Random Variables: Distribution and Expectation

Recall our setup of a probabilistic experiment as a procedure of drawing a sample from a set of possible values, and assigning a probability for each possible outcome of the experiment. For example, if we toss a fair coin n times, then there are 2^n possible outcomes, each of which is equally likely and has probability $\frac{1}{2^n}$.

Now suppose we want to make a measurement in our experiment. For example, we can ask what is the number of heads in n coin tosses; call this number X. Of course, X is not a fixed number, but it depends on the actual sequence of coin flips that we obtain. For example, if $n = 4$ and we observe the outcome $\omega = HTHH$, then is $X = 3$; whereas if we observe the outcome $\omega = HTHT$, then the is $X = 2$. In this example of n coin tosses, we only know that X is an integer between 0 and n, but we do not know what its exact value is until we observe which outcome of n coin flips is realized and count how many heads there are. Because every possible outcome is assigned a probability, the value X also carries with it a probability for each possible value it can take. The table below lists all the possible values X can take in the example of $n = 4$ coin tosses, along with their respective probabilities.

outcomes ω	value of X (# heads)	probability of occurring
$TTTT$	0	$1/16$
$HTTT$, $THTT$, $TTHT$, $TTTH$	1	$4/16$
$HHTT$, $HTHT$, $HTTH$, $THHT$, $THTH$, $TTHH$	2	$6/16$
$HHHT$, $HHTH$, $HTHH$, $THHH$	3	$4/16$
$HHHH$	4	$1/16$

Such a value X that depends on the outcome of the probabilistic experiment is called a *random variable* (abbreviated *r.v.*). As we see from the example above, a random variable X typically does not have a definitive value, but instead only has a probability *distribution* over the set of possible values X can take, which is why it is called random. So the question "What is the number of heads in n coin tosses?" does not exactly make sense because the answer X is a random variable. But the question "What is the *typical* number of heads in n coin tosses?" makes sense — it is asking what is the average value of X (the number of heads) if we repeat the experiment of tossing n coins multiple times. This average value is called the *expectation* of X, and is one of the most useful summary (also called *statistics*) of an experiment.

1 Random Variables

Before we formalize the above notions, let us consider another example to enforce our conceptual understanding of a random variable.

Example: Fixed Points of Permutations

Question: Suppose we collect the homeworks of n students, randomly shuffle them, and return them to the students. How many students receive their own homework?

离散数学与概率论
2020年秋季课程笔记 **笔记15**

随机变量：分布与期望

回顾我们对概率实验的设定：从一组可能的值中抽取样本，并为每个可能的结果分配一个概率。例如，如果我们抛一枚公平硬币n次，则有2n种可能的结果，每种结果的可能性相等，概率为12n。现在假设我们想在实验中进行一次测量。例如，我们可以问n次抛硬币中有多少次是正面；把这个数量记为X。当然，X不是一个固定数字，它取决于我们实际获得的硬币翻转序列。例如，当n = 4且我们观察到结果ω= HTHH时，X = 3；而如果我们观察到结果ω= HTHT，则X = 2。在这个抛n次硬币的例子中，我们只知道X是0到n之间的整数，但在观察到具体的n次抛硬币结果并统计正面数量之前，我们不知道它的准确值。由于每个可能的结果都被赋予了概率，X的值也因此携带了其可能取值的相应概率。下表列出了n = 4次抛硬币示例中X可能取的所有值及其相应的概率。

结果ω	X的值（正面数量）	发生的概率
$TTTT$	0	1/16
$HTTT$, $THTT$, $TTHT$, $TTTH$	1	4/16
$HHTT$, $HTHT$, $HTTH$, $THHT$, $THTH$, $TTHH$	2	6/16
$HHHT$, $HHTH$, $HTHH$, $THHH$	3	4/16
$HHHH$	4	1/16

这种依赖于概率实验结果的数值X被称为随机变量（简写为r.v.）。正如我们在上面的例子中看到的，随机变量X通常没有确定的值，而是仅在X可能取的值集合上有一个概率分布，这就是它被称为“随机”的原因。因此，“n次抛硬币中有多少次正面？”这个问题并不确切成立，因为答案X是一个随机变量。但“n次抛硬币中正面的典型数量是多少？”这个问题是有意义的——它是在问如果我们重复多次抛n枚硬币的实验，X（正面数量）的平均值是多少。这个平均值被称为X的期望，是实验中最有用的总结之一（也称为统计量）。

1 随机变量

在形式化上述概念之前，让我们再考虑另一个例子来加强我们对随机变量概念的理解。

示例：排列的不动点

问题：假设我们收集了n名学生的作业，随机打乱后返还给学生。有多少学生收到自己的作业？

Here the probability space consists of all $n!$ permutations of the homeworks, each with equal probability $\frac{1}{n!}$. If we label the homeworks as $1,2,\dots,n$, then each sample point is a permutation $\pi = (\pi_1,\dots,\pi_n)$ where π_i is the homework that is returned to the i-th student.

As in the coin flipping case above, our question does not have a simple numerical answer (such as 4), because the number depends on the particular permutation we choose (i.e., on the sample point). Let us call the number of fixed points X_n, which is a random variable. Recall the table in an earlier note that lists the value taken by X_3 for different permutations of $\{1,2,3\}$.

Formal Definition of a Random Variable

We now formalize the concepts discussed above.

Definition 15.1 (Random Variable)**.** *A random variable X on a sample space Ω is a function $X\colon \Omega \to \mathbb{R}$ that assigns to each sample point $\omega \in \Omega$ a real number $X(\omega)$.*

Until further notice, we will restrict our attention to random variables that are discrete, i.e., they take values in a range that is finite or countably infinite. This means even though we define X to map Ω to $\mathbb{R}$, the actual set of values $\{X(\omega)\colon \omega \in \Omega\}$ that X takes is a discrete subset of $\mathbb{R}$.

A random variable can be visualized in general by the picture in Figure 1.[1] Note that the term "random variable" is really something of a misnomer: it is a function so there is nothing random about it and it is definitely not a variable! What is random is which sample point of the experiment is realized and hence the value that the random variable maps the sample point to.

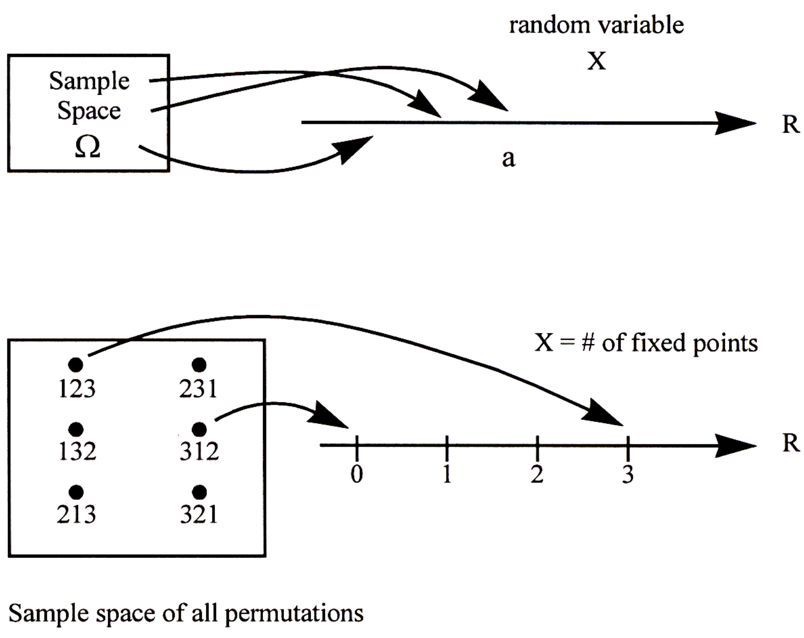


Figure 1: Visualization of how a random variable is defined on the sample space.

[1]The figures in this note are inspired by figures in Chapter 2 of *Introduction to Probability* by D. Bertsekas and J. Tsitsiklis.

