MATH 170A/B: Introduction to Probability Theory
Lecture Note

Hanbaek Lyu

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF CALIFORNIA, LOS ANGELES, CA 90095

Email address: hlyu@math.ucla.edu

WWW.HANAEKLYU.COM

Contents

Chapter 1. Sample space and Probability measures 3
1. Basic set theory 3
2. Probability measure and probability space 5
3. Conditional probability 9
4. Partitioning the sample space and Bayes' theorem 11
5. Independence 14

Chapter 2. Random variables 17
1. Discrete random variables 17
2. Expectation and variance of sums of RVs 19
3. Binomial, geometric, and Poisson RVs 21
4. Continuous Random Variables 23
5. Uniform, exponential, and normal RVs 24

Chapter 3. Joint distributions and conditioning 28
1. Joint probability mass functions 28
2. Joint probability density functions 30
3. Conditional expectation 32
4. Conditional expectation as an estimator 34
5. Conditional variance 35
6. Bayesian inference 37

Chapter 4. Random variable as a function of another random variable 41
1. Functions of one or two RVs 41
2. Sums of independent RVs – Convolution 44
3. Covariance and Correlation 47
4. Variance of sum of RVs 49

Chapter 5. Transforms of RVs 53
1. Moment generating function 53
2. Two important theorems about MGFs 55
3. MGF of sum of independent RVs 56
4. Sum of random number of independent RVs 57

Chapter 6. Elementary limit theorems 59
1. Overview of limit theorems 59
2. Bounding tail probabilities 60
3. The WLLN and convergence in probability 63
4. Central limit theorem 68
5. The SLLN and almost sure convergence 71

Chapter 7. Elementary Stochastic Processes 75
1. The Bernoulli process 75
2. The Poisson process 79
3. Discrete-time Markov chains 85

CHAPTER 1

Sample space and Probability measures

Many things in life are uncertain. Can we 'measure' and compare such uncertainty so that it helps us to make more informed decision? Probability theory provides a systematic way of doing so.

1. Basic set theory

The basic language of probability theory is provided by a branch of mathematics called *set theory*. Even though it has a lot of fascinating stories in it, we will only be needing the most basic concepts.

A *set* is a collection of abstract elements. If a set Ω contains n elements $x_1, x_2, \cdots, x_n$, then we write

$$\Omega = \{x_1, x_2, \cdots, x_n\}. \tag{1}$$

For each $i = 1, 2, \cdots, n$, we write $x_i \in \Omega$, meaning that x_i is an element of Ω. A subcollection A of the elements of Ω is called a *subset* of Ω, and we write $A \subseteq \Omega$ or $\Omega \supseteq A$. If x_i is an element of A we write $x_i \in A$; otherwise, $x_i \notin A$. When we describe A, either we list all of its elements as in (1), or we use the following conditional statement

$$A = \{x \in \Omega \mid x \text{ has property } P\}. \tag{2}$$

For instance, if $\Omega = \{1,2,3,4,5,6\}$ and $A = \{2,4,6\}$, then we can also write

$$A = \{x \in \Omega \mid x \text{ is even}\}. \tag{3}$$

Ω is a subset of itself. A subset of Ω containing no element is called the *empty set*, and is denoted by $\emptyset$.

Let A, B be two subsets of Ω. Define their *union* $A \cup B$ and *intersection* $A \cap B$ by

$$A \cup B = \{x \in \Omega \mid x \in A \text{ or } x \in B\} \tag{4}$$

$$A \cap B = \{x \in \Omega \mid x \in A \text{ and } x \in B\}. \tag{5}$$

We write $A = B$ if they consists of the same elements.

Exercise 1.1. Let Ω be a set and let $A, B \subseteq \Omega$. Show that $A = B$ if and only if $A \subseteq B$ and $B \subseteq A$.

Exercise 1.2. Let Ω be a set and let $A \subseteq \Omega$. Define the *complement* of A, denoted by A^c, by

$$A^c = \{x \in \Omega \mid x \notin A\}. \tag{6}$$

Show that $A \cup A^c = \Omega$ and $A \cap A^c = \emptyset$.

Exercise 1.3. Let Ω be a set and let $A, B \subseteq \Omega$. Define a subset $A \setminus B$ of Ω by

$$A \setminus B = \{x \in \Omega \mid x \in A \text{ and } x \notin B\}. \tag{7}$$

Show that $A \setminus B = A \cap B^c$.

