Chapter 4

Language of Probabilities: Discrete Random Variables

This chapter is about another element of the language of probabilities, namely, *random variables*. Random variables provide the vocabulary for talking and reasoning about random quantities in probability models. For instance, the height and the weight of a randomly picked student from a class can be formulated as random variables in a model which describes the experiment of picking a student from the class at random.

In this chapter, we focus on *discrete* random variables, that is, those with discrete range of values. Those with continuous range of values and those which do not fit in this division will be dealt with later. Moreover, our discussion in this and the following chapter will be limited to individual random variables, or collections of random variables that are statistically independent of one another. Inter-dependence between random variables will be discussed later.

4.1 What is a random variable

Example 4.1.1 (Height and weight of a random student)**.** Consider the height (in centimeters) and the weight (in kilograms) of a student chosen at random from a class.

(Q) How should we incorporate these quantities in our model of probabilities?

The model for the experiment of picking a student from a class at random is described by

- (sample space) Ω: set of all students in the class,
- (measure of probabilities) $\mathbb{P}(\omega) = 1/|\Omega|$ for each outcome $\omega \in \Omega$ (i.e., all students are equally likely to be chosen).

In this framework,

- Each student $a \in \Omega$, has a height $H(a)$ and a weight $W(a)$.

The two functions H and W are examples of random variables. ○

Terminology. A *random variable* (sometimes abbreviated as *RV*) is a numerical quantity which is determined by the outcome of an experiment. Mathematically, a random variable X is simply a function that assigns a number $X(a)$ to each outcome $a \in \Omega$. The set of *possible values* that X can take is sometimes referred to as its *range*.

More examples.

(1) Experiment: roll two 6-sided dice. An example of a random variable:

$$X \coloneqq \langle \text{sum of the two numbers on the dice} \rangle \ .$$

(Q) What are the possible values of X?

[A] $2, 3, 4, \ldots, 12$.

(2) Experiment: flip a coin 10 times. An example of a random variable:

$$X \coloneqq \langle \text{number of heads} \rangle \ .$$

Ⓠ What are the possible values of X?

🅰 $0, 1, 2, \ldots, 10$.

③ Experiment: repeat flipping a coin until there is a head. An example of a random variable:

$$N \coloneqq \langle \text{total number of flips} \rangle \ .$$

Ⓠ What are the possible values of N?

🅰 $1, 2, 3, \ldots$.

④ Experiment: repeat flipping a coin until there are two heads. Two examples of random variables:

$$N_1 \coloneqq \langle \text{the number of flips until and including the 1st head} \rangle \ ,$$
$$N_2 \coloneqq \langle \text{the number of flips until and including the 2nd head} \rangle \ .$$

Ⓠ What are the possible values of N_1 and N_2?

🅰 Possible values of N_1: $1, 2, 3, \ldots$.
Possible values of N_2: $2, 3, 4, \ldots$.

⑤ Experiment: draw a number from the interval $[-1, 1]$ at random. An example of a random variable:

$$X \coloneqq \langle \text{square of the drawn number} \rangle \ .$$

Ⓠ What are the possible values of X?

🅰 $[0, 1]$.

Convention. We use capital letters (such as $X, N, A, \ldots$) for naming random variables and use small letters (such as $x, n, a, \ldots$) to refer to their possible values.

Following this convention would greatly help us not get confused when dealing with random variables. We will often need to refer to events such as $\{X = x\}$ or $\{X \leq x\}$ where x is a possible value for a random variable X. For instance, this convention allows us to define a function such as

$$p(x) \coloneqq \mathbb{P}(X = x)$$

concisely and unambiguously. This is simply a function that

- with input 1, gives out $p(1) = \mathbb{P}(X = 1)$ (the probability of the event $\{X = 1\}$),
- with input 10, gives out $p(10) = \mathbb{P}(X = 10)$ (the probability of the event $\{X = 10\}$),
- with input 2.5, gives out $p(2.5) = \mathbb{P}(X = 2.5)$ (the probability of the event $\{X = 2.5\}$),
- $\ldots$

Notation. The notation $\mathbb{P}(X = x)$ is a compact way to refer to the probability that $X = x$. In other words, this is the probability of the event $\{X = x\}$, which consists of all possible outcomes $\omega \in \Omega$ for which $X(\omega) = x$. Therefore, in our notation,

$$\mathbb{P}(X = x) = \mathbb{P}(\{X = x\}) = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) = x\}\right) \ .$$

Types of random variables. In this chapter, we study random variables that have only finitely many, or countably many possible values. Such random variables are known as *discrete* random variables. Of the five examples above, all except ⑤ are discrete. Handling discrete random variables is significantly simpler than the other types of random variables. In particular, the distribution of a discrete random variable can simply be described by identifying the probability that the random variable takes each of its possible values.

Example ⑤ is an example of another type of random variables, known as *continuous* random variables. The possible values of a continuous random variable range over entire intervals.[1] While working with discrete random variables involves calculating finite or infinite sums and counting objects, dealing with continuous random variables involves calculating integrals and using calculus.

Not every random variable falls into one of the two categories of discrete and continuous random variables. However, in many scenarios which we encounter in practice, discrete and continuous random variables seem to be sufficient in modeling random quantities. In a later chapter, we will discuss other types of random variables.

[1] The exact definition of what is called a continuous random variable will be discussed later.

4.2 Distribution of a random variable: part I

4.2.1 Describing the distribution

Example 4.2.1 (Rolling two dice)**.** Suppose we roll two fair dice. Let X denote the sum of the two numbers that appear on the dice.

An appropriate model for this experiment can be given by

- (sample space) $\Omega := \{\boxdot\boxdot, \boxdot\boxdot, \ldots, \boxdot\boxdot\}$,
- (measure of probabilities) $\mathbb{P}(\omega) := 1/|\Omega| = 1/36$ for each $\omega \in \Omega$ (i.e., all outcomes are equally likely).

In this model, X is a random variable. For instance, when the outcome is $\boxdot\boxdot$, the value of X is 8, that is, $X(\boxdot\boxdot) = 8$. In general, for each possible outcome (a, b) (where a is the number appearing on the first die, and b is the number appearing on the second die), we have $X\big((a, b)\big) := a + b$. Table 4.1 shows the value of X for each possible outcome.

X:	⚀	⚁	⚂	⚃	⚄	⚅
⚅	7	8	9	10	11	12
⚄	6	7	8	9	10	11
⚃	5	6	7	8	9	10
⚂	4	5	6	7	8	9
⚁	3	4	5	6	7	8
⚀	2	3	4	5	6	7

die #2 / die #1

Table 4.1: Sum of the numbers on two dice

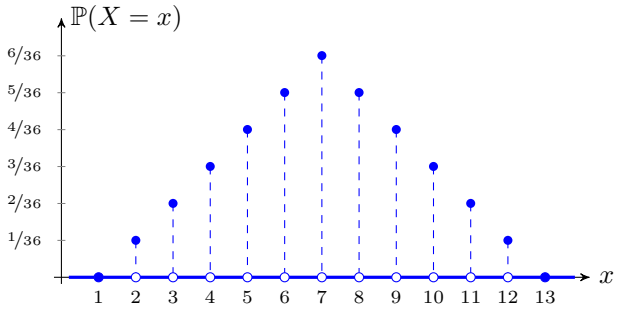


Figure 4.1: Distribution of the sum of the numbers on two dice

The possible values of X are

$$2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 .$$

However, observe that these values are not equally likely. For instance, $X = 2$ happens only if the outcome is $\boxdot\boxdot$, whereas there are six possible outcomes for which $X = 7$, namely,

$$X(\boxdot\boxdot) = X(\boxdot\boxdot) = X(\boxdot\boxdot) = X(\boxdot\boxdot) = X(\boxdot\boxdot) = X(\boxdot\boxdot) = 7 .$$

This means that it is six times more likely that $X = 7$ than $X = 2$.

Q What is the probability that X takes each of its possible values?

A

x :	2	3	4	5	6	7	8	9	10	11	12
$\mathbb{P}(X = x)$:	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

From Table 4.1, we see that there is one outcome realizing $X = 2$, two outcomes realizing $X = 3, \ldots,$ and one outcome realizing $X = 12$. Since each individual outcome has probability $1/36$, we have $\mathbb{P}(X = 2) = 1/36$, $\mathbb{P}(X = 3) = 2/36, \ldots, \mathbb{P}(X = 12) = 1/36$.

The function $p(x) := \mathbb{P}(X = x)$ is called the *probability mass function* of the random variable X. In this example, $p(x)$ is given by the above table when x is a possible value of X, and $p(x) := 0$ when x is not one of the possible values. Figure 4.1 shows the graph of $p(x)$. ○

Example 4.2.2 (Flipping a coin 10 times)**.** Suppose we flip a coin 10 times. Let X denote the number of heads. Let p be the parameter of the coin, indicating the chance of getting a head in one flip.

The model for this experiment is described by

- (sample space) $\Omega := \{\mathtt{H}, \mathtt{T}\}^{10}$,
- (measure of probabilities) $\mathbb{P}$ is given by

$$\mathbb{P}(\mathtt{HHHHHHHHHH}) = p^{10} , \quad \mathbb{P}(\mathtt{HHHHHHHHHT}) = p^9(1-p) , \quad \ldots, \quad \mathbb{P}(\mathtt{TTTTTTTTTT}) = (1-p)^{10} .$$

In this model, X is a random variable given by

$$X(a_1a_2\cdots a_{10}) \coloneqq \langle \#\text{ of H's in } a_1a_2\cdots a_{10}\rangle$$

for each outcome $a_1a_2\cdots a_{10} \in \Omega$.

The possible values of X are $0, 1, 2, \ldots, 10$.

Q What is the probability that X takes each of its possible values?

A For $x = 0, 1, 2, \ldots, 10$, we have

$$\mathbb{P}(X = x) = \binom{10}{x} p^x (1-p)^{10-x} .$$

There are $\binom{10}{x}$ outcomes with exactly x heads and $10 - x$ tails, each of which has probability $p^x(1-p)^{10-x}$.