这里的概率空间由所有n!个作业排列组成，每个排列具有相等的概率1 $\frac{}{n!}$.
如果我们把作业标记为1,2,...,n，则每个样本点是一个排列π= (π1,...,πn)，其中πi是返还给第i个学生的作业。

如同上面抛硬币的例子，我们的问题没有一个简单的数值答案（例如4），因为该数值取决于我们选择的具体排列（即样本点）。让我们称不动点的数量为Xn，这是一个随机变量。回想早先笔记中列出的不同{1,2,3}排列下X3取值的表格。

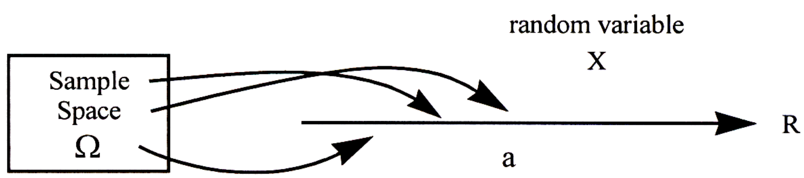
随机变量的形式化定义

我们现在正式化上述讨论的概念。

定义15.1（随机变量）。样本空间Ω上的随机变量X是一个函数X：Ω→R，它为每个样本点ω∈Ω分配一个实数X(ω)。

在进一步说明之前，我们将注意力限制在离散型随机变量上，即它们取值于有限或可数无限的范围内。这意味着即使我们将X定义为从Ω到R的映射，X实际取的值集合{X(ω): ω∈Ω}仍是R的一个离散子集。

一般情况下，随机变量可以通过图1中的图像进行可视化。1请注意，“随机变量”这个术语实际上有些名不副实：它是一个函数，因此本身并不随机，也肯定不是一个变量！真正随机的是实验中实现的是哪一个样本点，从而决定了随机变量映射到的值。



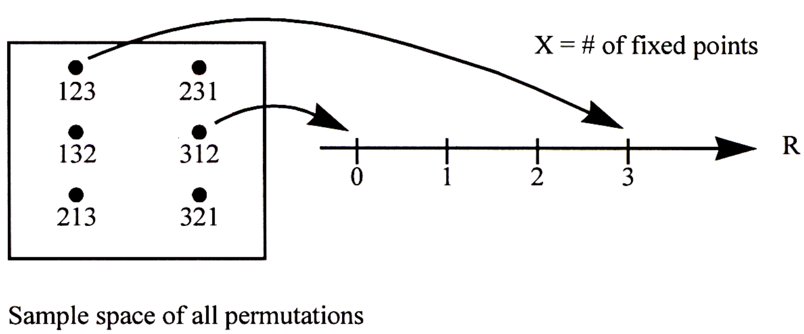


图1：随机变量在样本空间上定义的可视化示意图。

1本笔记中的图受到D. Bertsekas和J. Tsitsiklis所著《概率导论》第2章中图的启发。

2 Probability Distribution

When we introduced the basic probability space in an earlier note, we defined two things:

1. The sample space Ω consisting of all the possible outcomes (sample points) of the experiment.
2. The probability of each of the sample points.

Analogously, there are two important things about any random variable:

1. The set of values that it can take.
2. The probabilities with which it takes on the values.

Since a random variable is defined on a probability space, we can calculate these probabilities given the probabilities of the sample points. Let a be any number in the range of a random variable X. Then the set

$$\{\omega \in \Omega : X(\omega) = a\}$$

is an *event* in the sample space (simply because it is a subset of Ω). We usually abbreviate this event to simply "$X = a$". Since $X = a$ is an event, we can talk about its probability, $\mathbb{P}[X = a]$. The collection of these probabilities, for all possible values of a, is known as the *distribution* of the random variable X.

Definition 15.2 (Distribution)**.** *The distribution of a discrete random variable X is the collection of values $\{(a, \mathbb{P}[X = a]) : a \in \mathscr{A}\}$, where $\mathscr{A}$ is the set of all possible values taken by X.*

Thus, the distribution of the random variable X in our permutation example above is:

$$\mathbb{P}[X=0] = \frac{1}{3}, \qquad \mathbb{P}[X=1] = \frac{1}{2}, \qquad \mathbb{P}[X=3] = \frac{1}{6},$$

and $\mathbb{P}[X = a] = 0$ for all other values of a.

The distribution of a random variable can be visualized as a bar diagram, shown in Figure 2. The x-axis represents the values that the random variable can assume. The height of the bar at a value a is the probability $\mathbb{P}[X = a]$. Each of these probabilities can be computed by looking at the probability of the corresponding event in the sample space.

Note that the collection of events $X = a$, for $a \in \mathscr{A}$, satisfy two important properties:

- Any two events $X = a_1$ and $X = a_2$ with $a_1 \neq a_2$ are disjoint.
- The union of all these events is equal to the entire sample space Ω.

The collection of events thus form a *partition* of the sample space (see Figure 2). Both properties follow directly from the fact that X is a function defined on Ω, i.e., X assigns a unique value to each and every possible sample point in Ω. As a consequence, the sum of the probabilities $\mathbb{P}[X = a]$ over all possible values of a is exactly equal to 1. So when we sum up the probabilities of the events $X = a$, we are really summing up the probabilities of all the sample points.

2 概率分布

当我们早先介绍基本概率空间时，我们定义了两件事：

1. 实验所有可能结果（样本点）组成的样本空间Ω。
2. 每个样本点的概率。

类似地，任何随机变量都有两个重要的方面：

1. 它可以取的值的集合。
2. 它取各个值的概率。

由于随机变量定义在概率空间上，因此在已知样本点概率的情况下，我们可以计算这些概率。设a为随机变量X取值范围内的任意数，则集合

$$\{\omega\in\Omega: X(\omega) = a\}$$

是样本空间中的一个事件（因为它只是Ω的一个子集）。我们通常将此事件简写为X = a。由于X = a是一个事件，我们可以讨论它的概率P[X = a]。对于所有可能的a值，这些概率的集合称为随机变量X的分布。定义15.2（分布）。离散随机变量X的分布是值的集合{(a,P[X = a]) : a ∈A }，其中A是X取的所有可能值的集合。

因此，上述排列示例中随机变量X的分布为：

$$\mathbb{P}[X=0]=\frac{1}{3}, \qquad \mathbb{P}[X=1]=\frac{1}{2}, \qquad \mathbb{P}[X=3]=\frac{1}{6},$$

而对于所有其他a值，P[X = a] = 0。

随机变量的分布可以用条形图表示，如图2所示。横轴表示随机变量可以取的值。在值a处的条形高度为概率P[X = a]。这些概率中的每一个都可以通过查看样本空间中相应事件的概率来计算。

注意，对于a ∈A，事件集合X = a满足两个重要性质：

- 任意两个事件X = a1和X = a2，当a1 ≠ a2时，它们互不相交。
- 这些事件的并集等于整个样本空间Ω。

因此，这些事件的集合构成了样本空间的一个划分（见图2）。这两个性质都直接来源于X是定义在Ω上的一个函数这一事实，即X为Ω中每个可能的样本点分配唯一的值。因此，对所有可能的a值，P[X = a]的概率之和恰好等于1。当我们把X = a这些事件的概率加起来时，实际上是在对所有样本点的概率求和。

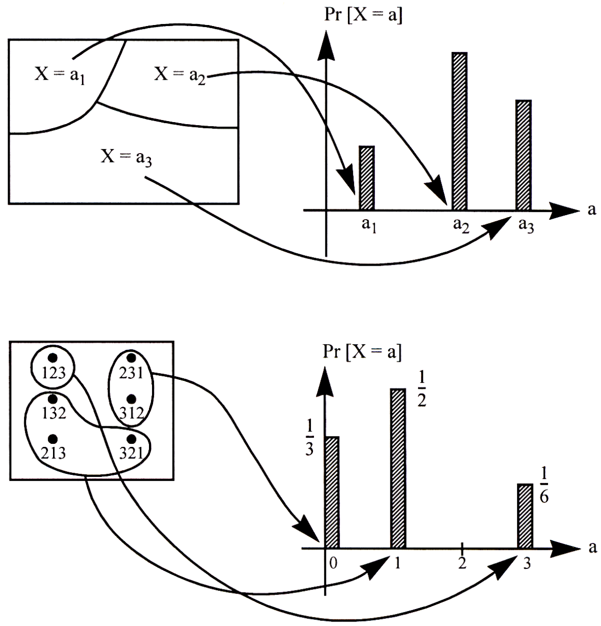


Figure 2: Visualization of how the distribution of a random variable is defined.

