y) = E(X) + E(Y) = 3.5 + 3.5 = 7 \quad (17)$$

图 8 展示了两颗色子的联合样本空间  $(X, Y)$ ，它包含 36 个等可能的组合。也就是说，图 8 中每个点被取到的概率为  $1/36 (= 1/6 \times 1/6)$ 。

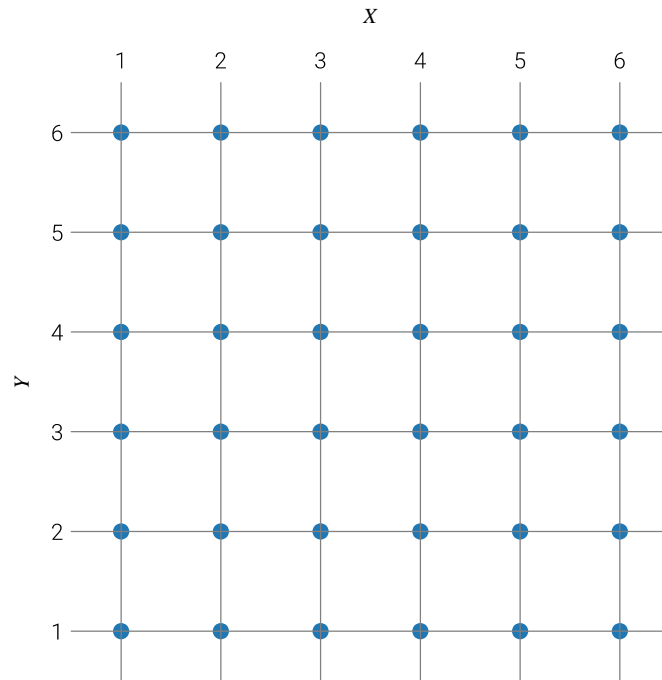
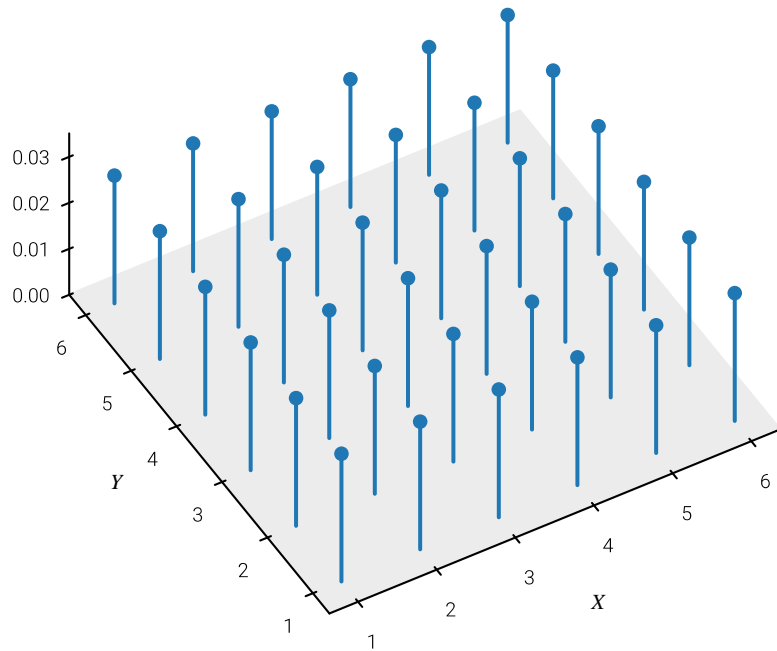