The probability mass function of the random variable X is thus

$$p_X(x) \coloneqq \mathbb{P}(X = x) = \begin{cases} \binom{10}{x} p^x (1-p)^{10-x} & \text{if } x \in \{0, 1, 2, \ldots, 10\}, \\ 0 & \text{otherwise.} \end{cases}$$

Another function associated to the random variable X is its so-called *cumulative distribution function*, which is defined as

$$F_X(x) \coloneqq \mathbb{P}(X \leq x) .$$

Q What is the value of $F_X(x)$ for each $x \in \mathbb{R}$?

A

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{if } x < 0, \\ \displaystyle\sum_{k=0}^{\lfloor x \rfloor} \binom{10}{k} p^k (1-p)^{10-k} & \text{if } 0 \leq x \leq 10, \\ 1 & \text{if } 10 < x. \end{cases}$$

If $x < 0$, then the event $X \leq x$ never happens, hence $\mathbb{P}(X \leq x) = 0$. If $x > 10$, then the event $X \leq x$ always happens, thus $\mathbb{P}(X \leq x) = 1$. If $x = 0, 1, \ldots, 10$, then the event $X \leq x$ happens if either $X = 0$, or $X = 1$, or ..., $X = x$. Note that if $0 \leq x \leq 10$ but x is not an integer, then the events $X \leq x$ is the same as the event $X \leq \lfloor x \rfloor$, where $\lfloor x \rfloor$ denotes the largest integer k satisfying $k \leq x$. For instance, $X \leq 2.7$ happens precisely when $X \leq 2$ happens. ○

Terminology. The *probability mass function* (often abbreviated as *pmf*) of a discrete random variable X is a function $p_X(x)$ with one input argument x, defined as

$$p_X(x) \coloneqq \mathbb{P}(X = x) \qquad \text{for each } x \in \mathbb{R}.$$

It describes the *distribution* of X, that is, how likely it is for X to take each value. Based on the interpretation of the probabilities as "idealized frequencies", the value of $p_X(x)$ is roughly the proportion of times in which $X = x$ if we repeat the experiment many many times.

The distribution of a discrete random variable X can alternatively be described by its *cumulative distribution function* (often abbreviated as *cdf*), which is defined as

$$F_X(x) \coloneqq \mathbb{P}(X \leq x) \qquad \text{for each } x \in \mathbb{R}.$$

Relationship between the pmf and the cdf. The two functions $p_X(x)$ and $F_X(x)$ carry the same information about a discrete random variable X: if we know one of them, we can find the other.

Q If we know the pmf of a discrete random random variable X, how can we find its cdf?

A

$$F_X(x) = \sum_{y : y \leq x} p_X(y) \qquad \text{for each } x \in \mathbb{R}.$$

Q If we know the cdf of a discrete random variable X, how can we find its pmf?

A Suppose we know the cdf and we want to find the value of the pmf $p_X(x)$ for some value x. Let y be another value which is strictly smaller than x but very close to it (say, $y = x - 0.001$). Then, the probability that $y < X < x$ will be small, and hence

$$\begin{aligned} p_X(x) = \mathbb{P}(X = x) &= \mathbb{P}(X \le x) - \mathbb{P}(X < x) \\ &\approx \mathbb{P}(X \le x) - \mathbb{P}(X \le y) = F_X(x) - F_X(y) \,. \end{aligned}$$

As we choose y closer and closer to x, the approximation becomes more and more accurate, and in the limit $y \nearrow x$, we get an exact equality. In fact, if y is close enough to x such that X has no possible value between y and x, then the approximation is already an exact equality and we can calculate $p_X(x)$ as $F_X(x) - F_X(y)$.

While the meaning of the pmf is clear, the relevance of the cdf is less obvious. If the pmf and the cdf carry the same information, why not stick to the simpler, more intuitive one? In some scenarios, working with the cdf simplifies the calculations. However, the main advantage of the cdf is that its domain of applicability is much broader. As we will see later, the pmf is only useful in describing the distribution of *discrete* random variables, whereas the cdf can be used for any type of random variable. This will be clarified in the following chapters.

Example 4.2.3 (Flipping until a head comes up)**.** Suppose we repeat flipping a coin until it comes up heads. Let N denote the total number of flips in this experiment. As usual, we let p denote the bias parameter of the coin, that is, the probability that it comes up heads in one flip.

The model for this experiment can be described by

- (sample space) $\Omega := \{\mathtt{H}, \mathtt{TH}, \mathtt{T}^2\mathtt{H}, \mathtt{T}^3\mathtt{H}, \ldots\}$,
- (measure of probabilities) $\mathbb{P}(\mathtt{T}^k\mathtt{H}) = (1-p)^k p$ for each $k = 0, 1, 2, \ldots$.

In this model, N is a random variable given by

$$N(\mathtt{T}^k\mathtt{H}) := k + 1 \qquad \text{for each } k = 0, 1, 2, \ldots.$$

The possible values of N are $1, 2, 3, \ldots$.

Q What is the probability mass function of N?

A

$$p(n) := \mathbb{P}(N = n) = \begin{cases} (1-p)^{n-1}p & \text{if } n = 1, 2, 3, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

Q What is the cumulative distribution function of N?

A1 For $n = 1, 2, 3, \ldots$, the cdf is given by

$$F(n) := \mathbb{P}(N \le n) = \sum_{k=1}^{n} (1-p)^{k-1} p \,.$$

This is a geometric sum,[2] which can be calculated as

$$= \frac{1 - (1-p)^n}{1 - (1-p)} p = 1 - (1-p)^n \,.$$

For general values of n (not necessarily positive, not necessarily an integer), we have

$$F(n) := \mathbb{P}(N \le n) = \begin{cases} 0 & \text{if } n < 1, \\ 1 - (1-p)^{\lfloor n \rfloor} & \text{if } n \ge 1. \end{cases}$$

A2 When $n = 1, 2, 3, \ldots$, we can use the following trick to simplify the calculation, hence avoiding the geometric sum:

$$F(n) := \mathbb{P}(N \le n) = 1 - \mathbb{P}(N > n)$$

Observe that $\{N > n\}$ is simply the event that there are no H's in the first n flips, hence $\mathbb{P}(N > n) = (1-p)^n$. Therefore,

$$F(n) = 1 - (1-p)^n \,.$$

The value of $F(n)$ for other values of n (not necessarily positive, not necessarily an integer) is as in the previous answer. ○

[2] A quick review of geometric sums comes in Interlude 4.A.

Interlude 4.A (Review of geometric sums)**.** A *geometric sum* is the sum of a finite number of terms in which the ratio between every two consecutive terms is the same. For instance, in the sum

$$S = 3 + \frac{3}{2} + \frac{3}{4} + \frac{3}{8} + \cdots + \frac{3}{2^{10}} ,$$

each term is half the previous term. The value of a geometric sum can be found using a similar trick as in the case of geometric series. Namely, observe that

$$S = 3 + \frac{1}{2}\Big(\underbrace{3 + \frac{3}{2} + \frac{3}{8} + \cdots + \frac{3}{2^9}}_{S - 3/2^{10}}\Big) .$$

Hence, $S = 3 + (1/2)(S - 3/2^{10})$. Solving this equation, we obtain $S = 6 \times (1 - 1/2^{11}) = 5.9970703125$.

More generally, a geometric sum has the form

$$S = a + ar + ar^2 + ar^3 + \cdots ar^n = \sum_{k=0}^{n} ar^k , \qquad (\approx)$$

where a is the starting term, r is the common ratio of the consecutive terms, and there are $n + 1$ terms in total. The value of the sum $(\approx)$ can be found as in the example above. ◇

4.2.2 Bernoulli random variables

A random variable with exactly two possible values 0 and 1 is called a *Bernoulli random variable*.[3] This is the simplest type of random variable. The distribution of a Bernoulli random variable X is identified by a single parameter p, indicating the probability of $X = 1$. The probability of $X = 0$ is then simply $1 - p$. The probability mass function of a Bernoulli random variable with parameter p is thus

$$p(x) \coloneqq \begin{cases} p & \text{if } x = 1, \\ 1 - p & \text{if } x = 0, \\ 0 & \text{otherwise.} \end{cases}$$

The distribution of X is known as the *Bernoulli distribution* with parameter p.

Bernoulli random variables naturally arise in coin flipping experiments, but also in many other contexts. Here are some examples of Bernoulli random variables in three different contexts:

(B1) Experiment: flip a coin 5 times. For $k = 1, 2, \ldots, 5$, let

$$X_k \coloneqq \begin{cases} 1 & \text{if the } k\text{'th flip comes up heads,} \\ 0 & \text{otherwise.} \end{cases}$$

Then, $X_1, X_2, \ldots, X_5$ are Bernoulli random variables.

(Q) What are the parameters of these Bernoulli random variables?

[A] The same as the parameter of the coin (i.e., the chance of getting a head in one flip).

(B2) Experiment: roll two fair dice. Let

$$X \coloneqq \begin{cases} 1 & \text{if the two dice show the same number,} \\ 0 & \text{otherwise.} \end{cases}$$

This is a Bernoulli random variable.

(Q) What is the parameter of this Bernoulli random variable?

[A] $1/6$. There are 6 possibilities ⚀⚀, ⚁⚁, ⚂⚂, ⚃⚃, ⚄⚄, ⚅⚅, each occurring with probability $1/36$.

(B3) Experiment: pick a student from a class at random. Let

$$H \coloneqq \begin{cases} 1 & \text{if the height of the picked student is} \geq 170\,\text{cm,} \\ 0 & \text{otherwise.} \end{cases}$$

This is a Bernoulli random variable.

[3] Named after mathematician Jacob Bernoulli (1655–1705).

Ⓠ What is the parameters of this Bernoulli random variable?

Ⓐ The proportion of students in the class with height ≥ 170 cm.