Bernoulli Distribution

A simple yet very useful probability distribution is the *Bernoulli* distribution of a random variable which takes value in $\{0,1\}$:

$$\mathbb{P}[X=i] = \begin{cases} p, & \text{if } i=1, \\ 1-p, & \text{if } i=0, \end{cases}$$

where $0 \le p \le 1$. We say that X is distributed as a *Bernoulli* random variable with parameter p, and write

$$X \sim \text{Bernoulli}(p).$$

Binomial Distribution

Let us return to our coin tossing example above, where we defined our random variable X to be the number of heads. More formally, consider the random experiment consisting of n independent tosses of a biased coin that shows H with probability p. Each sample point ω is a sequence of tosses, and $X(\omega)$ is defined to be the number of heads in ω. For example, when $n=3$, $X(THH)=2$.

To compute the distribution of X, we first enumerate the possible values that X can take. They are simply $0,1,\ldots,n$. Then we compute the probability of each event $X=i$ for $i=0,1,\ldots,n$. The probability of the event $X=i$ is the sum of the probabilities of all the sample points with exactly i heads (for example, if $n=3$ and $i=2$, there would be three such sample points $\{HHT,HTH,THH\}$). Any such sample point has probability $p^i(1-p)^{n-i}$, since the coin flips are independent. There are exactly $\binom{n}{i}$ of these sample points. Hence,

$$\mathbb{P}[X=i] = \binom{n}{i} p^i (1-p)^{n-i}, \qquad \text{for } i=0,1,\ldots,n. \tag{1}$$

This distribution, called the *binomial* distribution, is one of the most important distributions in probability.

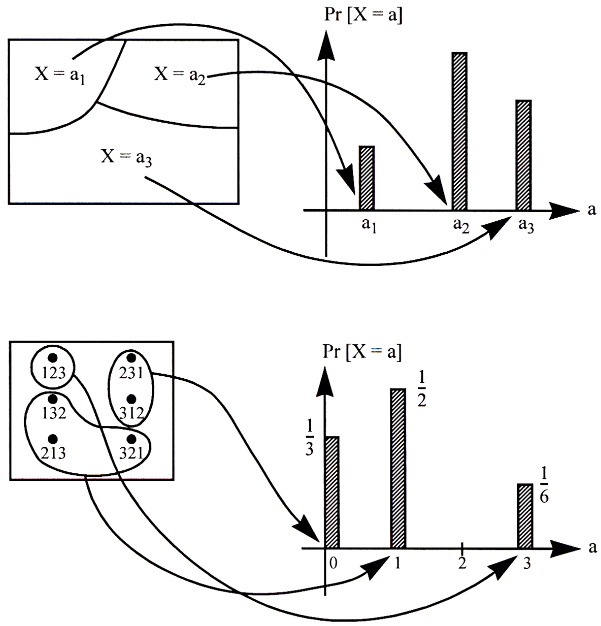


图2：随机变量分布定义的可视化示意图。

伯努利分布

一种简单但非常有用的概率分布是取值于{0,1}的随机变量的伯努利分布：

$$\mathbb{P}[X=i]=\begin{cases}\text{p，若 i = 1,}\\ \text{1−p, 若i = 0,}\end{cases}$$

其中0 ⩽p ⩽1。我们说X服从参数为p的伯努利随机变量分布，并记作X ~Bernoulli(p)。

二项分布

让我们回到上面的抛硬币例子，在那里我们将随机变量X定义为正面出现的次数。更正式地说，考虑一个由n次独立投掷一枚偏置硬币组成的随机实验，该硬币出现H的概率为p。每个样本点ω是一个投掷序列，X(ω)被定义为ω中正面的数量。例如，当n = 3时，X(THH) = 2。

为了计算X的分布，我们首先枚举X可能取的所有值，它们就是0,1,...,n。然后我们计算每个事件X = i（i = 0,1,...,n）的概率。事件X = i的概率等于所有恰好包含i个正面的样本点的概率之和（例如，当n = 3且i = 2时，会有三个这样的样本点{HHT,HTH,THH}）。任何一个这样的样本点的概率为pi(1−p)n−i，因为硬币投掷是相互独立的。恰好有$\binom{n}{i}$这些样本点的概率之和。

因此，

$$\mathbb{P}[X=i]=\binom{n}{i}\text{pi(1−p)n−i，其中i = 0,1,...,n。} \qquad (1)$$

这个分布被称为二项分布，是概率论中最重要的分布之一。

A random variable with this distribution is called a *binomial* random variable, and we write

$$X \sim \text{Bin}(n, p),$$

where n denotes the number of trials and p the probability of success (observing an H in the example). An example of a binomial distribution is shown in Figure 3. Notice that due to the properties of X mentioned above, it must be the case that $\sum_{i=0}^{n} \mathbb{P}[X = i] = 1$, which implies that $\sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} = 1$. This provides a probabilistic proof of the Binomial Theorem from an earlier note where we saw it combinatorially, for $a = p$ and $b = 1 - p$.

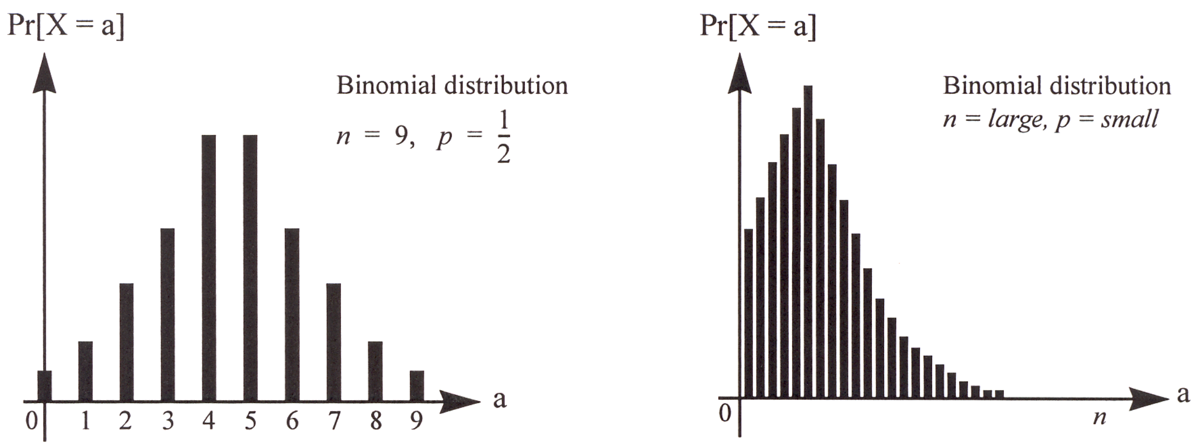


Figure 3: The binomial distributions for two choices of (n, p).

Although we define the binomial distribution in terms of an experiment involving tossing coins, this distribution is useful for modeling many real-world problems. Consider for example the error correction problem studied earlier. Recall that we wanted to encode n packets into $n + k$ packets such that the recipient can reconstruct the original n packets from any n packets received. But in practice, the number of packet losses is random, so how do we choose k, the amount of redundancy? If we model each packet getting lost with probability p and the losses are independent, then if we transmit $n + k$ packets, the number of packets received is a random variable X with binomial distribution: $X \sim \text{Bin}(n + k, 1 - p)$ (we are tossing a coin $n + k$ times, and each coin turns out to be a head (packet received) with probability $1 - p$). So the probability of successfully decoding the original data is:

$$\mathbb{P}[X \geq n] = \sum_{i=n}^{n+k} \mathbb{P}[X = i] = \sum_{i=n}^{n+k} \binom{n+k}{i} (1-p)^i p^{n+k-i}.$$

Given fixed n and p, we can choose k such that this probability is no less than, say, 0.99.