Union and intersection can be defined among more than two subsets of Ω. Let $A_1, A_2, \cdots, A_k$ be subsets of Ω. Define the union and intersection of A_i's by

$$\bigcup_{i=1}^{k} A_i = \Big\{x \in \Omega \mid x \in A_i \text{ for some } i \in \{1,2,\cdots,k\}\Big\}, \tag{8}$$

$$\bigcap_{i=1}^{k} A_i = \Big\{x \in \Omega \mid x \in A_i \text{ for all } i \in \{1,2,\cdots,k\}\Big\}. \tag{9}$$

Exercise 1.4 (de Morgan's law)**.** Let Ω be a set and let $A_1, \cdots, A_k \subseteq \Omega$. Show that

$$\left(\bigcup_{i=1}^{k} A_i\right) = \bigcap_{i=1}^{k} A_i^c. \tag{10}$$

Exercise 1.5. Let Ω be a set and let $B, A_1, A_2 \cdots \subseteq \Omega$. Show that

$$B \cap \left(\bigcup_{i=1}^{\infty} A_i\right) = \bigcup_{i=1}^{\infty} B \cap A_i, \tag{11}$$

$$B \cup \left(\bigcap_{i=1}^{\infty} A_i\right) = \bigcap_{i=1}^{\infty} B \cup A_i. \tag{12}$$

Given two sets A and B, we can form their *cartesian product* $A \times B$ by the set of all pairs of elements in A and B. That is,

$$A \times B = \{(a, b) \mid a \in A,\, b \in B\}. \tag{13}$$

We write $A^2 = A \times A$, $A^3 = A \times A \times A$, and so on.

Lastly, all of the above discussion can be extended when the grounding set Ω consists of infinitely many elements. For instance, the set of all integers $\mathbb{Z}$, the set of all natural numbers $\mathbb{N}$, and the set of all real numbers $\mathbb{R}$. A set Ω containing infinitely many elements is said to be *countably infinite* if there is a one-to-one correspondence between Ω and $\mathbb{N}$; it is said to be *uncountably infinite* otherwise. One of the most well-known example of uncountably infinite set is $\mathbb{R}$, which is proved by the celebrated Cantor's diagonalization argument.

Exercise 1.6 (Binary expansion)**.** Let $[0,1] \subseteq \mathbb{R}$ be the unit interval, which is also called the *continuum*.

(i) Given an element $x \in [0,1]$, show that there exists a unique binary expansion of x. That is, there exists a sequence of integers $x_1, x_2, \cdots$ from $\{0,1\}$ such that

$$x = 0.x_1x_2x_3\cdots \tag{14}$$

$$:= \frac{x_1}{2} + \frac{x_2}{2^2} + \frac{x_3}{2^3} + \cdots. \tag{15}$$

(Hint: Divide $[0,1]$ into $[0,1/2) \cup [1/2,1]$. Then $x_1 = 0$ if x belongs to the first half, and $x_1 = 1$ otherwise. Subdivide the interval and see which half it belongs to, and so on.)

(ii) Given a binary expansion $0.x_1x_2x_3\cdots$, define a sequence of real numbers $y_1, y_2, \cdots$ by

$$y_n = 0.x_1x_2\cdots x_n \tag{16}$$

$$= \frac{x_1}{2} + \frac{x_2}{2^2} + \cdots + \frac{x_n}{2^n}. \tag{17}$$

Show that the sequence $(y_n)_{n\geq 1}$ is non-decreasing and bounded above by 1. Conclude that there exists a limit $y := \lim_{n\to\infty} y_n$.

Remark 1.7. Binary expansion in fact gives a one-two-one correspondence between the unit interval $[0,1]$ and the infinite product $\{0,1\}^{\mathbb{N}} = \{0,1\} \times \{0,1\} \times \cdots$ of 0's and 1's.

Exercise 1.8 (Cantor's diagonalization argument)**.** Let $[0,1] \subseteq \mathbb{R}$ be the unit interval.