Bernoulli random variables can often be used to indicate whether a certain event has happened or not. As we will see, this sometimes allows us express other types of random variables in much simpler terms.

4.2.3 Graphs of pmf and cdf

To understand the concepts of the probability mass function and the cumulative distribution function, it is instructive to know how their graphs look like.

Example 4.2.4 (Bernoulli RV with parameter p)**.** Let X be a Bernoulli random variable with parameter p, that is,

$$X = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1-p. \end{cases}$$

Ⓠ How does the pmf of X look like?

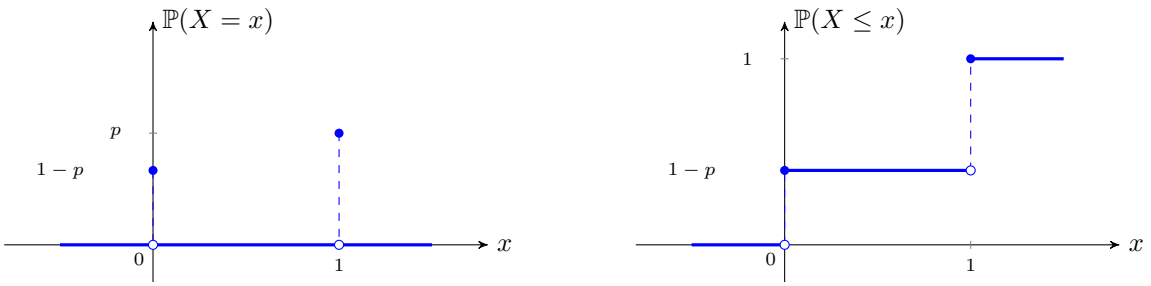
Ⓐ See Figure 4.2a. Note that $\mathbb{P}(X = x) = 0$ unless x is one of the two possible values.

Ⓠ How does the cdf of X look like?

Ⓐ See Figure 4.2b. Note that

- If $x < 0$, then the event $X \leq x$ never happens, hence $\mathbb{P}(X \leq x) = 0$.
- If $0 \leq x < 1$, then $X \leq x$ if and only if $X = 0$. Therefore, $\mathbb{P}(X \leq x) = \mathbb{P}(X = 0) = 1 - p$.
- If $x \geq 1$, then the event $X \leq x$ always happens, hence $\mathbb{P}(X \leq x) = 1$.

○



(a) (b)

Figure 4.2: The distribution of a Bernoulli RV with parameter p. (a) pmf (b) cdf

Example 4.2.5 (Number on a die)**.** Suppose we roll a fair die. Let X be the number appearing on the die. Hence,

$$X = \begin{cases} 1 & \text{with probability } 1/6, \\ 2 & \text{with probability } 1/6, \\ \vdots & \quad \vdots \\ 6 & \text{with probability } 1/6. \end{cases}$$

Ⓠ How does the pmf of X look like?

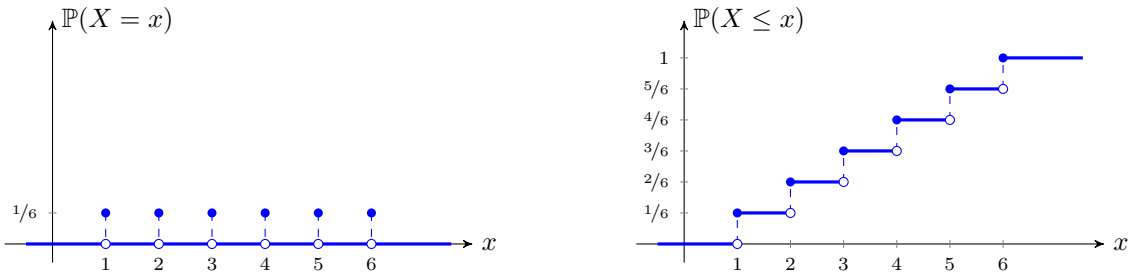
Ⓐ See Figure 4.3a. Note that $\mathbb{P}(X = x) = 0$ unless x is one of the six possible values.

Ⓠ How does the cdf of X look like?

Ⓐ See Figure 4.3b. Note that

- If $x < 1$, then the event $X \le x$ never happens, hence $\mathbb{P}(X \le x) = 0$.
- If $1 \le x < 2$, then the event $X \le x$ happens if and only if $X = 1$. Therefore, $\mathbb{P}(X \le x) = \mathbb{P}(X = 1) = 1/6$.
- If $2 \le x < 3$, then the event $X \le x$ happens if either $X = 1$ or $X = 2$. Therefore, $\mathbb{P}(X \le x) = \mathbb{P}(X = 1) + \mathbb{P}(X = 2) = 2/6$.
- ...
- If $x \ge 1$, then the event $X \le x$ always happens, hence $\mathbb{P}(X \le x) = 1$.

○



(a) (b)

Figure 4.3: The distribution of the number on a fair die. (a) pmf (b) cdf

Shape of the cdf (discrete RVs). Let $F(x)$ be the cdf of a *discrete* random variable X.[4] Then,

(i) $F(x)$ is non-decreasing.

(ii) $F(x) = 0$ if x is smaller than all the possible values of X.[5]

(iii) $F(x) = 1$ if x is larger than all the possible values of X.[6]

(iv) At each possible value x of X, the graph of F has an upward jump. The amount of the jump is $\mathbb{P}(X = x)$.

(v) In between the possible values of X, $F(x)$ is constant.[7]

Exercise. By examining the above two examples or by using mathematical reasoning, argue that the above properties indeed hold for the cdf of every discrete random variable.

4.2.4 Independence of random variables

Example 4.2.6 (Flipping a coin 5 times)**.** Consider again the experiment of flipping a coin 5 times. For $k = 1, 2, \ldots, 5$, consider the Bernoulli random variables

$$X_k := \begin{cases} 1 & \text{if the } k\text{'th flip comes up heads,} \\ 0 & \text{otherwise.} \end{cases}$$

The random variables $X_1, X_2, \ldots, X_5$ indicate the results of the 5 flips. Since the flips do not affect one another, the values of $X_1, X_2, \ldots, X_5$ are independent of one another.

Q How should we formulate the fact that $X_1, X_2, \ldots, X_5$ are independent?

A The independence of $X_1, X_2, \ldots, X_5$ is perfectly captured by the fact that the events

$$\{X_1 = 1\}, \quad \{X_2 = 1\}, \quad \{X_3 = 1\}, \quad \{X_4 = 1\}, \quad \{X_5 = 1\}$$

are independent. ○

[4] The cdf of non-discrete random variables has some but no all the above properties. For instance, for a continuous random variable, the cdf is a continuous function.

[5] If no such x exists, then we still have $F(x) \to 0$ as $x \to -\infty$.

[6] If no such x exists, then we still have $F(x) \to 1$ as $x \to \infty$.

[7] In particular, if x_1, x_2 are two consecutive possible values of X, then $F(x) = F(x_1)$ whenever $x_1 \le x < x_2$.

Terminology. Two discrete random variables X and Y are said to be (statistically) *independent* if for every values x, y, the events

$$\{X = x\} \qquad \text{and} \qquad \{Y = y\}$$

are independent. This means that the value that X takes does not provide us with any information about the value that Y takes, and vice versa.

The independence of more than two random variables can be formulated analogously.

Exercise. Verify that the random variables $X_1, X_2, \ldots, X_5$ in Example 4.2.6 are independent in the above sense. For instance, you need to show that the events

$$\{X_1 = 1\}, \quad \{X_2 = 1\}, \quad \{X_3 = 1\}, \quad \{X_4 = 1\}, \quad \{X_5 = 0\}$$

are also independent, and similarly for any combination of 0s and 1s as values.

4.3 Expected value of a random variable: part I

Example 4.3.1 (Number on a die)**.** Let us go back to the experiment of rolling a fair die (Example 4.2.5). Suppose we repeat rolling the die many many times.

Q What will be the average of the numbers appearing on the die?

A Let X denote the number appearing on the die. This is a random variable, where

$$X = \begin{cases} 1 & \text{with probability } 1/6, \\ 2 & \text{with probability } 1/6, \\ \vdots & \quad \vdots \\ 6 & \text{with probability } 1/6. \end{cases}$$

Suppose we repeat the experiment of rolling the die n times, where n is very large. If x_k denotes the number we observe on the die in the k'th experiment, then

$$\begin{aligned} \big\langle \text{average of the values of } X \big\rangle &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &= \frac{N_n(X=1)\cdot 1 + N_n(X=2)\cdot 2 + \cdots + N_n(X=6)\cdot 6}{n}\,, \end{aligned}$$

where as usual, we have written $N_n(X = i)$ to denote the number of times among these repeated experiments in which $X = i$. Note that each of the terms in the sum $x_1 + x_2 + \cdots + x_n$ is either 1, or 2, ..., or 6. In the above equality, we have simply grouped the 1s together, the 2s together, and so on.[8] The latter can now be written as

$$= \frac{N_n(X=1)}{n}\cdot 1 + \frac{N_n(X=2)}{n}\cdot 2 + \cdots + \frac{N_n(X=6)}{n}\cdot 6\,.$$

The ratio $N_n(X = 1)/n$ is the proportion of times in which $X = 1$. Based on the interpretation of probabilities as "idealized frequencies", the latter proportion approximates $\mathbb{P}(X = 1)$. Similarly,

$$\frac{N_n(X=1)}{n} \approx \mathbb{P}(X=1)\,, \quad \frac{N_n(X=2)}{n} \approx \mathbb{P}(X=2)\,, \quad \cdots \quad \frac{N_n(X=6)}{n} \approx \mathbb{P}(X=6)\,.$$

Therefore,

$$\begin{aligned} \big\langle \text{average of the values of } X \big\rangle &\approx \mathbb{P}(X=1)\cdot 1 + \mathbb{P}(X=2)\cdot 2 + \cdots + \mathbb{P}(X=6)\cdot 6 \\ &= \frac{1}{6}\cdot 1 + \frac{1}{6}\cdot 2 + \cdots + \frac{1}{6}\cdot 6 = 3.5\,. \end{aligned}$$

[8] For instance, if $n = 10$ and the values of X in 10 repeated experiments are $3, 1, 5, 3, 2, 1, 1, 6, 4, 5$, then

$$\begin{aligned} \frac{3+1+5+3+2+1+1+6+4+5}{10} &= \frac{(1+1+1)+2+(3+3)+4+(5+5)+6}{10} \\ &= \frac{3\times 1 + 1\times 2 + 2\times 3 + 1\times 4 + 2\times 5 + 1\times 6}{10} \end{aligned}$$

In this case, $N_{10}(X=1)=3$, $N_{10}(X=2)=1$, $N_{10}(X=3)=2$, $N_{10}(X=4)=1$, $N_{10}(X=5)=2$ and $N_{10}(X=6)=1$.