Hypergeometric Distribution

Consider an urn containing $N = B + W$ balls, where B balls are black and W are white. Suppose you randomly sample $n \leq N$ balls from the urn *with* replacement, and let X denote the number of black balls in your sample. What is the probability distribution of X? Since the probability of seeing a black ball is B/N for each draw, independently of all other draws, X follows the binomial distribution $\text{Bin}(n, p)$, where $p = B/N$.

What if you randomly sample $n \leq N$ balls from the urn *without* replacement? In this case, the probability of seeing a black ball in the i-th draw depends on the colors of the $i - 1$ balls already drawn; that is, unlike

具有这种分布的随机变量称为二项随机变量，我们记作X ~Bin(n, p)，

其中n表示试验次数，p表示成功的概率（在例子中即观察到H的概率）。图3展示了一个二项分布的例子。注意，根据上述X的性质，必然有Σni=0P[X = i] = 1，这意味着Σni=0 $\binom{n}{i}$ pi(1−p)n−i = 1。这提供了一种来自早期笔记中二项式定理的概率性证明，当时我们是从组合角度看到它的，其中a = p且b = 1−p。

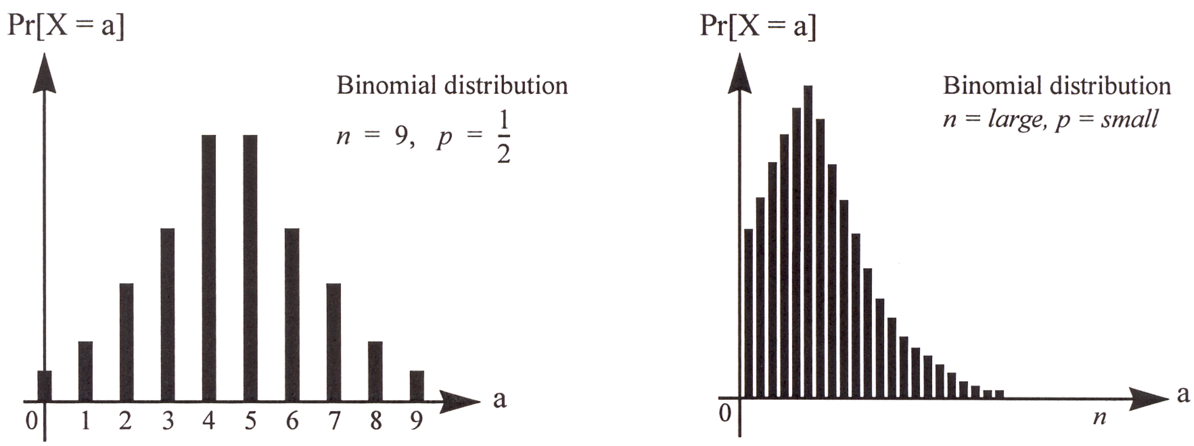


图3：两种不同(n, p)取值下的二项分布。

尽管我们是通过抛硬币实验来定义二项分布的，但这种分布在建模许多现实世界问题时非常有用。例如考虑之前研究过的纠错问题。回想一下，我们希望将n个数据包编码为n + k个数据包，使得接收方可以从任意接收到的n个数据包中重建原始的n个数据包。但在实际中，丢失的数据包数量是随机的，那么我们应如何选择冗余量k？如果我们假设每个数据包以概率p丢失，且丢失事件相互独立，那么当我们发送n+k个数据包时，接收到的数据包数量是一个随机变量X，服从二项分布：X ~Bin(n+k,1−p)（相当于抛一枚硬币n+k次，每次结果为正面（数据包被成功接收）的概率为1−p）。因此成功解码原始数据的概率为：

$$\mathbb{P}[X \geq n] = \sum_{i=n}^{n+k} \mathbb{P}[X = i] = \sum_{i=n}^{n+k} \binom{n+k}{i} (1-p)^i p^{n+k-i}.$$

给定固定的n和p，我们可以选择k使得该概率至少达到例如0.99。

超几何分布

考虑一个包含N = B +W个球的瓮，其中有B个黑球和W个白球。假设你从瓮中有放回地随机抽取n ≤N个球，并令X表示你样本中黑球的数量。X的概率分布是什么？由于每次抽取看到黑球的概率都是B/N，且各次抽取相互独立，因此X服从二项分布Bin(n, p)，其中p = B/N。

如果你从瓮中无放回地随机抽取n ≤N个球会怎样？在这种情况下，第i次抽取得到黑球的概率取决于前i−1个已抽出球的颜色；也就是说，与

in the case of sampling with replacement, you don't have independence. The probability distribution of the number Y of black balls in this setting can be found as follows. Consider a sequence ω of n draws where the first k balls are black and the next $n-k$ balls are white. The probability of this particular sequence of draws can be computed as

$$\begin{aligned}\mathbb{P}[\omega] &= \frac{B}{N} \times \frac{B-1}{N-1} \times \cdots \times \frac{B-k+1}{N-k+1} \times \frac{W}{N-k} \times \frac{W-1}{N-k-1} \times \cdots \times \frac{W-(n-k)+1}{N-n+1} \qquad (2)\\ &= \frac{\frac{B!}{(B-k)!}\frac{W!}{(W-(n-k))!}}{\frac{N!}{(N-n)!}} \\ &= \frac{\frac{B!}{(B-k)!}\frac{(N-B)!}{(N-B-(n-k))!}}{\frac{N!}{(N-n)!}}.\end{aligned}$$

Furthermore, any other sequence ω' consisting of k blacks balls and $n-k$ white balls has exactly the same probability as ω: if we carry out a similar calculation as in (2) for the sequence of draws in ω', the numerators appearing in $\mathbb{P}[\omega']$ will be some permutation of the numerators in the first line of (2), while the denominators will be exactly the same as in (2). Since there are $\binom{n}{k}$ distinct sequences of length n consisting of k black balls and $n-k$ white balls, we obtain

$$\mathbb{P}[Y=k] = \binom{n}{k}\frac{\frac{B!}{(B-k)!}\frac{(N-B)!}{(N-B-(n-k))!}}{\frac{N!}{(N-n)!}} = \frac{\binom{B}{k}\binom{N-B}{n-k}}{\binom{N}{n}},$$

for $k \in \{0,1,\ldots,n\}$. (Note that $\binom{m}{j} = 0$ if $j > m$, so $\mathbb{P}[Y=k] \neq 0$ only if $\max(0, n+B-N) \leq k \leq \min(n,B)$.) This probability distribution is called the *hypergeometric distribution* with parameters N, B, n, and we write

$$Y \sim \text{Hypergeometric}(N,B,n).$$

3 Multiple Random Variables and Independence

Often one is interested in multiple random variables on the same sample space. Consider, for example, the sample space of flipping two coins. One could define many random variables: for example a random variable X indicating the number of heads in a sequence of coin tosses, or a random variable Y indicating the number of tails, or a random variable Z indicating whether the first is H or not. Note that for each sample point, any random variable has a specific value: e.g., for $\omega = HTT$, we have $X(\omega) = 1$, $Y(\omega) = 2$, and $Z(\omega) = 1$.

The concept of a distribution can then be extended to probabilities for the combination of values for multiple random variables.