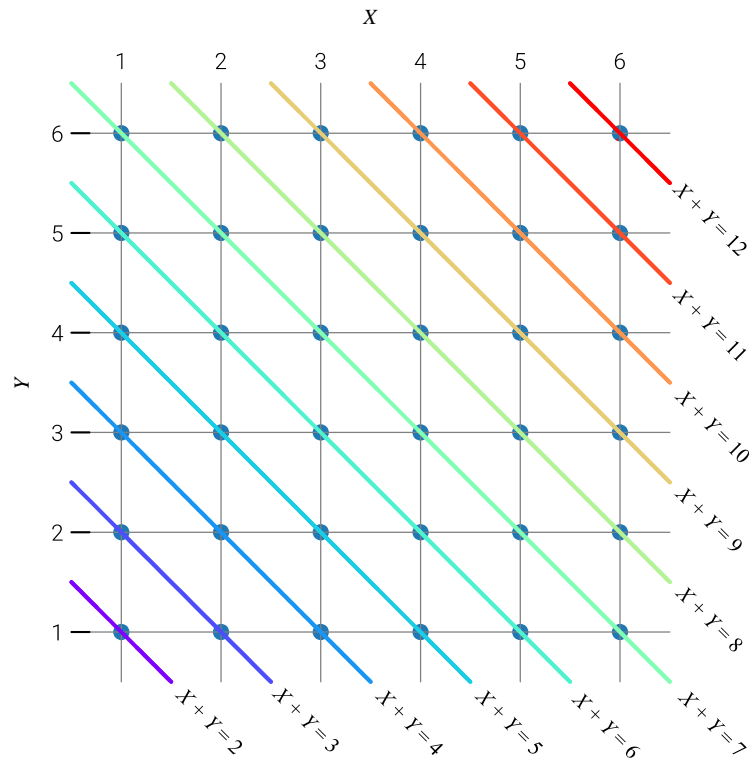
图 8. 掷两颗均匀色子试验色子点数分别为  $X$  和  $Y$ ， $(X, Y)$  样本空间

图 10 中，每一个蓝色点都代表两颗色子点数的一个组合  $(X, Y)$ 。这些点根据  $X + Y$  的值被分布在不同的等高线上，而每一条等高线对应一个固定的和值。

例如，等高线最左下侧对应和值为 2，中间位置对应和值为 7，最右上侧对应和值为 12。由于两颗色子各有 6 种点数，因此整体样本空间由 36 个等可能的点构成，可以用图 9 所示的三维火柴梗可视化。

图 9. 掷两颗均匀色子试验色子点数分别为  $X$  和  $Y$ ,  $(X, Y)$  样本空间的三维火柴梗图

观察这些等高线，可以直观地看出某些和值出现的可能性更高。比如，等高线对应和值为 7 的那一行会包含最多蓝点，它的组合方式最多， $(1, 6)$ 、 $(2, 5)$ 、 $(3, 4)$ 、 $(4, 3)$ 、 $(5, 2)$ 、 $(6, 1)$ ，因此出现概率也最大；而和值为 2 或 12 的等高线只包含一个蓝点，因此概率最小。

图 10. 掷两颗均匀色子试验色子点数分别为  $X$  和  $Y$ ,  $X + Y$  的样本空间



基于这种观察，我们可以很容易计算出  $S = X + Y$  的概率质量函数。写成分段函数形式

$$p_s(s) = \begin{cases} \frac{s-1}{36}, & s = 2, 3, 4, 5, 6, 7 \\ \frac{13-s}{36}, & s = 8, 9, 10, 11, 12 \end{cases} \quad (18)$$

利用 (18)，我们也可以计算  $S = X + Y$  的期望值

$$\begin{aligned} E(X+Y) = E(S) &= \sum_{s=2}^{12} s \cdot p_s(s) = 2 \times \frac{1}{36} + 3 \times \frac{2}{36} + 4 \times \frac{3}{36} + \cdots + 10 \times \frac{3}{36} + 11 \times \frac{2}{36} + 12 \times \frac{1}{36} \\ &= 7 \end{aligned} \quad (19)$$

答案和 (17) 一致。

图 11 所示为  $X + Y$  的概率分布和期望。可以看到，虽然  $X$  和  $Y$  的分布是离散均匀的，但  $X + Y$  的分布呈三角形，7 是出现概率最高的和值，而期望值也恰好位于这个分布的中心位置。