The value

$$\mathbb{E}[X] \coloneqq \mathbb{P}(X=1)\cdot 1 + \mathbb{P}(X=2)\cdot 2 + \cdots + \mathbb{P}(X=6)\cdot 6 = 3.5$$

is called the *expected value* of X. It can be thought of as the "idealized average" of X in many many repeated experiments. ○

Example 4.3.2 (Bernoulli RV with parameter p)**.** Consider the experiment of flipping a coin with parameter p. Define a random variable X by setting $X = 1$ if the coin comes up heads and $X = 0$ if the coin comes up tails. Then, X is a Bernoulli random variable with parameter p (Example 4.2.4).

Suppose we repeat the experiment many many times?

Ⓠ What will be the average value of X in these repeated experiments?

A1 Suppose we repeat the experiment n times, where n is very large. As in the previous example, let x_k denote the value of X in the k'th experiment. Let $N_n(X=0)$ and $N_n(X=1)$ denote the number of times in which $X=0$ and $X=1$, respectively. Then,

$$\begin{aligned}\big\langle \text{average of the values of } X \big\rangle &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &= \frac{N_n(X=0)\cdot 0 + N_n(X=1)\cdot 1}{n} \\ &= \frac{N_n(X=0)}{n}\cdot 0 + \frac{N_n(X=1)}{n}\cdot 1 \\ &\approx \mathbb{P}(X=0)\cdot 0 + \mathbb{P}(X=1)\cdot 1 = p\,.\end{aligned}$$

A2 Observe that the average value of X is simply the proportion of times in which the coin comes up heads. According to the interpretation of probabilities as "idealized frequencies", the latter proportion approximates $\mathbb{P}(\mathtt{H}) = p$.

In this example, the *expected value* of X is the "idealized average"

$$\mathbb{E}[X] \coloneqq \mathbb{P}(X=0)\cdot 0 + \mathbb{P}(X=1)\cdot 1 = p\,.$$

○

Terminology. The *expected value* (or *expectation*, or *mean*) of a discrete random variable X, denoted by $\mathbb{E}[X]$, is the quantity

$$\mathbb{E}[X] \coloneqq \sum_x \mathbb{P}(X=x)\cdot x\,, \tag{$\varnothing_1$}$$

where the sum runs over all the possible values of X. In words, the expected value of X is the average of the possible values of X weighted by their probabilities.

If Ω denotes the underlying sample space, then the expected value of X can also be calculated as[9]

$$\mathbb{E}[X] = \sum_{\omega\in\Omega} \mathbb{P}(\omega)\cdot X(\omega)\,, \tag{$\varnothing_2$}$$

In words, the expected value of X is the average of the values that X takes for each possible outcome, weighted by its probability.

The expected value of a random variable X is often denoted by the Greek letter μ, or by μ_X if we want to emphasize the dependence on X.

Interpretation. The expected value of a random variable X is understood as the "idealized average" of the values of X in many many repeated experiments. More specifically, suppose that we repeat the experiment n times, and let x_k denote the value of X in the k'th experiment. Then,

$$\mathbb{E}[X] \approx \frac{x_1 + x_2 + \cdots + x_n}{n}$$

[9] To be precise, this is true if Ω is finite or countable.

when n is large, and the approximation becomes more and more accurate as $n \to \infty$. As usual, let $N_n(E)$ denote the number of time in these repeated experiments in which the event E happens. Expression ($\varnothing_1$) is obtained by grouping the terms in the sum $x_1 + x_2 + \cdots + x_n$ based on the possible values of X, and noting that

$$\frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_x N_n(X = x) \cdot x}{n} .$$

Expression ($\varnothing_2$) is obtained by grouping the terms in the sum $x_1 + x_2 + \cdots + x_n$ based on the outcome of the experiment, and noting that

$$\frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{\omega \in \Omega} N_n(\omega) \cdot X(\omega)}{n} .$$

Example 4.3.3 (A game of chance)**.** To "meditate the sacred geometry of chance",[10] two curious friends Ismael and Leyla are playing the following silly game. They roll two fair dice.

- If the two dice show the same number, Ismael gives 6000 LL to Leyla.
- If not, Leyla gives 1000 LL to Ismael.

Q Is this game fair?

A Let G denote Leyla's gain (in Lebanese Liras) in one game. This is a random variable with[11]

$$G \coloneqq \begin{cases} +6000 & \text{with probability } 1/6, \\ -1000 & \text{with probability } 5/6. \end{cases}$$

The expected gain of Leyla in one game is therefore

$$\mathbb{E}[G] = \mathbb{P}(G = +6000) \cdot (+6000) + \mathbb{P}(G = -1000) \cdot (-1000) = \frac{1}{6} \cdot 6000 + \frac{5}{6} \cdot (-1000) = \frac{1000}{6} .$$

This means that if Ismael and Leyla play this game many many times, then Leyla's average gain per game will be roughly $1000/6$ LL. Therefore, the game is biased in Leyla's favor.

Ismael and Leyla decide to modify the game to make it fair. In the modified game, the 6000 LL which Ismael pays to Leyla is changed to a different amount which we denote by x. Thus,

- If the two dice show the same number, Ismael gives x LL to Leyla.
- If not, Leyla gives 1000 LL to Ismael.

Q For which value of x is this game fair?

A In the modified game, Leyla's expected gain is

$$\mathbb{E}[G] = \mathbb{P}(G = x) \cdot x + \mathbb{P}(G = -1000) \cdot (-1000) = \frac{1}{6} \cdot x + \frac{5}{6} \cdot (-1000) .$$

In order for the game to be fair, Leyla's expected gain must be zero, hence

$$\frac{1}{6} \cdot x + \frac{5}{6} \cdot (-1000) = 0 .$$

Solving for x, we obtain $x = 5000$. ○

Example 4.3.4 (Variant of China's one-child policy[12])**.** In order to limit the population growth in China, during 1979–2016 the Chinese government had implemented the so-called *one-child policy*, limiting the number of children a couple could have to one.[13] An alternative policy which was considered was the *one-son policy*:

- *One-son policy*: As long as a couple have only female children, they are allowed to have more children.

There was however a concern that such a policy could affect the ratio of male to female in the population. Indeed, following this policy, no family would have more than one son, whereas many families would have several daughters.

[10] As Sting put it . . .
[11] A negative gain corresponds to a loss. Ismael's gain in one game is simply $-G$.
[12] This example is taken from the book *Elementary Probability for Applications* by Rick Durrett.
[13] In practice, the policy allowed for various exceptions and adjustments. See One-child policy (Wikipedia) and the references therein for more information.

Ⓠ Would the implementation of the one-son policy affect the ratio of male to female in the country?

To answer this question, we make a probability model. To avoid getting lost in details, we make the following simplifying assumptions:

I. In each birth, the chance of delivering male and female children are the same.[14]

II. The sex of successive children are independent.

III. There are no intersex children.[15]

IV. There are no twins, triplets, and so on.[16]

V. Each family will keep on having children until they have a son.

Note that under these assumptions, each family will have exactly one son.

Ⓠ What is the average number of daughters per family under the one-son policy with the above assumptions.

Ⓐ The probability model for the births in one family can be given by

- (sample space) $\Omega := \{\mathtt{M}, \mathtt{FM}, \mathtt{F}^2\mathtt{M}, \mathtt{F}^3\mathtt{M}, \dots\}$,
- (measure of probabilities) $\mathbb{P}(\mathtt{F}^k\mathtt{M}) = (1/2)^k(1/2) = (1/2)^{k+1}$ for each $k = 0, 1, 2, \dots$.

Let N denote the number of daughter in this family. This is a random variable where $N(\mathtt{F}^k\mathtt{M}) = k$. The possible values of N are $0, 1, 2, \dots$, and its distribution is given by

$$\mathbb{P}(N = k) = \begin{cases} (1/2)^{k+1} & \text{if } k = 0, 1, 2, \dots, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, the expected value of N is

$$\mathbb{E}[N] = \sum_{k=0}^{\infty} \mathbb{P}(N = k) \cdot k = \sum_{k=0}^{\infty} (1/2)^{k+1} k \, .$$

Calculating the latter sum,[17] we obtain

$$\mathbb{E}[N] = 1 \, .$$

This is the expected number of daughters in one family. Since the number of families is very large, the average number of daughters per family is approximately $\mathbb{E}[N] = 1$.

The above calculation shows that, under assumptions I–V, the one-son policy would not change the male-to-female ratio in the population.

While assumptions I–IV may be considered as reasonable simplifications, assumption V appears more suspicious. Would the conclusion remain valid if assumption V does not hold? To examine this, let us consider a modified model in which each family stops having children once they have three daughters. In other words, let us replace assumption V with the following:

V′. Each family will keep on having children until they have a son, or if they have three daughters.

Note that with this modified stopping strategy, the number of sons per family is not always one anymore.

Ⓠ What are the average numbers of daughters and sons per family under the one-son policy with the modified assumption.