Definition 15.3. *The joint distribution for two discrete random variables X and Y is the collection of values $\{((a,b), \mathbb{P}[X=a, Y=b]) : a \in \mathscr{A}, b \in \mathscr{B}\}$, where $\mathscr{A}$ is the set of all possible values taken by X and $\mathscr{B}$ is the set of all possible values taken by Y.*

When given a joint distribution for X and Y, the distribution $\mathbb{P}[X=a]$ for X is called the *marginal distribution* for X, and can be found by "summing" over the values of Y. That is,

$$\mathbb{P}[X=a] = \sum_{b \in \mathscr{B}} \mathbb{P}[X=a, Y=b].$$

在有放回抽样的情况下，事件之间并不独立。这种情形下，抽到黑球的数量Y的概率分布可如下计算。考虑一个长度为n的抽取序列ω，其中前k个球是黑色，后n−k个球是白色。该特定抽取序列的概率可以计算为

$$\begin{aligned}\mathbb{P}[\omega] &= \frac{B}{N}\times\frac{B-1}{N-1}\times\cdots\times\frac{B-k+1}{N-k+1}\times\frac{W}{N-k}\times\frac{W-1}{N-k-1}\times\cdots\times\frac{W-(n-k)+1}{N-n+1} \\ &= \frac{\frac{B!}{(B-k)!}\frac{W!}{(W-(n-k))!}}{\frac{N!}{(N-n)!}} \\ &= \frac{\frac{B!}{(B-k)!}\frac{(N-B)!}{(N-B-(n-k))!}}{\frac{N!}{(N-n)!}}.\end{aligned} \tag{2}$$

此外，任何由k个黑球和n−k个白球组成的其他序列ω0具有与ω完全相同的概率：如果我们对ω0中的抽取序列进行类似于(2)式的计算，P[ω0]中的分子将是(2)式第一行分子的一个排列，而分母则与(2)式完全相同。由于长度为n且包含k个黑球和n−k个白球的不同序列共有nk个，因此我们得到
( )

$$\mathbb{P}[Y=k] = \binom{n}{k}\frac{\frac{B!}{(B-k)!}\frac{(N-B)!}{(N-B-(n-k))!}}{\frac{N!}{(N-n)!}} = \frac{\binom{B}{k}\binom{N-B}{n-k}}{\binom{N}{n}},$$

对于k ∈{0,1,...,n}。（注意，( ) 当j > m时mj = 0，因此P[Y = k] = 0仅当max(0,n+B−N) ≤k ≤min(n,B)成立。）这种概率分布称为参数为N,B,n的超几何分布，我们记作

Y ∼Hypergeometric(N,B,n)。

3 多个随机变量与独立性

人们常常关注同一样本空间上的多个随机变量。例如，考虑抛两枚硬币的样本空间。可以定义许多随机变量：比如随机变量X表示一系列抛硬币中正面的次数，或随机变量Y表示反面的次数，或随机变量Z表示第一次是否为正面H。注意，对于每个样本点，任何随机变量都有一个特定的值：例如，对于ω= HTT，我们有X(ω) = 1，Y(ω) = 2，Z(ω) = 1。

因此，分布的概念可以扩展到多个随机变量组合取值的概率。

定义15.3。两个离散随机变量X和Y的联合分布是集合{((a,b),P[X = a,Y = b]) : a ∈A , b ∈B}，其中A是X所有可能取值的集合，B是Y所有可能取值的集合。

当给定X和Y的联合分布时，X的分布P[X = a]称为X的边缘分布，可以通过对Y的所有取值求和得到。也就是说，

$$\mathbb{P}[X=a] = \sum_{b\in\mathscr{B}} \text{P[X = a,Y = b]。}$$

The marginal distribution for Y is analogous, as is the notion of a joint distribution for any number of random variables.

A joint distribution over random variables $X_1,\dots,X_n$ (for example, X_i could be the value of the i-th roll of a sequence of n die rolls) is $\mathbb{P}[X_1 = a_1,\dots,X_n = a_n]$, where $a_i \in \mathscr{A}_i$ and $\mathscr{A}_i$ is the set of possible values for X_i. The marginal distribution for X_i is simply the distribution for X_i and can be obtained by summing over all the possible values of the other variables, but in some cases can be derived more simply. We proceed to one such case.

Independence for random variables is defined in an analogous fashion to independence for events:

Definition 15.4 (Independence)**.** *Random variables X and Y on the same probability space are said to be* independent *if the events $X = a$ and $Y = b$ are independent for all values a,b. Equivalently, the joint distribution of independent r.v.'s decomposes as*

$$\mathbb{P}[X = a, Y = b] = \mathbb{P}[X = a]\mathbb{P}[Y = b], \quad \forall a,b.$$

Mutual independence of more than two r.v.'s is defined similarly.

A very important example of independent random variables are indicator random variables for independent events. If I_i denotes the indicator r.v. for the i-th toss of a coin being H, then $I_1,\dots,I_n$ are mutually independent random variables. This example motivates the commonly used phrase "*independent and identically distributed (i.i.d.)* set of random variables." In this example, $\{I_1,\dots,I_n\}$ is a set of i.i.d. indicator random variables.

4 Expectation

The distribution of a r.v. contains *all* the probabilistic information about the r.v. In most applications, however, the complete distribution of a r.v. is very hard to calculate. For example, consider the homework permutation example with $n = 20$. In principle, we would have to enumerate $20! \approx 2.4 \times 10^{18}$ sample points, compute the value of X at each one, and count the number of points at which X takes on each of its possible values (though in practice we could streamline this calculation a bit)! Moreover, even when we can compute the complete distribution of a r.v., it is often not very informative.

For these reasons, we seek to *summarize* the distribution into a more compact, convenient form that is also easier to compute. The most widely used such form is the *expectation* (or *mean* or *average*) of the r.v.

Definition 15.5 (Expectation)**.** *The* <u>*expectation*</u> *of a discrete random variable X is defined as*

$$\mathbb{E}[X] = \sum_{a\in\mathscr{A}} a \times \mathbb{P}[X = a], \tag{3}$$

where the sum is over all possible values taken by the r.v.

(**Technical Note.** Expectation is well defined provided that the sum on the right hand side of (3) is absolutely convergent, i.e., $\sum_{a\in\mathscr{A}} |a| \times \mathbb{P}[X = a] < \infty$. There are random variables for which expectations do not exist.)

For our simpler permutation example with only 3 students, the expectation is

$$\mathbb{E}[X] = \left(0 \times \frac{1}{3}\right) + \left(1 \times \frac{1}{2}\right) + \left(3 \times \frac{1}{6}\right) = 0 + \frac{1}{2} + \frac{1}{2} = 1.$$

That is, the expected number of fixed points in a permutation of three items is exactly 1.

Y的边缘分布类似，同样也可以推广到任意多个随机变量的联合分布概念。

随机变量X1,...,Xn上的联合分布（例如，Xi可以表示第i次掷骰子的结果
n次骰子投掷序列）为P[X1 = a1,...,Xn = an]，其中ai ∈Ai，Ai是Xi可能取值的集合。Xi的边缘分布就是Xi自身的分布，可通过对其它变量所有可能取值求和得到，但在某些情况下可以更简便地推导。我们接下来讨论这样一个情况。

随机变量的独立性定义与事件独立性的定义类似：定义15.4（独立性）。定义在同一概率空间上的

随机变量X和Y，若对所有取值a,b，事件X = a和Y = b都是独立的，则称X和Y是独立的。等价地，其联合
独立随机变量的联合分布可分解为

P[X = a,Y = b] = P[X = a]P[Y = b], ∀a,b。

两个以上随机变量的相互独立性以类似方式定义。

独立随机变量的一个非常重要的例子是独立事件的指示随机变量。如果Ii表示第i次抛硬币结果为正面H的指示随机变量，那么I1,...,In是相互独立的随机变量。这个例子引出了常用的术语“独立同分布（i.i.d.）”随机变量集合。在这个例子中，{I1,...,In}是一组独立同分布的指示随机变量。

4 期望

一个随机变量的分布包含了关于该随机变量的所有概率信息。但在大多数应用中，完整计算一个随机变量的分布是非常困难的。例如，考虑n = 20的学生作业排列示例。原则上，我们需要枚举20! ≈2.4 ×1018个样本点，计算每个点上X的取值，并统计X取各个可能值的样本点数量（尽管实践中我们可以稍微优化这一计算过程）！此外，即使我们能够计算出随机变量的完整分布，它往往也不够直观易懂。

出于这些原因，我们希望将分布总结为一种更紧凑、方便的形式，也更容易计算。最广泛使用的形式之一就是随机变量的期望（或均值、平均值）。定义15.5（期望）。离散随机变量X的期望定

义为

$$\mathbb{E}[X] = \sum_{a \in \mathscr{A}} a \times \mathbb{P}[X = a], \tag{3}$$

其中求和覆盖了随机变量所有可能取的值。

（技术说明：当公式(3)右侧的求和绝对收敛时，即Σa∈A |a|×P[X = a] < ∞，期望是有明确定义的。存在一些期望不存在的随机变量。）对于我们只有3个学生的简单排列示例，期望值为

$$\mathbb{E}[X] = \left(0 \times \frac{1}{3}\right) + \left(1 \times \frac{1}{2}\right) + \left(3 \times \frac{1}{6}\right) = 0 + \frac{1}{2} + \frac{1}{2} = 1.$$

也就是说，在三个元素的排列中，固定点的期望数量恰好是1。

The expectation can be seen in some sense as a "typical" value of the r.v. (though note that $\mathbb{E}[X]$ may not actually be a value that X can take). The question of how typical the expectation is for a given r.v. is a very important one that we shall return to in a later lecture.