(i) Suppose $[0,1]$ is countably infinite. Then we can enumerate all of its elements by $a_1, a_2, a_3, \cdots$. Using (i), we can write each a_i's by its unique binary expansion:

$$a_1 = 0.a_{11}a_{12}a_{13}a_{14}\cdots \tag{18}$$

$$a_2 = 0.a_{21}a_{22}a_{23}a_{24}\cdots \tag{19}$$

$$a_3 = 0.a_{31}a_{32}a_{33}a_{34}\cdots \tag{20}$$

$$a_4 = 0.a_{41}a_{42}a_{43}a_{44}\cdots \tag{21}$$

$$\vdots \tag{22}$$

Now let $\alpha \in [0,1]$ be such that

$$\alpha = 0.\bar{a}_{11}\bar{a}_{22}\bar{a}_{33}\bar{a}_{44}\cdots, \tag{23}$$

where $\bar{0} = 1$ and $\bar{1} = 0$. Show that $\alpha \neq a_i$ for all $i \geq 1$. Hence we have found an element of $[0,1]$ that is not among the list $a_1, a_2, a_3, \cdots$.

(ii) Conclude that $[0,1]$ is uncountably infinite. Since $[0,1] \subseteq \mathbb{R}$, it follows that $\mathbb{R}$ is also uncountably infinite.

2. Probability measure and probability space

We begin with idealizing our situation. Let Ω be a finite set, called *sample space*. This is the collection of all possible outcomes that we can observe (think of six sides of a die). We are going to perform some experiment on Ω, and the outcome could be any subset E of Ω, which we call an *event*. Let us denote the collection of all events $E \subseteq \Omega$ by 2^Ω. A *probability measure* on Ω is a function $\mathbb{P} : 2^\Omega \to [0,1]$ such that for each event $E \subseteq \Omega$, it assigns a number $\mathbb{P}(E) \in [0,1]$ and satisfies the following properties:

(i) $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$.

(ii) If two events $E_1, E_2 \subseteq \Omega$ are disjoint, then $\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_1)$.

In words, $\mathbb{P}(E)$ is our quantization of how likely it is that the event E occurs out of our experiment.

Remark 2.1. For general sample space Ω (not necessarily finite), not every subset of Ω can be an event. Precise definition of the collection of 'events' for the general case is beyond the scope of this course. On the other hand, the axiom (ii) for the probability measure needs to be replaced with the following countable version:

(ii)' For a countable collection of disjoint events $A_1, A_2, \cdots \subseteq \Omega$,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i). \tag{24}$$

Exercise 2.2. Let $\mathbb{P}$ be a probability measure on sample space Ω. Show the following.

(i) Let $E = \{x_1, x_2, \cdots x_k\} \subseteq \Omega$ be an event. Then $\mathbb{P}(E) = \sum_{i=1}^{k} \mathbb{P}(\{x_i\})$.

(ii) $\sum_{x\in\Omega} \mathbb{P}(\{x\}) = 1$.

If $\mathbb{P}$ is a probability measure on sample space Ω, we call the pair $(\Omega, \mathbb{P})$ a *probability space*. This is our idealized world where we can precisely measure uncertainty of all possible events. Of course, there could be many (in fact, infinitely many) different probability measures on the same sample space.

Exercise 2.3 (coin flip)**.** Let $\Omega = \{H, T\}$ be a sample space. Fix a parameter $p \in [0,1]$, and define a function $\mathbb{P}_p : 2^\Omega \to [0,1]$ by $\mathbb{P}_p(\emptyset) = 0$, $\mathbb{P}_p(\{H\}) = p$, $\mathbb{P}_p(\{T\}) = 1-p$, $\mathbb{P}_p(\{H,T\}) = 1$. Verify that $\mathbb{P}_p$ is a probability measure on Ω for each value of p.

A typical way of constructing a probability measure is to specify how likely it is to see each individual element in Ω. Namely, let $f : \Omega \to [0,1]$ be a function that sums up to 1, i.e., $\sum_{x\in\Omega} f(x) = 1$. Define a function $\mathbb{P} : 2^\Omega \to [0,1]$ by

$$\mathbb{P}(E) = \sum_{\omega\in E} f(\omega). \tag{25}$$

Then this is a probability measure on Ω, and f is called a *probability distribution* on Ω. For instance, the probability distribution on $\{H,T\}$ we used to define $\mathbb{P}_p$ in Exercise 2.3 is $f(H) = p$ and $f(T) = 1-p$.