Ⓐ With the modified assumption, the model is changed to the following:

- (sample space) $\Omega' := \{\mathtt{M}, \mathtt{FM}, \mathtt{F}^2\mathtt{M}, \mathtt{F}^3\}$,
- (measure of probabilities)

$$\mathbb{P}(\mathtt{M}) = 1/2 \, , \qquad \mathbb{P}(\mathtt{FM}) = 1/4 \, , \qquad \mathbb{P}(\mathtt{F}^2\mathtt{M}) = 1/8 \, , \qquad \mathbb{P}(\mathtt{F}^3) = 1/8 \, .$$

[14] In reality, the male to female ratio at birth for humans is around 1.05. The bias is compensated by the fact that the infant mortality rate for boys is slightly higher than for girls. See Human sex ratio (Wikipedia) and the references therein for more information.

[15] In reality, between 0.018% to 1.7% of infants at birth do not fit into either of the two categories of male and female (depending on the definition). See Intersex (Wikipedia) and the references therein for more information.

[16] See Multiple birth (Wikipedia) for statistics on multiple births.

[17] See Interlude 4.B for a review of how to calculate such series.

Let $N_{\mathtt{F}}$ and $N_{\mathtt{M}}$ denote the number of daughters and sons in the family, respectively. Then,

$$\mathbb{E}[N_{\mathtt{F}}] = (1/2)\cdot 0 + (1/4)\cdot 1 + (1/8)\cdot 2 + (1/8)\cdot 3 = 7/8\,, \qquad \mathbb{E}[N_{\mathtt{M}}] = (1/2)\cdot 1 + (1/4)\cdot 1 + (1/8)\cdot 1 + (1/8)\cdot 0 = 7/8\,.$$

Therefore, the average numbers of daughters and sons per family will both be roughly $7/8$.

We conclude that, under the modified assumption, the policy would not affect the male-to-female ratio in the population.

In fact, under assumptions I–IV, the balance of male to female in the population cannot change, no matter what strategy the couples follow to decide when to stop having children. Indeed, imagine a couple who are about to have a baby. Ismael and Leyla make the following bet (compare with Example 4.3.3):

- If the baby is a girl, Ismael gives 1\$ to Leyla.
- If the baby is a boy, Leyla gives 1\$ to Ismael.

Let G denote Leyla's gain in the game. Then, under assumption I, we have $\mathbb{E}[G] = 0$, which means that the game is fair. Therefore, if Ismael and Leyla repeat playing this game many times, Leyla's average gain per game will be approximately zero. Leyla may want to come up with a strategy for when to stop playing the game in order to maximize her gain. However, it can be shown that in a fair game, no matter what stopping strategy Leyla uses, her expected total gain will remain zero.[18]

Surprisingly, the one-child policy seems to have lead to an increase in the male-to-female ratio in China, but for altogether different reasons. This imbalance has been associated (among other potential causes) to widespread sex-selective abortions among some traditional segments of the society who have preference for having sons. In the presence of such sex-selective abortions, assumption I is no longer valid.[19] ○

The problem of finding the expected value of a discrete random variable X amounts to the calculation of the sum $\sum_x \mathbb{P}(X = x)x$. In practice, direct calculation of the sum can be quite tedious if not challenging. An important fact which is often of great help in this regard is the linearity of expectation.

Linearity of expectation.

- If X is a (discrete) random variable and $a, b \in \mathbb{R}$ arbitrary numbers, then

$$\mathbb{E}[aX + b] = a\,\mathbb{E}[X] + b\,.$$

- If X and Y are two (discrete) random variables, then

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

To be more precise, the above statements hold *provided* that the expectations $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ exist! In general, the expected value of a (discrete) random variable need not exist. In particular, the sum $\sum_x \mathbb{P}(X = x)x$ defining the expected value might diverge. An example in which this happens will come later (Example 4.4.6). In the meanwhile, you can try to cook up an example of your own.

Q Why do the above two identities hold?

A Let Ω denote the underlying sample space. Then,

$$\mathbb{E}[aX + b] = \sum_{\omega\in\Omega} \mathbb{P}(\omega)\big(aX(\omega) + b\big) = a\overbrace{\sum_{\omega\in\Omega} \mathbb{P}(\omega)X(\omega)}^{\mathbb{E}[X]} + b\overbrace{\sum_{\omega\in\Omega} \mathbb{P}(\omega)}^{1} = a\,\mathbb{E}[X] + b\,.$$

Similarly,

$$\mathbb{E}[X + Y] = \sum_{\omega\in\Omega} \mathbb{P}(\omega)\big(X(\omega) + Y(\omega)\big) = \overbrace{\sum_{\omega\in\Omega} \mathbb{P}(\omega)X(\omega)}^{\mathbb{E}[X]} + \overbrace{\sum_{\omega\in\Omega} \mathbb{P}(\omega)Y(\omega)}^{\mathbb{E}[Y]} = \mathbb{E}[X] + \mathbb{E}[Y]\,.$$

---

[18] The mathematical theorem explaining this is known as the optional stopping theorem. Certain realistic assumptions are needed (for instance, the assumption that Leyla's initial wealth is finite). The optional stopping theorem is the starting point of the applications of probability theory in finance.

[19] See Missing women of China (Wikipedia) and the references therein for more information.

Interlude 4.B (Arithmetico-geometric series)**.** Consider the series

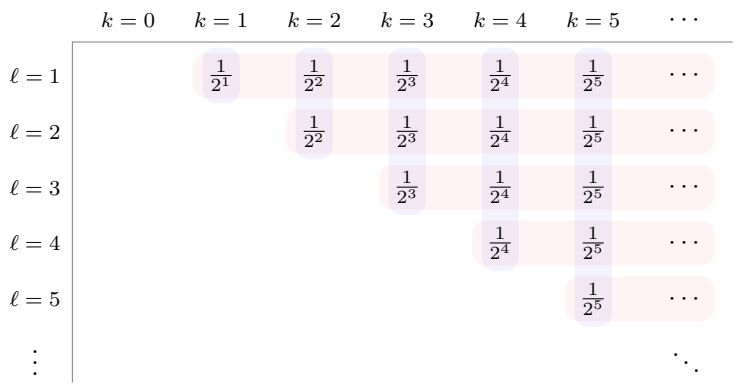
$$S = \sum_{k=0}^{\infty} \frac{k}{2^k} = \frac{0}{1} + \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{4}{16} + \cdots . \tag{♋}$$

This looks almost like a geometric series, except that the k'th term is multiplied by k.

Ⓠ How can we calculate such a series?

A1 (Method 1)

Consider the following infinite triangular array:



Observe that S is simply the sum of all the entries in this array. Indeed, the sum of the entries in the k'th column is $k/2^k$, and hence the sum of all the entries in the array is $\sum_{k=0}^{\infty} k/2^k = S$.

Now, let us add up the entries row by row. The sum of the entries in the ℓ'th row is a geometric series with starting term $1/2^\ell$ and common ratio $1/2$, and hence evaluates to $1/2^{\ell-1}$. Therefore, the sum of all the entries in the array is $\sum_{\ell=1}^{\infty} 1/2^{\ell-1}$. This is again a geometric series, this time with starting term 1 and common ratio $1/2$, and hence evaluates to 2.

We conclude that $S = 2$.

A2 (Method 2: via generating functions)

Let us define an auxiliary function $g(z)$ with one argument as

$$g(z) \coloneqq \sum_{k=0}^{\infty} z^k = \frac{1}{1-z} .$$

The series is convergent if and only if $-1 < z < 1$, so $g(z)$ is defined only for such values. Let us now take the derivative of this function:

$$g'(z) = \sum_{k=1}^{\infty} k z^{k-1} = \frac{1}{(1-z)^2} .$$

Note that the derivative can be taken in two ways: by taking the derivative of each term in the series, or by taking the derivative of the evaluated sum $1/(1-z)$. Note also that the derivative of $z^0 = 1$ is zero, and that is why we have started the series for $g'(z)$ from $k = 1$ rather than $k = 0$.

Now, observe that the series in $g'(z)$ resembles the series in S. In fact, S can be written in terms of $g'(z)$ evaluated at $z = 1/2$. Namely,

$$S = \frac{1}{2} g'(1/2) = \frac{1}{2} \cdot \frac{1}{(1-1/2)^2} = 2 .$$

The function $g(z)$ is an example of a so-called *generating function*. The method of generating functions is very powerful, but requires a degree of creativity in choosing the right function.

The series (♋) is an example of an *arithmetico-geometric series* (also known as *Gabriel's staircase*). An arithmetico-geometric series is a series of the form

$$S = \sum_{k=0}^{\infty} (a + kb) r^k = a + (a+b)r + (a+2b)r^2 + (a+3b)r^3 + (a+4b)r^4 + \cdots .$$

When $0 < r < 1$, the series converges and either of the above methods can be used to calculate it.[20] ◇

[20] See Gariel's staircase (MathWorld) for a beautiful illustration of the first method.

4.4 Interesting examples: computing the expected value

Example 4.4.1 (A binomial RV)**.** Take a coin with bias parameter p, and consider the experiment of flipping the coin n times. Let X denote the number of heads. The possible values of X are $0, 1, 2, \ldots, n$. The probability mass function of X is given by

$$p_X(k) \coloneqq \mathbb{P}(X = k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{if } k = 0, 1, 2, \ldots, n, \\ 0 & \text{otherwise,} \end{cases} \qquad (\binom{\circ}{\circ})$$

(see Example 4.2.2). The latter is called the *binomial distribution* with parameters n and p. A random variable with this distribution is called a *binomial random variable* with parameters n and p.

Ⓠ What is the expected value of X?

A1 (Method 1: via generating functions)

We have

$$\mathbb{E}[X] = \sum_{k=0}^{n} \mathbb{P}(X = k) k = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \cdot k \,.$$

How can we calculate the latter sum?