Here is a physical interpretation of the expectation of a random variable: imagine carving out a wooden cutout figure of the probability distribution as in Figure 4. Then the expected value of the distribution is the balance point (directly below the center of gravity) of this object.

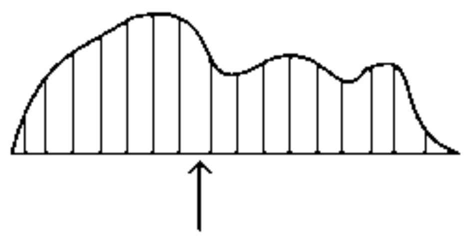


Figure 4: Physical interpretation of expected value as the balance point.

4.1 Examples

1. **Single die.** Throw a fair die once and let X be the number that comes up. Then X takes on values $1,2,\ldots,6$ each with probability $\frac{1}{6}$, so

$$\mathbb{E}[X] = \frac{1}{6}(1+2+3+4+5+6) = \frac{21}{6} = \frac{7}{2}.$$

   Note that X never actually takes on its expected value $\frac{7}{2}$.

2. **Two dice.** Throw two fair dice and let X be the sum of their scores. Then the distribution of X is

a	2	3	4	5	6	7	8	9	10	11	12
$\mathbb{P}[X=a]$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

   The expectation is therefore

$$\mathbb{E}[X] = \left(2\times\frac{1}{36}\right) + \left(3\times\frac{1}{18}\right) + \left(4\times\frac{1}{12}\right) + \cdots + \left(12\times\frac{1}{36}\right) = 7.$$

3. **Roulette.** A roulette wheel is spun (recall that a roulette wheel has 38 slots: the numbers $1,2,\ldots,36$, half of which are red and half black, plus 0 and 00, which are green). You bet \$1 on Black. If a black number comes up, you receive your stake plus \$1; otherwise you lose your stake. Let X be your net winnings in one game. Then X can take on the values $+1$ and -1, and $\mathbb{P}[X=1] = \frac{18}{38}$, $\mathbb{P}[X=-1] = \frac{20}{38}$. Thus,

$$\mathbb{E}[X] = \left(1\times\frac{18}{38}\right) + \left(-1\times\frac{20}{38}\right) = -\frac{1}{19};$$

   i.e., you expect to lose about a nickel per game. Notice how the zeros tip the balance in favor of the casino!

在某种意义上，期望可以被视为随机变量的一个典型值（但请注意，E[X]实际上可能并不是X可以取到的值）。对于给定的随机变量，期望有多“典型”是一个非常重要的问题，我们将在后续讲座中再次讨论。

这里给出随机变量期望的一个物理解释：想象按照图4中的概率分布雕刻出一个木制轮廓图形。那么该分布的期望值就是这个物体的平衡点（位于重心正下方）。

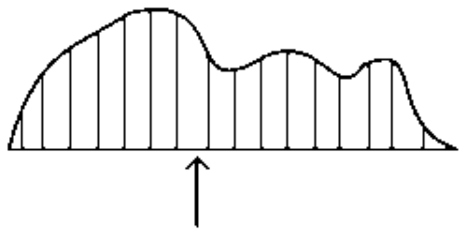


图4：期望值的物理意义，即平衡点。

4.1 示例

1. 单次掷骰子。掷一次公平的骰子，令X为出现的点数。那么X取值1,2,...,6，每个值的概率均为 $\frac{1}{6}$, 所以

$$\mathbb{E}[X] = \frac{1}{6}(1+2+3+4+5+6) = \frac{21}{6} = \frac{7}{2}.$$

注意X实际上从未取到其期望值 $\frac{7}{2}$。

2. 两个骰子。掷两个公平的骰子，令X为它们点数之和。则X的分布为

a	2	3	4	5	6	7	8	9	10	11	12
$\mathbb{P}[X=a]$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

因此期望值为

$$\mathbb{E}[X] = \left(2\times\frac{1}{36}\right) + \left(3\times\frac{1}{18}\right) + \left(4\times\frac{1}{12}\right) + \cdots + \left(12\times\frac{1}{36}\right) = 7.$$

3. 轮盘赌。旋转一个轮盘（轮盘共有38个槽位：数字1,2,...,36，其中一半为红色，一半为黑色，加上0和00，这两个为绿色）。你下注1美元押黑色。如果出现黑色数字，你拿回本金并额外赢得1美元；否则你失去本金。令X为你单局游戏的净赢钱。那么X可能取值为+1和−1，且P[X = 1] = $\frac{18}{38}$, $\mathbb{P}[X=-1] = \frac{20}{38}$.

因此，

$$\mathbb{E}[X] = \left(1\times\frac{18}{38}\right) + \left(-1\times\frac{20}{38}\right) = -\frac{1}{19};$$

也就是说，你平均每局游戏会输掉大约五分钱。注意这些零是如何让天平偏向赌场的！

4.2 Linearity of Expectation

So far, we have computed expectations by brute force: i.e., we have written down the whole distribution and then added up the contributions for all possible values of the r.v. The real power of expectations is that in many real-life examples they can be computed much more easily using a simple shortcut. The shortcut is the following:

Theorem 15.1. *For any two random variables X and Y on the same probability space, we have*

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

Also, for any constant c, we have

$$\mathbb{E}[cX] = c\,\mathbb{E}[X].$$

Proof. We first rewrite the definition of expectation in a more convenient form. Recall from Definition 15.5 that

$$\mathbb{E}[X] = \sum_{a\in\mathscr{A}} a \times \mathbb{P}[X=a].$$

Consider a particular term $a \times \mathbb{P}[X=a]$ in the above sum. Notice that $\mathbb{P}[X=a]$, by definition, is the sum of $\mathbb{P}[\omega]$ over those sample points ω for which $X(\omega)=a$. Furthermore, we know that every sample point $\omega \in \Omega$ is in exactly one of these events $X=a$. This means we can write out the above definition in a more long-winded form as

$$\mathbb{E}[X] = \sum_{\omega\in\Omega} X(\omega) \times \mathbb{P}[\omega]. \tag{4}$$

This equivalent definition of expectation will make the present proof much easier (though it is usually less convenient for actual calculations). Applying (4) to $\mathbb{E}[X+Y]$ gives:

$$\begin{aligned}
\mathbb{E}[X+Y] &= \textstyle\sum_{\omega\in\Omega} (X+Y)(\omega) \times \mathbb{P}[\omega] \\
&= \textstyle\sum_{\omega\in\Omega} (X(\omega)+Y(\omega)) \times \mathbb{P}[\omega] \\
&= \textstyle\sum_{\omega\in\Omega} \Big(X(\omega) \times \mathbb{P}[\omega]\Big) + \sum_{\omega\in\Omega} \Big(Y(\omega) \times \mathbb{P}[\omega]\Big) \\
&= \mathbb{E}[X] + \mathbb{E}[Y].
\end{aligned}$$

In the last step, we used (4) twice.

This completes the proof of the first equality. The proof of the second equality is much simpler and is left as an exercise. □

Theorem 15.1 is very powerful: it says that the expectation of a sum of r.v.'s is the sum of their expectations, with no assumptions about the r.v.'s. We can use Theorem 15.1 to conclude things like $\mathbb{E}[3X-5Y] = 3\,\mathbb{E}[X] - 5\,\mathbb{E}[Y]$, regardless of whether or not X and Y are independent. This important property is known as linearity of expectation.