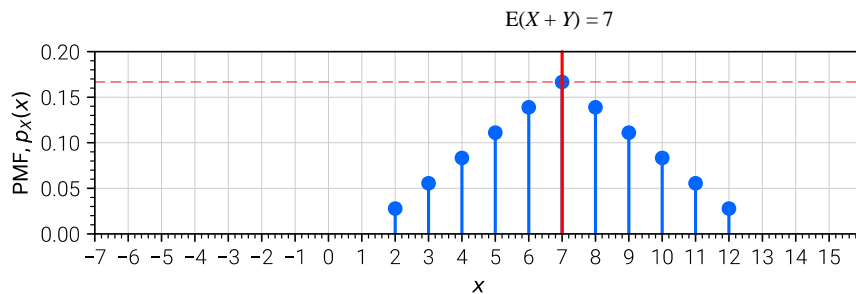


图 11. 掷两颗均匀色子试验色子点数分别为  $X$  和  $Y$ ， $X + Y$  的分布和期望

## 线性组合

在前面讨论了“加上常数”“乘以倍数”以及“多个随机变量相加”的期望性质之后，我们可以把这些结论进一步整合成一个更一般、也更常见的情形：**随机变量的线性组合** (linear combination of random variables)。

许多统计学与机器学习模型（如线性回归、主成分分析、因子分析等算法）都涉及形如

$$Y = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n = \sum_{i=1}^n a_i X_i \quad (20)$$

如图 12 所示，这种组合方式被称为随机变量的线性组合。

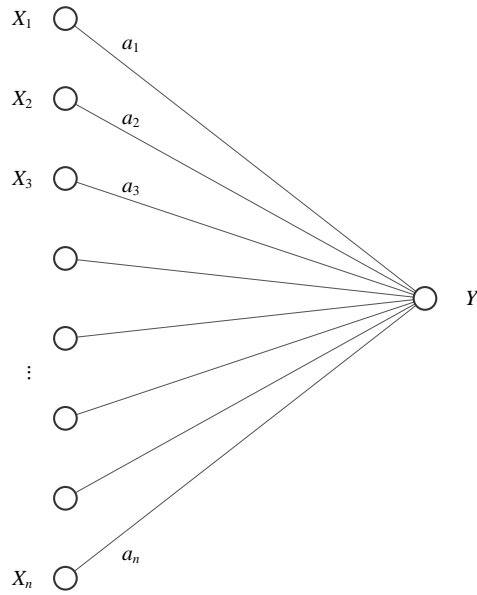


图 12. 随机变量的线性组合

线性组合的期望有一个非常重要且优雅的性质：无论这些随机变量是否独立，只要各自的期望存在，就必然有

$$E(Y) = E(a_1 X_1 + a_2 X_2 + \cdots + a_n X_n) = a_1 E(X_1) + a_2 E(X_2) + \cdots + a_n E(X_n) \quad (21)$$

即

$$E\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i E(X_i) \quad (22)$$

特别地，当  $n = 2$  时，上式可以写成：

$$E(a_1 X_1 + a_2 X_2) = a_1 E(X_1) + a_2 E(X_2) \quad (23)$$

此外，引入常数项，(21) 可以写成

$$E(Y) = E(a_1 X_1 + a_2 X_2 + \cdots + a_n X_n + c) = a_1 E(X_1) + a_2 E(X_2) + \cdots + a_n E(X_n) + c \quad (24)$$

即

$$E\left(\sum_{i=1}^n a_i X_i + c\right) = \sum_{i=1}^n a_i E(X_i) + c \quad (25)$$