Recalling the binomial identity,[21] let us define the generating function

$$g(z) \coloneqq \sum_{k=0}^{n} \binom{n}{k} z^k = (1+z)^n \,.$$

Taking the derivative, we get

$$g'(z) = \sum_{k=1}^{n} k \cdot \binom{n}{k} z^{k-1} = n(1+z)^{n-1} \,.$$

Note that the derivative can be taken in two ways, using either of the two expressions for $g(z)$. Note also that the derivative of $z^0 = 1$ is zero, and that is why we have started the series for $g'(z)$ from $k = 1$ rather than $k = 0$.

Now, observe that the first expression for $g'(z)$ resembles the expression for $\mathbb{E}[X]$. In fact, we can write $\mathbb{E}[X]$ as

$$\begin{aligned} \mathbb{E}[X] &= (1-p)^{n-1} p \sum_{k=0}^{n} \binom{n}{k} \Big(\frac{p}{1-p}\Big)^{k-1} \cdot k \\ &= (1-p)^{n-1} p \cdot g'\Big(\frac{p}{1-p}\Big) \\ &= (1-p)^{n-1} p \cdot n\Big(1 + \frac{p}{1-p}\Big)^{n-1} \qquad \text{(using the second expression for } g'(z)\text{)} \\ &= (1-p)^{n-1} p \cdot n\Big(\frac{1}{1-p}\Big)^{n-1} \\ &= pn \,. \end{aligned}$$

A2 (Method 2: using the linearity of expectation)

Define Bernoulli random variables $X_1, X_2, \ldots, X_n$, where

$$X_i \coloneqq \begin{cases} 1 & \text{if the } i\text{'th flip comes up heads,} \\ 0 & \text{otherwise.} \end{cases}$$

These are independent Bernoulli random variables with parameter p. Observe that

$$X = X_1 + X_2 + \cdots + X_n \,.$$

Therefore, by the linearity of expectation,

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}[X_1 + X_2 + \cdots + X_n] \\ &= \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n] \\ &= \underbrace{p + p + \cdots + p}_{n \text{ times}} = np \,. \end{aligned}$$

○

[21] The binomial identity is the algebraic identity $(a+b)^n = \sum_{k=1}^{n} \binom{n}{k} a^k b^{n-k}$, which holds for every two numbers $a, b \in \mathbb{R}$ and every non-negative integer n.

Binomial random variables. A *binomial random variable* with parameters n and p is a random variable X which can be represented as a sum

$$X = X_1 + X_2 + \cdots + X_n \ ,$$

where $X_1, X_2, \ldots, X_n$ are independent Bernoulli random variables with the same parameter p. The distribution of a binomial random variable is called a *binomial distribution*.[22] Here are two generic examples of binomial random variables in different contexts:

(Bin1) Experiment: flip a coin 10 times. Let X be the number of heads. In this context, X_i indicates whether the i'th flip comes up heads or tails.

More generally, the coin flips could be replaced with any sort of random *trials* which are either successful or unsuccessful. The results of the repeated trials must be independent of one another.

(Bin2) Experiment: we have a jar with N balls, K of which are blue and the remaining $N-K$ red. At random, we draw an (ordered) sample of n balls with replacement (see Example 3.4.6 and Figure 3.11). Let X be the number of blue balls drawn. In this context, X_i indicates whether the i'th drawn ball is blue or red.

(Q) What are the binomial parameters of X?

[A] The first parameter is n, the second is $p \coloneqq K/N$.

Example 4.4.2 (A hypergeometric RV)**.** Suppose that we have a jar with N balls, K of which are blue, and the remaining $N-K$ are red. At random, we draw an (unordered) sample of n balls without replacement (see Example 3.4.7 and Figure 3.11). For simplicity, let us assume that $n \leq K$ and $n \leq N-K$, which means there are at least n balls of each color in the jar. Let X be the number of blue balls drawn. The possible values of X are $0, 1, 2, \ldots, n$. The probability mass function of X is given by

$$p_X(k) \coloneqq \mathbb{P}(X=k) = \begin{cases} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} & \text{if } k = 0,1,2,\ldots,n, \\ 0 & \text{otherwise.} \end{cases}$$

The latter is called the *hypergeometric distribution* with parameters N, K and n. A random variable with this distribution is called a *hypergeometric random variable* with parameters N, K and n.[23]

(Q) What is the expected value of X?

[A1] (Method 1: via generating functions; optional)

We have

$$\mathbb{E}[X] = \sum_{k=0}^{n} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \cdot k = \frac{1}{\binom{N}{n}} \sum_{k=0}^{n} \binom{K}{k}\binom{N-K}{n-k} \cdot k \ .$$

The latter sum can be calculated using the method of generating functions, but in a more complicated way. Let us define a function $h(u,v)$ of two real arguments as

$$h(u,v) = (1+u)^K (1+v)^{N-K}$$

Expanding the two terms using the binomial identity, we get

$$h(u,v) = \left[\sum_{k=0}^{K} \binom{K}{k} u^k\right]\left[\sum_{\ell=0}^{N-K} \binom{N-K}{\ell} v^\ell\right] = \sum_{k=0}^{K}\sum_{\ell=0}^{N-K} \binom{K}{k}\binom{N-K}{\ell} u^k v^\ell \ .$$

Taking the partial derivative of h with respect to its first argument, we get

$$h_1(u,v) \coloneqq \frac{\partial}{\partial u} h(u,v) = K(1+u)^{K-1}(1+v)^{N-K} = \sum_{k=1}^{K}\sum_{\ell=0}^{N-K} \binom{K}{k}\binom{N-K}{\ell} \cdot k u^{k-1} v^\ell \ .$$

[22] Alternatively, the binomial distribution with parameters n and p can be defined by its pmf, which is given in $\binom{\circ}{\circ}$.

[23] The assumptions $n \leq K$ and $n \leq N-K$ can be relaxed with appropriate adjustments.

Note that the first sum starts from $k = 1$ rather than $k = 0$, because the derivative of $u^0 = 1$ is zero. Note also that the second expression for $h_1(u, v)$ has a resemblance to the expression for $\mathbb{E}[X]$. Let us now substitute u for v and define

$$g(u) \coloneqq h_1(u, u) = K(1+u)^{N-1} = \sum_{k=1}^{K} \sum_{\ell=0}^{N-K} \binom{K}{k} \binom{N-K}{\ell} \cdot k u^{k+\ell-1} \,.$$

Observe that the two expressions for $g(u)$ are polynomials in variable u. In order for two polynomials in u to be equal, their corresponding coefficients for the powers of u must be equal. Identifying the coefficients of u^{n-1} in the two polynomials, we obtain

$$K\binom{N-1}{n-1} = \sum_{k=1}^{n} \binom{K}{k} \binom{N-K}{n-k} \cdot k \,.$$

Dividing by $\binom{N}{n}$, we find that

$$\mathbb{E}[X] = \frac{1}{\binom{N}{n}} \sum_{k=0}^{n} \binom{K}{k} \binom{N-K}{n-k} \cdot k = \frac{K\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{K\frac{(N-1)!}{(n-1)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = \boxed{\frac{Kn}{N}} \,.$$

A2 (Method 2: using the linearity of expectation)

Imagine drawing the balls one by one. At the end, we ignore the order of the drawn balls and obtain an unordered sample. However, this way of drawing the sample allows us represent X as

$$X = X_1 + X_2 + \cdots X_n \,,$$

where

$$X_i \coloneqq \begin{cases} 1 & \text{if the } i\text{'th drawn ball is blue,} \\ 0 & \text{otherwise.} \end{cases}$$

Note that

- Each X_i is a Bernoulli random variable.
- $X_1, X_2, \ldots, X_n$ are *not* independent.

Q What is the Bernoulli parameter of X_i?

A1 The parameter of X_1 is clearly K/N. To find the parameter of X_2, we divide the possibilities based on the value of X_1. Namely, by the principle of total probability,

$$\begin{aligned} \langle \text{parameter of } X_2 \rangle &= \mathbb{P}(X_2 = 1) \\ &= \mathbb{P}(X_1 = 0)\,\mathbb{P}(X_2 = 1 \mid X_1 = 0) + \mathbb{P}(X_1 = 1)\,\mathbb{P}(X_2 = 1 \mid X_1 = 1) \\ &= \frac{N-K}{N} \cdot \frac{K}{N-1} + \frac{K}{N} \cdot \frac{K-1}{N-1} \\ &= \frac{NK - K^2 + K^2 - K}{N(N-1)} \\ &= \frac{K}{N} \,. \end{aligned}$$

With the same approach but more tediously, we can verify that the parameter of X_3 is K/N as well. What about the rest?

A2 The parameter of X_i is K/N for every $i = 0, 1, 2, \ldots, n$. Imagine drawing the balls without looking at their colors. Once all the balls are drawn, pick the i'th ball and observe its color. Clearly, the chance that the ball is blue is K/N.

Therefore, by the linearity of expectation,

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}[X_1 + X_2 + \cdots + X_n] \\ &= \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n] \\ &= \underbrace{K/N + K/N + \cdots + K/N}_{n \text{ times}} = \boxed{n \cdot \frac{K}{N}} \,. \end{aligned}$$

Let us emphasize that, unlike in Example 4.4.1, here the random variables $X_1, X_2, \ldots, X_n$ are not independent. Nevertheless, the above computation is still valid. The linearity of expectation holds irrespective of whether the summands are independent of not. ○

Example 4.4.3 (A geometric RV)**.** Take a coin with bias parameter p. Repeat flipping the coin until it comes up heads. Let N denote the total number of flips in this experiment. The possible values of N are $1, 2, \ldots$. The probability mass function of N is given by

$$p_N(k) \coloneqq \mathbb{P}(N = k) = \begin{cases} (1-p)^{k-1}p & \text{if } k = 1, 2, 3, \ldots, \\ 0 & \text{otherwise,} \end{cases}$$

(see Example 4.2.3). This is called the *geometric distribution* with parameter p. A random variable with this distribution is called a *geometric random variable* with parameter p.

Q What is the expected value of N?