Important caveat: Theorem 15.1 does **not** *say that $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$, or that $\mathbb{E}\left[\frac{1}{X}\right] = \frac{1}{\mathbb{E}[X]}$, etc. These claims are not true in general. It is only* sums *and* differences *and* constant multiples *of random variables that behave so nicely.*

4.2 期望的线性性

到目前为止，我们都是通过暴力计算期望值：即写出整个分布，然后将随机变量所有可能取值的贡献相加。期望的真正威力在于，在许多现实生活例子中，它们可以通过一个简单的捷径更轻松地计算出来。这个捷径如下：

定理15.1。对于同一概率空间上的任意两个随机变量X和Y，我们有E[X +Y] =

E[X]+E[Y]。

此外，对于任意常数c，我们有

$$E[cX] = cE[X]。$$

证明。我们首先将期望的定义改写成一个更方便的形式。回顾定义15.5，

$$\mathbb{E}[X] = \sum_{a\in\mathscr{A}} a\times P[X = a]。$$

考虑上述求和中的特定项a×P[X = a]。注意到根据定义，P[X = a]是在那些满足X(ω) = a的样本点ω上P[ω]的总和。此外，我们知道每个样本点ω∈Ω恰好属于一个事件X = a。这意味着我们可以将上述定义以更详细的形式展开为

$$\mathbb{E}[X] = \sum_{\omega\in\Omega} X(\omega)\times P[\omega]。 \tag{4}$$

这个期望的等价定义将使当前的证明更加容易（尽管在实际计算中通常不太方便）。将(4)应用于E[X +Y]可得：

$$\begin{aligned} E[X+Y] &= \Sigma_{\omega\in\Omega}(X+Y)(\omega)\times P[\omega] = \Sigma_{\omega\in\Omega}(X(\omega)+Y(\omega))\times P[\omega] \\ &= \left(\Sigma_{\omega\in\Omega}X(\omega)\times P[\omega]\right) + \left(\Sigma_{\omega\in\Omega}Y(\omega)\times P[\omega]\right) = E[X]+E[Y]。\end{aligned}$$

在最后一步中，我们两次使用了(4)。

这完成了第一个等式的证明。第二个等式的证明要简单得多，留作练习。

□

定理15.1非常强大：它指出，无论随机变量之间是否存在任何假设，一组随机变量之和的期望等于它们各自期望之和。我们可以用定理15.1得出诸如E[3X −5Y] = 3E[X]−5E[Y]之类的结论，无论X和Y是否独立。这一重要性质被称为期望的线性性。

重要提醒：定理15.1并未说明E[XY] = E[X]E[Y]，或 $\mathbb{E}\left[\frac{1}{X}\right] = \frac{1}{\mathbb{E}[X]}$，等等。这些断言通常并不成立。只有随机变量的和、差以及常数倍才具有这种性质。这些表现得如此良好的随机变量。

4.3 Applications of Linearity of Expectation

Now let us see some examples of Theorem 15.1 in action.

1. **Two dice again.** Here is a much less painful way of computing $\mathbb{E}[X]$, where X is the sum of the scores of the two dice. Note that $X = Y_1 + Y_2$, where Y_i is the score on die i. We know from example 1 in Section 4.1 that $\mathbb{E}[Y_1] = \mathbb{E}[Y_2] = \frac{7}{2}$. So, by Theorem 15.1, we have $\mathbb{E}[X] = \mathbb{E}[Y_1] + \mathbb{E}[Y_2] = 7$.

2. **More roulette.** Suppose we play the roulette game mentioned in Section 4.1 $n \geq 1$ times. Let X_n be our expected net winnings. Then $X_n = Y_1 + Y_2 + \cdots + Y_n$, where Y_i is our net winnings in the i-th play. We know from earlier that $\mathbb{E}[Y_i] = -\frac{1}{19}$ for each i. Therefore, by Theorem 15.1, $\mathbb{E}[X_n] = \mathbb{E}[Y_1] + \mathbb{E}[Y_2] + \cdots + \mathbb{E}[Y_n] = -\frac{n}{19}$. For $n = 1000$, $\mathbb{E}[X_n] = -\frac{1000}{19} \approx -53$, so if you play 1000 games, you expect to lose about \$53.

3. **Fixed points of permutations.** Let us return to the homework permutation example with an arbitrary number n of students. Let X_n denote the number of students who receive their own homework after shuffling (or equivalently, the number of fixed points). To take advantage of Theorem 15.1, we need to write X_n as a *sum* of simpler r.v.'s. Since X_n *counts* the number of times something happens, we can write it as a sum using the following useful trick:

$$X_n = I_1 + I_2 + \cdots + I_n, \qquad \text{where } I_i = \begin{cases} 1, & \text{if student } i \text{ gets their own homework,} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

   [You should think about this equation for a moment. Remember that all the I_i's are random variables. What does an equation involving random variables mean? What we mean is that, *at every sample point* ω, we have $X_n(\omega) = I_1(\omega) + I_2(\omega) + \cdots + I_n(\omega)$. Why is this true?]

   A Bernoulli random variable such as I_i is called an indicator random variable of the corresponding event (in this case, the event that student i gets their own homework). For indicator r.v.'s, the expectation is particularly easy to calculate. Specifically,

$$\mathbb{E}[I_i] = (0 \times \mathbb{P}[I_i = 0]) + (1 \times \mathbb{P}[I_i = 1]) = \mathbb{P}[I_i = 1].$$

   In our case, we have

$$\mathbb{P}[I_i = 1] = \mathbb{P}[\text{student } i \text{ gets their own homework}] = \frac{1}{n}.$$

   We can now apply Theorem 15.1 to (5), yielding

$$\mathbb{E}[X_n] = \mathbb{E}[I_1] + \mathbb{E}[I_2] + \cdots + \mathbb{E}[I_n] = n \times \frac{1}{n} = 1.$$

   So, we see that the expected number of students who get their own homeworks in a class of size n is 1. That is, the expected number of fixed points in a random permutation of n items is always 1, regardless of n.

4. **Coin tosses.** Toss a fair coin $n \geq 1$ times. Let the r.v. X_n be the number of heads observed. As in the previous example, to take advantage of Theorem 15.1 we write

$$X_n = I_1 + I_2 + \cdots + I_n,$$

4.3 期望线性性质的应用

现在让我们看一些定理15.1的应用实例。

1. 再谈两个骰子。这里有一种更简便的方法来计算E[X]，其中X是两个骰子点数之和。注意X = Y1 +Y2，其中Yi是第i个骰子的点数。我们从第4.1节的例子1中知道E[Y1] = E[Y2] = $\frac{7}{2}$。因此，根据定理15.1，我们有E[X] = E[Y1]+E[Y2] = 7。

2. 更多轮盘赌。假设我们玩第4.1节中提到的轮盘赌游戏n ⩾1次。令Xn为我们预期的净赢钱。则Xn = Y1 +Y2 + · · ·+Yn，其中Yi为我们第i次下注的净赢钱。我们从前文知道E[Yi] = $-\frac{1}{19}$ 对每个i成立。因此，根据定理15.1，E[Xn] = E[Y1]+E[Y2]+· · ·+E[Yn] = $-\frac{n}{19}$。当n = 1000时，E[Xn] = $-\frac{1000}{19}$ ≈−53，因此如果你玩1000次游戏，你预计会输掉大约 53 美元。

3. 排列的不动点。让我们回到任意数量 n 名学生的作业排列例子。令 Xn 表示洗牌后收到自己作业的学生人数（或等价地，不动点的数量）。为了利用定理 15.1，我们需要将 Xn 写成一些更简单随机变量的和。由于 Xn 统计的是某件事情发生的次数，我们可以使用以下有用技巧将其写成一个和：

$$X_n = I_1 + I_2 + \cdots + I_n, \text{其中 } I_i = \begin{cases} 1\text{，如果学生 } i \text{ 拿到自己的作业，否则} \\ \text{为 } 0\text{。} \end{cases} \tag{5}$$