Example 2.4 (Uniform probability measure)**.** Let $\Omega = \{1, 2, \cdots, m\}$ be a sample space and let $\mathbb{P}$ be the *uniform probability measure* on Ω, that is,

$$\mathbb{P}(\{x\}) = 1/m \qquad \forall x \in \Omega. \tag{26}$$

Then for the event $A = \{1,2,3\}$, we have

$$\mathbb{P}(A) = \mathbb{P}(\{1\} \cup \{2\} \cup \{3\}) \tag{27}$$

$$= \mathbb{P}(\{1\}) + \mathbb{P}(\{2\}) + \mathbb{P}(\{3\}) \tag{28}$$

$$= \frac{1}{m} + \frac{1}{m} + \frac{1}{m} = \frac{3}{m} \tag{29}$$

Likewise, if $A \subseteq \Omega$ is any event and if we let $|A|$ denote the size (number of elements) of A, then

$$\mathbb{P}(A) = \frac{|A|}{m}. \tag{30}$$

For example, let $\Omega = \{1,2,3,4,5,6\}^2$ be the sample space of a roll of two fair dice. Let A be the event that the sum of two dice is 5. Then

$$A = \{(1,4),(2,3),(3,2),(4,1)\}, \tag{31}$$

so $|A| = 4$. Hence $\mathbb{P}(A) = 4/36 = 1/9$. ▲

Exercise 2.5. Show that the function $\mathbb{P} : 2^{\Omega} \to [0,1]$ defined in (25) is a probability measure on Ω. Conversely, show that every probability measure on a finite sample space Ω can be defined in this way.

Remark 2.6 (General probability space)**.** A probability space does not need to be finite, but we need a more careful definition in that case. For example, if we take Ω to be the unit interval $[0,1]$, then we have to be careful in deciding which subset $E \subseteq \Omega$ can be an 'event': not every subset of Ω can be an event. A proper definition of general probability space is out of the scope of this course.

Exercise 2.7. Let $(\Omega, \mathbb{P})$ be a probability space and let $A \subseteq \Omega$ be an event. Show that $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

Example 2.8 (Roll of two dice)**.** Suppose we roll two dice and let X and Y be the outcome of each die. Say all possible joint outcomes are equally likely. The sample space for the roll of a single die can be written as $\{1,2,3,4,5,6\}$, so the sample space for rolling two dice can be written as $\Omega = \{1,2,3,4,5,6\}^2$. In picture, think of 6 by 6 square grid and each node represents a unique outcome (x,y) of the roll. By assumption, each out come has probability $1/36$. Namely,

$$\mathbb{P}((X,Y) = (x,y)) = 1/36 \qquad \forall 1 \le x, y \le 6. \tag{32}$$

This gives the uniform probability distribution on our sample space Ω (see Example 2.4).

We can compute various probabilities for this experiment. For example,

$$\mathbb{P}(\text{at least one die is even}) = 1 - \mathbb{P}(\text{both dice are odd}) \tag{33}$$

$$= 1 - \mathbb{P}(\{1,3,5\} \times \{1,3,5\}) = 1 - \frac{9}{36} = 3/4, \tag{34}$$

where for the first equality we have used the complementary probability in Exercise 2.7.

Now think about the following question: *What is the most probable value for the sum $X+Y$?* By considering diagonal lines $x + y = k$ in the 2-dimensional plane for different values of k, we find

$$\mathbb{P}(X+Y=k) = \frac{\#\text{ of intersections between the line } x+y=k \text{ and } \Omega}{36}. \tag{35}$$

From example, $\mathbb{P}(X+Y=2) = 1/36$ and $\mathbb{P}(X+Y=7) = 6/36 = 1/6$. Moreover, from Figure 1, it is clear that the number of intersections is maximized when the diagonal line $x+y=k$ passes through the extreme points $(1,6)$ and $(6,1)$. Hence 7 is the most probable value for $X+Y$ with the probability being $1/6$.

▲

Exercise 2.9 (Roll of three dice)**.** Suppose we roll three dice and all possible joint outcomes are equally likely. Identify the sample space $\Omega = \{1,2,3,4,5,6\}^3$ as a $(6 \times 6 \times 6)$ 3-dimensional integer lattice, and let X, Y, and Z denote the outcome of each die.

(i) Write down the probability distribution on Ω.