A (via generating functions)

We have

$$\mathbb{E}[N] = \sum_{k=1}^{\infty} \mathbb{P}(N = k)k = \sum_{k=1}^{\infty} (1-p)^{k-1}p \cdot k \, .$$

Note that the latter resembles an arithmetico-geometric series. As in Interlude 4.B, let us consider the generating function

$$g(z) \coloneqq \sum_{k=0}^{\infty} z^k = \frac{1}{1-z} \, ,$$

defined for $-1 < z < 1$. Taking the derivative, we have

$$g'(z) = \sum_{k=1}^{\infty} k z^{k-1} = \frac{1}{(1-z)^2} \, .$$

Now, we have

$$\mathbb{E}[N] = p\, g'(1-p) = \frac{p}{\big(1-(1-p)\big)^2} = \frac{p}{p^2} = 1/p \, .$$

○

Geometric RVs are memoryless. Let N be a geometric random variable with parameter p. Then, for every two integers $m, k \geq 1$, we have

$$\begin{aligned} \mathbb{P}(N = m + k \mid N > m) &= \frac{\mathbb{P}(N = m + k \text{ and } N > m)}{\mathbb{P}(N > m)} = \frac{\mathbb{P}(N = m + k)}{\mathbb{P}(N > m)} \\ &= \frac{(1-p)^{m+k-1}p}{(1-p)^m} = (1-p)^{k-1}p \\ &= \mathbb{P}(N = k) \, . \end{aligned}$$

In other words, the conditional distribution of $N - m$ given $N > m$ is the same as the distribution of N. This phenomenon is referred to as the *memorylessness* of the geometric random variables, and has a simple explanation based on the coin flipping experiment. Namely, thinking of N as the number of flips till the first head comes up, the memorylessness of N simply says the following:

- If we learn that in the first m flips no head has come up, then the number flips till the first head counted from after the m'th flip is again geometrically distributed with the same parameter.

But this is rather obvious, given that the coin flips are independent.

Example 4.4.4 (A negative binomial RV)**.** Take a coin with bias parameter p. Let r be a positive integer. We repeat flipping the coin until we get r heads (see Example 3.4.8). Let X denote the number of tails. The possible values of X are $0, 1, 2, \ldots$. The probability mass function of X is given by

$$p_X(k) \coloneqq \mathbb{P}(X = k) = \begin{cases} \binom{k+r-1}{r-1} p^r (1-p)^k & \text{if } k = 0, 1, 2, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

The latter is called the *negative binomial distribution* with parameters r and p. A random variable with this distribution is called a *negative binomial random variable* with parameters r and p.

Observe that this is *essentially* a generalization of the previous example. In particular, if $r = 1$, then $X + 1$ is a geometric random variable.

Q What is the expected value of X?

A1 (Method 1: via generating functions; optional)

This is left to you as an exercise. You can use the negative binomial identity[24]

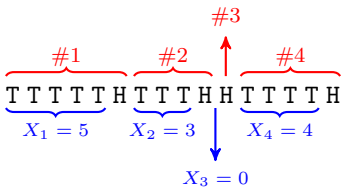
$$\frac{1}{(1-z)^r} = \sum_{k=0}^{\infty} \binom{k+r-1}{r-1} z^k \,.$$

A2 (Method 2: using the linearity of expectation)

Imagine simulating this experiment with the help of r students. Each student has a coin with the same parameter.

- Student $\#1$: Repeats flipping his/her coin until it comes of heads. Let X_1 denote the number of tails he/she gets.
- Student $\#2$: Repeats flipping his/her coin until it comes of heads. Let X_2 denote the number of tails he/she gets.
- ...
- Student $\#r$: Repeats flipping his/her coin until it comes of heads. Let X_r denote the number of tails he/she gets.

We simulate the original experiment by first reading the outcome of the coin flips made by the first student, then reading the outcomes of coin flips made by the second students, and so on and so forth. Here is an example, in the case $r = 4$:



Clearly,

$$X = X_1 + X_2 + \cdots + X_r \,.$$

Furthermore,

- $X_1, X_2, \ldots, X_r$ are independent,
- For each i, $X_i + 1$ is a geometric random variable with parameter p. In particular, $\mathbb{E}[X_i + 1] = 1/p$, which implies $\mathbb{E}[X_i] = 1/p - 1$.

Therefore, by the linearity of expectation,

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[X_1 + X_2 + \cdots + X_r] \\ &= \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_r] \\ &= \underbrace{(1/p - 1) + (1/p - 1) + \cdots + (1/p - 1)}_{r \text{ times}} = r(1/p - 1) = \frac{r(1-p)}{p} \,.\end{aligned}$$

○

Example 4.4.5 (A Poisson RV)**.** A *Poisson random variable*[25] with parameter $\mu > 0$ is a random variable X with possible values $0, 1, 2, \ldots$ and probability mass function

$$p_X(k) \coloneqq \mathbb{P}(X = k) = \begin{cases} \mathrm{e}^{-\mu} \dfrac{\mu^k}{k!} & \text{if } k = 0, 1, 2, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

The latter is called the *Poisson distribution* with parameter p. We will later discuss the interpretation of such random variables. For now, note that p_X is non-negative and

$$\sum_{k=0}^{\infty} p_X(k) = \sum_{k=0}^{\infty} \mathrm{e}^{-\mu} \frac{\mu^k}{k!} = \mathrm{e}^{-\mu} \sum_{k=0}^{\infty} \frac{\mu^k}{k!} = \mathrm{e}^{-\mu} \mathrm{e}^{\mu} = 1 \,,$$

so p_X is indeed a valid probability mass function.

[24] See Newton's generalized binomial theorem.
[25] Named after mathematician Siméon Denis Poisson (1781–1840).

Q What is the expected value of X?

A $\mathbb{E}[X] = \mu$. The derivation is left as an exercise. ○

Example 4.4.6 (Expected value need not exist!)**.** Let X be a geometric random variable with parameter p, that is,

$$\mathbb{P}(X = k) = \begin{cases} (1-p)^{k-1}p & \text{if } k = 1, 2, 3, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

Let r be a real number and consider a new random variable $Y \coloneqq r^X$. If we try to calculate the expected value of Y, we get

$$\mathbb{E}[Y] = \sum_{k=1}^{\infty} \mathbb{P}(X = k) r^k = \sum_{k=1}^{\infty} (1-p)^{k-1} p\, r^k = \frac{p}{1-p} \sum_{k=1}^{\infty} \big(r(1-p)\big)^k .$$

However, the latter sum (a geometric series) diverges if $|r(1-p)| \geq 1$. For instance, if $r = 1/(1-p)$, then the sum becomes $\sum_{k=1}^{\infty} 1$, which diverges to $+\infty$, and if $r = -1/(1-p)$, the sum turns to $\sum_{k=1}^{\infty} (-1)^k$ which is an oscillating divergent series. ○

4.5 Variance and standard deviation of a random variable

The expected value is an indication of the "center" of a distribution. The variance and standard deviation measure its "spread".

Terminology. The *variance* of a (discrete) random variable X with $\mathbb{E}[X] = \mu$ is defined as

$$\mathbb{V}\mathrm{ar}[X] \coloneqq \mathbb{E}[(X-\mu)^2] .$$

The *standard deviation* of X is the square root of its variance, that is,

$$\mathbb{SD}[X] \coloneqq \sqrt{\mathbb{V}\mathrm{ar}[X]} .$$

The standard deviation of a random variable X is often denoted by the Greek letter σ, or by σ_X if we want to emphasize its dependence on X. Thus, the variance of X is denoted by σ^2 or σ_X^2.

Interpretation. The variance of a random variable X is the abstract version of the variance of the values in a data set (Section 2.3). More specifically, suppose that we repeat the experiment in which X is defined n times, and let x_k denote the value of X in the k'th experiment. Then,

$$\mathbb{V}\mathrm{ar}[X] \approx \mathsf{var}(x_1, x_2, \ldots, x_n) = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2$$

when n is large, and the approximation becomes more and more accurate as $n \to \infty$.

Alternative expression. Expanding $(X-\mu)^2$ and using the linearity of expectation, we can write

$$\mathbb{V}\mathrm{ar}[X] = \mathbb{E}[X^2 - 2X\mu + \mu^2] = \mathbb{E}[X^2] - 2\mu \underbrace{\mathbb{E}[X]}_{\mu} + \mu^2 = \mathbb{E}[X^2] - \mu^2 .$$

Therefore, we get the following alternative expression:

$$\mathbb{V}\mathrm{ar}[X] = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2 .$$

Example 4.5.1 (A Bernoulli RV)**.** Let X be a Bernoulli random variable with parameter p, that is,

$$X = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1-p. \end{cases}$$

We know that $\mathbb{E}[X] = (1-p) \cdot 0 + p \cdot 1 = p$ (Example 4.3.2).

Q What is the variance of X?

A Observe that for a Bernoulli random variable, we always have $X^2 = X$, because $0^2 = 0$ and $1^2 = 1$. Therefore, using the second expression for the variance, we have

$$\mathbb{V}\mathrm{ar}[X] = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2 = p - p^2 = p(1-p) .$$

○

Before, we continue with the properties of variance, let us mention a fact about expectation.

Expectation of product of independent RVs. If X and Y are *independent* (discrete) random variables, then

$$\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

Ⓠ Why?

A

$$\begin{aligned}
\mathbb{E}[X \cdot Y] &= \sum_{x,y} \mathbb{P}(X = x, Y = y)xy \\
&= \sum_{x,y} \mathbb{P}(X = x)\,\mathbb{P}(Y = y)xy && \text{(by independence)} \\
&= \sum_{x} \sum_{y} \big[\,\mathbb{P}(X = x)x\big] \cdot \big[\,\mathbb{P}(Y = y)y\big] \\
&= \Big[\sum_{x} \mathbb{P}(X = x)x\Big] \cdot \Big[\sum_{y} \mathbb{P}(Y = y)y\Big] \\
&= \mathbb{E}[X] \cdot \mathbb{E}[Y]\,.
\end{aligned}$$

As an example, if X and Y are independent Bernoulli random variables with parameters p and q, respectively, then

$$\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y] = pq$$

which can alternatively be seen by noting that $X \cdot Y$ is itself a Bernoulli random variable with parameter pq.