[你应该花点时间思考一下这个等式。记住所有 Ii 都是随机变量。包含随机变量的等式意味着什么？我们的意思是，在每一个样本点 ω 上，都有 Xn(ω) = I1(ω)+I2(ω)+· · ·+In(ω)。为什么这是真的？]

像 Ii 这样的伯努利随机变量称为对应事件的指示随机变量（在本例中，即学生 i 拿到自己作业的事件）。对于指示随机变量，期望特别容易计算。具体来说，

$$E[I_i] = (0\times P[I_i = 0]) + (1\times P[I_i = 1]) = P[I_i = 1]\text{。}$$

在我们的情况下，有

$$P[I_i = 1] = P[\text{学生 } i \text{ 拿到自己的作业}] = \frac{1}{n}.$$

现在我们可以将定理 15.1 应用于 (5)，得到

$$\mathbb{E}[X_n] = \mathbb{E}[I_1] + \mathbb{E}[I_2] + \cdots + \mathbb{E}[I_n] = n \times \frac{1}{n} = 1.$$

因此，我们看到在一个大小为 n 的班级中，期望有 1 名学生拿到自己的作业。也就是说，在 n 个元素的随机排列中，不动点的期望数量恒为 1，与 n 无关。

4. 抛硬币。抛一枚公平硬币 n ⩾1 次。设随机变量 Xn 为观察到的正面次数。与前例类似，为了利用定理 15.1，我们写作

$$X_n = I_1 + I_2 + \cdots + I_n,$$

where I_i is the indicator r.v. of the event that the i-th toss is H. Since the coin is fair, we have

$$\mathbb{E}[I_i] = \mathbb{P}[I_i = 1] = \mathbb{P}[i\text{-th toss is } H] = \tfrac{1}{2}.$$

Using Theorem 15.1, we therefore get

$$\mathbb{E}[X_n] = \sum_{i=1}^{n} \frac{1}{2} = \frac{n}{2}.$$

In n tosses of a biased coin that shows H with probability p, $\mathbb{E}[X_n] = np$. (Check this.) So the expectation of a binomial r.v. $X \sim \text{Bin}(n, p)$ is equal to np. Note that it would have been harder to reach the same conclusion by computing this directly from the definition of expectation shown in (3).

5. **Balls and bins.** Throw m balls into n bins. Let the r.v. X be the number of balls that land in the first bin. Then X behaves exactly like the number of heads in m tosses of a biased coin with $\mathbb{P}[H] = \frac{1}{n}$ (why?). So, from the previous example, we get $\mathbb{E}[X] = \frac{m}{n}$. In the special case $m = n$, the expected number of balls in any bin is 1. If we wanted to compute this directly from the distribution of X, we would get into a messy calculation involving binomial coefficients.

   Here is another example on the same sample space. Let the r.v. Y_n be the number of empty bins. The distribution of Y_n is horrible to contemplate: to get a feel for this, you might like to write it down for $m = n = 3$ (i.e., 3 balls, 3 bins). However, computing the expectation $\mathbb{E}[Y_n]$ is easy using Theorem 15.1. As in previous two examples, we write

$$Y_n = I_1 + I_2 + \cdots + I_n, \tag{6}$$

   where I_i is the indicator r.v. of the event "bin i is empty". The expectation of I_i is easy to find:

$$\mathbb{E}[I_i] = \mathbb{P}[I_i = 1] = \mathbb{P}[\text{bin } i \text{ is empty}] = \left(1 - \frac{1}{n}\right)^m,$$

   as discussed earlier. Applying Theorem 15.1 to (6), we therefore obtain

$$\mathbb{E}[Y_n] = \sum_{i=1}^{n} \mathbb{E}[I_i] = n\left(1 - \frac{1}{n}\right)^m,$$

   a simple formula, quite easily derived. Let us see how it behaves in the special case $m = n$ (same number of balls as bins). In this case we get $\mathbb{E}[Y_n] = n\left(1 - \frac{1}{n}\right)^n$. Now the quantity $\left(1 - \frac{1}{n}\right)^n$ can be approximated (for large enough values of n) by the number $\frac{1}{e}$.[2] So we see that, for large n,

$$\mathbb{E}[Y_n] \approx \frac{n}{e} \approx 0.368n.$$

   The bottom line is that, if we throw (say) 1000 balls into 1000 bins, the expected number of empty bins is about 368.

[2] More generally, it is a standard fact that for any constant c,

$$(1 + \tfrac{c}{n})^n \to e^c \quad \text{as } n \to \infty.$$

We just used this fact in the special case $c = -1$. The approximation is actually very good even for quite small values of n. (Try it yourself!) E.g., for $n = 20$ we already get $(1 - \frac{1}{n})^n \approx 0.358$, which is very close to $\frac{1}{e} \approx 0.368$. The approximation gets better and better for larger n.

其中 Ii 是第 i 次投掷结果为 H 这一事件的指示随机变量。由于硬币是公平的，我们

有 E[Ii] = P[Ii = 1] = P[第 i 次投掷为 H] = 1 $\frac{}{2}$.

因此，使用定理 15.1 我们得到

$$\mathbb{E}[X_n] = \sum_{i=1}^{n} \frac{1}{2} = \frac{n}{2}.$$

在 n 次掷出正面概率为 p 的偏置硬币实验中，E[Xn] = np。（请验证这一点。）因此，服从二项分布的随机变量 X ~Bin(n, p) 的期望等于 np。注意，如果直接从 (3) 中所示的期望定义出发进行计算，要得出相同结论会更困难。

5. 球与箱子。将 m 个球扔进 n 个箱子中。设随机变量 X 表示落在第一个箱子中的球的数量。那么 X 的行为完全类似于掷一枚偏置硬币 m 次所出现正面的次数，其中 P[H] = 1 $\frac{}{n}$（为什么？）。因此，从前一个例子我们得到 E[X] = m $\frac{}{n}$。在 m = n 的特殊情况下，期望任意一个箱子中球的期望数量为 1。如果我们想直接从 X 的分布来计算这个值，将会陷入涉及二项式系数的复杂计算中。

这是同一样本空间上的另一个例子。设随机变量 Yn 为没有球的箱子数量。Yn 的分布非常复杂难以描述：为了体会这一点，你可以尝试写出当 m = n = 3（即 3 个球，3 个箱子）时的情形。然而，使用定理 15.1 计算期望 E[Yn] 却很容易。如同前面两个例子一样，我们写作

$$Y_n = I_1 + I_2 + \cdots + I_n, \tag{6}$$

其中 Ii 是事件“箱子 i 为空”的指示随机变量。Ii 的期望很容易求得：

$$\text{E[Ii] = P[Ii = 1] = P[箱子 i 为空]} = \left(1 - \frac{1}{n}\right)^m,$$

如前所述。将定理 15.1 应用于 (6)，我们因此得到

$$\mathbb{E}[Y_n] = \sum_{i=1}^{n} \mathbb{E}[I_i] = n\left(1 - \frac{1}{n}\right)^m,$$

一个简单的公式，很容易推导出来。让我们看看它在 m = n（球的数量与箱子数量相同）这一特殊情况下的表现。在这种情况下，我们得到 E[Yn] = n(1− $\frac{}{n}$)n。现在考虑量 1−1($\frac{}{n}$)n 可以是对于足够大的 n 值，近似于数值 1e.2。因此我们看到，当 n 很大时，E[Yn] ≈

$\frac{n}{e}$ ≈0.368n。

关键是，如果我们把（比如说）1000个球扔进1000个箱子中，空箱子的期望数量约为368。

2更一般地，这是一个标准结论：对于任意常数c，(1+

c $\frac{}{n}$)n →ec 当 n →∞。

我们刚刚在c = −1的特殊情况下使用了这一事实。该近似实际上即使对于相当小的n值也非常准确。（你自己试试！）例如，当n = 20时，我们已经得到(1−1 $\frac{}{n}$)n ≈0.358，这非常接近于1 $\frac{}{e}$ ≈0.368。该近似随着n增大而更精确，并且对于较大的n更好。