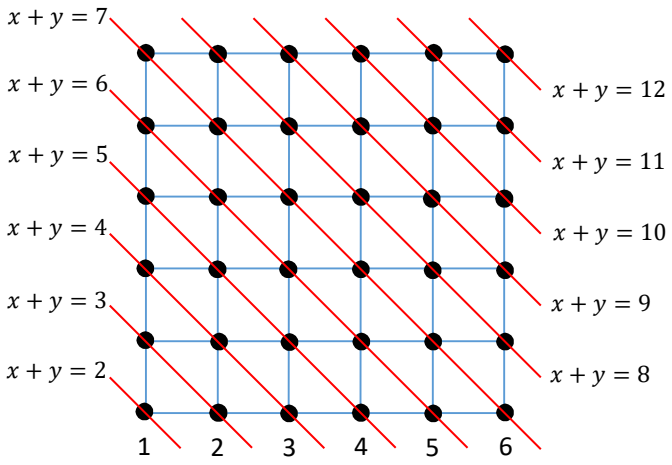

FIGURE 1. Sample space representation for roll of two independent fair dice and and events of fixed sum of two dice.

(ii) For each $k \geq 1$, show that

$$\mathbb{P}(X+Y+Z=k) = \frac{\#\text{ of intersections between the plane } x+y+z=k \text{ and } \Omega}{6^3}. \tag{36}$$

What are the minimum and maximum possible values for $X+Y+Z$?

(iii) Draw a cube for Ω and planes $x+y+z=k$ for $k=3,5,10,11,16,18$. Argue that the intersection gets larger as k increases from 3 to 10 and smaller as k goes from 11 to 18. Conclude that 10 and 11 are the most probable values for $X+Y+Z$.

(iv) Consider the following identity

$$(x+x^2+x^3+x^4+x^5+x^6)^3 \tag{37}$$

$$= x^{18}+3x^{17}+6x^{16}+10x^{15}+15x^{14}+21x^{13}+25x^{12}+27x^{11}+27x^{10} \tag{38}$$

$$+25x^9+21x^8+15x^7+10x^6+6x^5+3x^4+x^3 \tag{39}$$

Show that the coefficient of x^k in the right hand side equals the size of the intersection between Ω and the plane $x+y+z=k$. Conclude that

$$\mathbb{P}(X+Y+Z=10) = \mathbb{P}(X+Y+Z=11) = \frac{27}{6^3} = \frac{1}{8}. \tag{40}$$

(This way of calculating probabilities is called the generating function method.)

Exercise 2.10. Suppose Nate commutes to campus by Bruin bus, which arrives at his nearby bus stop every 10 min. Suppose each bus waits at the stop for 1 min. What is the probability that Nate takes no more than 3 min at the stop until he takes a bus? (Hint: Represent the sample space as a unit square in the coordinate plane)

The following are important properties of probability measure.

Theorem 2.11. *Let* $(\Omega, \mathbb{P})$ *be a probability space. These followings hold:*

(i) *(Monotonicity) For any events* $A \subseteq B$, $\mathbb{P}(A) \leq \mathbb{P}(B)$.
(ii) *(Subadditivity) For* $A \subseteq \bigcup_{i=1}^{\infty} A_i$, $\mathbb{P}(A) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.
(iii) *(Continuity from below) If* $A_1 \subseteq A_2 \subseteq \cdots$ *and* $A = \bigcup_{i=1}^{\infty} A_i$, *then* $\mathbb{P}(A) = \lim_{n\to\infty} \mathbb{P}(A_n)$.
(iv) *(Continuity from above) If* $A_1 \supseteq A_2 \supseteq \cdots$ *and* $A = \bigcap_{i=1}^{\infty} A_i$, *then* $\mathbb{P}(A) = \lim_{n\to\infty} \mathbb{P}(A_n)$.

PROOF. (i) Since $A \subseteq B$, write $B = A \cup (B \setminus A)$. Note that A and $B \setminus A$ are disjoint. Hence by the second axiom of probability measure, we get

$$\mathbb{P}(B) = \mathbb{P}(A \cup (B \setminus A)) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A), \tag{41}$$

where the last inequality uses the fact that $\mathbb{P}(B \setminus A) \geq 0$.