Some basic facts about the variance. Unlike expectation, variance is *not* linear. Nevertheless, we have the following properties, which can often help us when calculating the variance:

- If X is a (discrete) random variable and $a, b \in \mathbb{R}$ arbitrary numbers, then

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[aX + b] &= a^2\,\mathbb{V}\mathrm{ar}[X]\,, \\
\mathbb{SD}[aX + b] &= |a|\,\mathbb{SD}[X]\,.
\end{aligned}$$

  Ⓠ Why?

  A Let $\mu \coloneqq \mathbb{E}[X]$. Then, $\mathbb{E}[aX + b] = a\mu + b$, hence

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[aX + b] &= \mathbb{E}[(aX + b - a\mu - b)^2] \\
&= \mathbb{E}[a^2(X - \mu)^2] \\
&= a^2\,\mathbb{E}[(X - \mu)^2] && \text{(by linearity of expectation)} \\
&= a^2\,\mathbb{V}\mathrm{ar}[X]\,.
\end{aligned}$$

  For the second identity, note that $\sqrt{a^2} = |a|$.

  Interpretation:

  - Multiplication by a corresponds to *scaling* by $|a|$. which in turn results in the multiplication of the spread (measured by the standard deviation) by $|a|$.
  - Addition by b corresponds to *shifting* by b. Shifting does not change the spread.

  See Figure 4.4 for an example.

- If X and Y are two (discrete) independent random variables, then

$$\mathbb{V}\mathrm{ar}[X + Y] = \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Y]\,.$$

  Ⓠ Why?

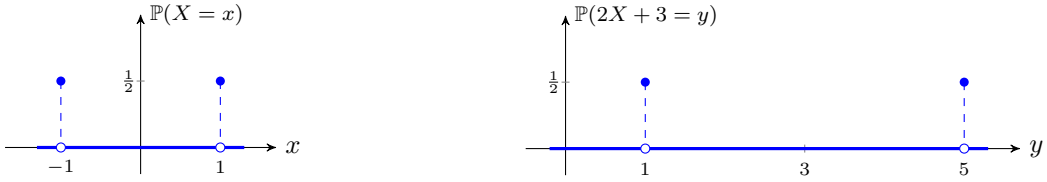
  A Let $\mu_X \coloneqq \mathbb{E}[X]$ and $\mu_Y \coloneqq \mathbb{E}[Y]$. Then,

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[X + Y] &= \mathbb{E}\left[(X + Y - \mu_X - \mu_Y)^2\right] \\
&= \mathbb{E}\left[(X - \mu_X)^2 + (Y - \mu_Y)^2 + 2(X - \mu_X)(Y - \mu_Y)\right] \\
&= \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Y] + 2\,\mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right] && \text{(by linearity of expectation)} \\
&= \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Y] + 2\underbrace{\mathbb{E}[X - \mu_X]}_{0}\underbrace{\mathbb{E}[Y - \mu_Y]}_{0} && \text{(by independence)} \\
&= \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Y]\,.
\end{aligned}$$

Note that if X and Y are not independent, then the above identity need not be satisfied. For instance, if $X = Y$ (hence, far from being independent), then

$$\mathbb{V}\mathrm{ar}[X + Y] = \mathbb{V}\mathrm{ar}[2X] = 4\,\mathbb{V}\mathrm{ar}[X] \neq 2\,\mathbb{V}\mathrm{ar}[X] = \mathbb{V}\mathrm{ar}[X] + \mathbb{V}\mathrm{ar}[Y]\,,$$

unless X is a constant.



pmf of X pmf of $2X + 3$

Figure 4.4: The distribution of a random variable X and its shifted and scaled version. Note that the distribution of $2X + 3$ is twice as spread as the distribution of X.

4.6 Interesting examples: computing the variance

Example 4.6.1 (A binomial RV)**.** Let X be a binomial random variable with parameters n and p. Thus, X has the form

$$X \coloneqq X_1 + X_2 + \cdots X_n$$

where $X_1, X_2, \ldots, X_n$ are independent Bernoulli random variables, all with the same parameter p. From Example 4.4.1, we know that $\mathbb{E}[X] = np$.

Q What is the variance of X?

A Recall that $\mathbb{V}\mathrm{ar}[X_i] = p(1-p)$ for each $i = 1, 2, \ldots, n$ (Example 4.5.1). Since $X_1, X_2, \ldots, X_n$ are independent, we have

$$\begin{aligned}\mathbb{V}\mathrm{ar}[X] &= \mathbb{V}\mathrm{ar}[X_1 + X_2 + \cdots + X_n] \\ &= \mathbb{V}\mathrm{ar}[X_1] + \mathbb{V}\mathrm{ar}[X_2] + \cdots + \mathbb{V}\mathrm{ar}[X_n] \\ &= \underbrace{p(1-p) + p(1-p) + \cdots + p(1-p)}_{n \text{ times}} = np(1-p)\,.\end{aligned}$$

○

Example 4.6.2 (A geometric RV)**.** Let X be a geometric random variable with parameter p, so that

$$\mathbb{P}(N = k) = \begin{cases}(1-p)^{k-1}p & \text{if } k = 1, 2, 3, \ldots, \\ 0 & \text{otherwise,}\end{cases}$$

From Example 4.4.3, we know that $\mathbb{E}[X] = 1/p$.

Q What is the variance of X?

A $\mathbb{V}\mathrm{ar}[X] = (1-p)/p^2$. The derivation is left as an exercise. You already know $\mathbb{E}[X]$, so it remains to find $\mathbb{E}[X^2]$. You can use the same generating function as in Example 4.4.3, but this time you need to differentiate twice. ○

Example 4.6.3 (A hypergeometric RV)**.** As in Example 4.4.2, suppose that we have a jar with N balls, K of which are blue, and the remaining $N - K$ are red. At random, we draw an (unordered) sample of n balls without replacement.

Let X be the number of blue balls drawn. This is a hypergeometric random variable with parameters N, K and n. We assume that $n \leq K$ and $n \leq N - K$.

From Example 4.4.2, we know that $\mathbb{E}[X] = nK/N$.

Q What is the variance of X?

Recall from Example 4.4.2 that X can be represented as

$$X = X_1 + X_2 + \cdots X_n \ ,$$

where

$$X_i \coloneqq \begin{cases} 1 & \text{if the } i\text{'th drawn ball is blue,} \\ 0 & \text{otherwise.} \end{cases}$$

However, note that $X_1, X_2, \ldots, X_n$ are *not* independent, whence $\mathbb{V}\mathrm{ar}[X]$ is *not* necessarily the sum of $\mathbb{V}\mathrm{ar}[X_i]$.

A

$$\mathbb{V}\mathrm{ar}[X] = \left(\frac{N-n}{N-1}\right) \cdot n \cdot \frac{K}{N}\left(1 - \frac{K}{N}\right) \ .$$

The derivation is still based on the above representation, but requires more careful calculation. We skip the derivation for now, and will return to it later when we discuss the covariance of random variables.

The above answer has the form

$$\mathbb{V}\mathrm{ar}[X] = \underbrace{\left(\frac{N-n}{N-1}\right)}_{\boxdot} \cdot \underbrace{n \cdot \frac{K}{N}\left(1 - \frac{K}{N}\right)}_{\boxplus} \ .$$

where $\boxplus$ is the variance if the sample was taken *with* replacement (see Example 4.6.1), and $\boxdot$ can be thought of as a "correction" factor needed when the sample is taken *without* replacement.

When $N, K, N-K \gg n$, the distinction between a sample with or without replacement is minute.[26] This is reflected in the above expression for $\mathbb{V}\mathrm{ar}[X]$: when $N \gg n$, we have $(N-n)/(N-1) \approx 1$, and the variance is approximately the same as in the case in which the sample is taken with replacement. ○

Example 4.6.4 (A negative binomial RV)**.** We have a coin with bias parameter p. We repeat flipping the coin until we get r heads. Let X denote the number of tails. This is a negative binomial random variable with parameters r and p.

From Example 4.4.4, we know that $\mathbb{E}[X] = r(1-p)/p$.

Q What is the variance of X?

A Recall from Example 4.4.4, that we could represent X as

$$X = X_1 + X_2 + \cdots + X_r \ ,$$

where,

- $X_1, X_2, \ldots, X_r$ are independent,
- For each i, $X_i + 1$ is a geometric random variable with parameter p. In particular, $\mathbb{V}\mathrm{ar}[X_i] = \mathbb{V}\mathrm{ar}[X_i + 1] = (1-p)/p^2$.

Therefore,

$$\begin{aligned} \mathbb{V}\mathrm{ar}[X] &= \mathbb{V}\mathrm{ar}[X_1 + X_2 + \cdots + X_n] \\ &= \mathbb{V}\mathrm{ar}[X_1] + \mathbb{V}\mathrm{ar}[X_2] + \cdots + \mathbb{V}\mathrm{ar}[X_n] && \text{(by independence)} \\ &= \underbrace{\frac{(1-p)}{p^2} + \frac{(1-p)}{p^2} + \cdots + \frac{(1-p)}{p^2}}_{n \text{ times}} = \frac{n(1-p)}{p^2} \ . \end{aligned}$$

○

Example 4.6.5 (A Poisson RV)**.** Let X be a Poisson random variable with parameter $\mu > 0$, that is,

$$\mathbb{P}(X = k) = \begin{cases} \mathrm{e}^{-\mu}\dfrac{\mu^k}{k!} & \text{if } k = 0, 1, 2, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

From Example 4.4.5, we know that $\mathbb{E}[X] = \mu$.

Q What is the variance of X?

A $\mathbb{V}\mathrm{ar}[X] = \mu$. The derivation is left as an exercise. ○

[26] Indeed, in this case, even if the sample is taken with replacement, the chance that there are repetitions will be small.