## 独立

在讨论期望时，一个非常重要的情形是两个随机变量之间的独立性。如果两个随机变量  $X$  和  $Y$  相互独立，那么它们的联合行为就能以一种非常简单的方式分解开来。讨论期望时，独立性带来一个关键结论，如果  $X$  和  $Y$  独立，

$$E(XY) = E(X)E(Y) \quad (26)$$

这表示两个独立随机变量的乘积，其期望可以写成各自期望的乘积。

(26) 等价于

$$E(XY) - E(X)E(Y) = 0 \quad (27)$$

还是用掷两颗均匀色子试验为例，设第一颗色子的点数为  $X$ ，第二颗为  $Y$ ，它们相互独立且都服从 1 到 6 的离散均匀分布，因此

$$E(X) = E(Y) = 3.5 \quad (28)$$

根据独立性的结论，有

$$E(XY) = E(X)E(Y) = 3.5 \times 3.5 = 12.25 \quad (29)$$

如果按照传统方法去枚举 36 个点并求加权平均，也能得到相同结果，但独立性让计算过程大幅简化。

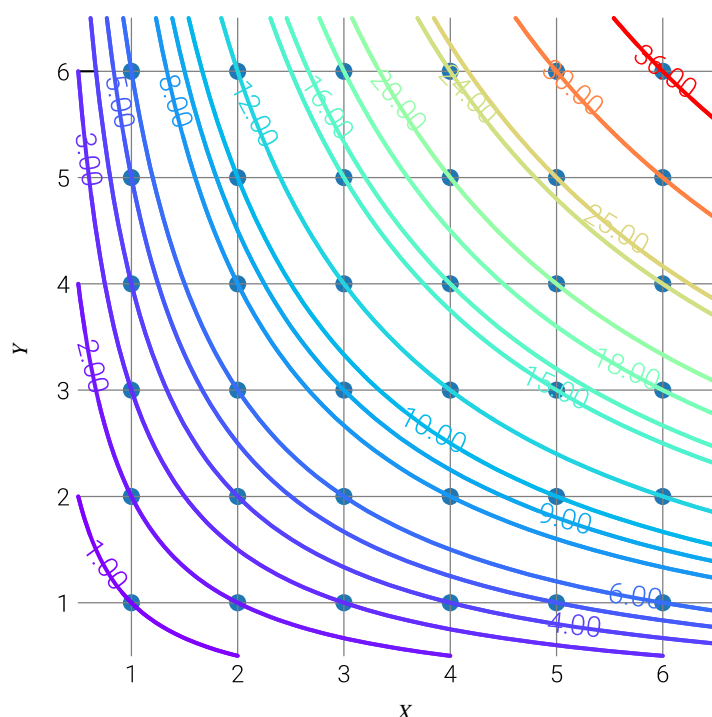


图 13. 掷两颗均匀色子试验色子点数分别为  $X$  和  $Y$ ， $XY$  的样本空间

然而，如果两个随机变量不独立，情况就完全不同了。

在这种情况下， $E(XY)$  一般不等于  $E(X)E(Y)$ 。两个量之间的依赖关系会改变乘积的平均值，而这个差异正是**协方差** (covariance) 概念的来源。协方差刻画两个随机变量是否倾向于同时变大、同时变小，或者一个变大时另一个倾向变小。更具体地，它度量的正是

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y) \quad (30)$$

如果这个差异为零，说明它们的“线性关联”被完全抵消，而在独立情况下，这个差异必然为零。



本书后续讲专门讲解协方差。



请大家用 DeepSeek/ChatGPT 等工具完成本节如下习题。

**Q1.** 同时抛一枚硬币 (正面为 1、反面为 0)，掷一颗色子 (点数为 1、2、3、4、5、6)，用随机变量  $X$ 、 $Y$  分别表示结果。请绘制样本空间，并计算

▶  $E(X + Y)$

▶  $E(2X + Y)$

▶  $E(XY)$

**Q2.** 使用 Python 模拟掷一枚均匀硬币 1000 次，计算实验均值，并与理论期望 0.5 比较。

**Q3.** 什么是 LOTUS (Law of the Unconscious Statistician)? 如何用 LOTUS 理解(15)?