(ii) The events A_i's are not necessarily disjoint, but we can cook up a collection of disjoint events B_i's so that $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$. Namely, we define

$$B_1 = A_1 \subseteq A_1 \tag{42}$$

$$B_2 = (A_1 \cup A_2) \setminus A_1 \subseteq A_2 \tag{43}$$

$$B_3 = (A_1 \cup A_2 \cup A_3) \setminus (A_1 \cup A_2) \subseteq A_3, \tag{44}$$

and so on. Then clearly B_i's are disjoint and their union is the same as the union of A_i's. Now by part (i) and axiom (ii)' of probability measure (See Remark 2.1), we get

$$\mathbb{P}(A) \leq \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(B_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i). \tag{45}$$

(iii) Define a collection of disjoint subsets B_i's similarly as in (ii). In this case, they will be

$$B_1 = A_1 \subseteq A_1 \tag{46}$$

$$B_2 = A_2 \setminus A_1 \subseteq A_2 \tag{47}$$

$$B_3 = A_3 \setminus A_2 \subseteq A_3, \tag{48}$$

and so on. Then $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$ and $A_n = \bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} B_i$. Hence we get

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(B_i) \tag{49}$$

$$= \lim_{k\to\infty} \sum_{i=1}^{n} \mathbb{P}(B_i) = \lim_{n\to\infty} \mathbb{P}(B_1 \cup \cdots \cup B_n) = \lim_{n\to\infty} \mathbb{P}(A_n). \tag{50}$$

(iv) Note that $A_1^c \subseteq A_2^c \subseteq \cdots$ and $A^c = \bigcup_{i=1}^{\infty} A_i^c$ by de Morgan's law (Exercise 1.4). Hence by part (iii), we deduce

$$\mathbb{P}(A^c) = \lim_{n\to\infty} \mathbb{P}(A_n^c). \tag{51}$$

But then by Exercise 2.7, this yields

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c) = 1 - \lim_{n\to\infty} \mathbb{P}(A_n^c) = \lim_{n\to\infty} \left(1 - \mathbb{P}(A_n^c)\right) = \lim_{n\to\infty} \mathbb{P}(A_n). \tag{52}$$

□

Some immediate consequences of Theorem 2.11 (ii) are given in the following exercise.

Exercise 2.12 (Union bound)**.** Let $(\Omega, \mathbb{P})$ be a probability space.

(i) For any $A, B \subseteq \Omega$ such that $A \subseteq B$, show that

$$\mathbb{P}(A) \leq \mathbb{P}(B). \tag{53}$$

(ii) For any $A, B \subseteq \Omega$, show that

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B). \tag{54}$$

(iii) Let $A_1, A_2, \cdots, A_k \subseteq \Omega$. By an induction on k, show that

$$\mathbb{P}\left(\bigcup_{i=1}^{k} A_i\right) \leq \sum_{i=1}^{k} \mathbb{P}(A_i). \tag{55}$$

(iv) (Countable subadditivity) Let $A_1, A_2, \cdots \subseteq \Omega$ be a countable collection of events. Show that

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i). \tag{56}$$

Exercise 2.13 (Inclusion-exclusion)**.** Let $(\Omega, \mathbb{P})$ be a probability space. Let $A_1, A_2, \cdots, A_k \subseteq \Omega$. Show the following.

(i) For any $A, B \subseteq \Omega$,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \tag{57}$$

(ii) For any $A, B, C \subseteq \Omega$,

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(B \cap C) - \mathbb{P}(C \cap A) + \mathbb{P}(A \cap B \cap C). \tag{58}$$

(iii)* Let $A_1, A_2, \cdots, A_k \subseteq \Omega$. Use an induction on k to show that

$$\mathbb{P}\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i_1=1}^{k} \mathbb{P}(A_{i_1}) - \sum_{1 \leq i_1 < i_2 \leq k} \mathbb{P}(A_{i_1} \cap A_{i_2}) \tag{59}$$

$$+ \sum_{1 \leq i_1 < i_2 < i_3 \leq k} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \cdots + (-1)^k \mathbb{P}\left(\bigcap_{i=1}^{k} A_i\right). \tag{60}$$

Remark 2.14. Later, we will show the general inclusion-exclusion in a much easier way using random variables and expectation.

3. Conditional probability

Consider two experiments on a probability space and the outcomes are recorded by X and Y. For instance, X could be the number of friends on Facebook and Y could be the number of connections on LinkedIn of a randomly chosen classmate. Perhaps it would be case that Y is large if X is large. Or maybe the opposite is true. In any case, the outcome of Y is most likely be affected by knowing something about X. This leads to the notion of 'conditioning'. For any two events E_1 and E_2 such that $\mathbb{P}(E_2) > 0$, we define

$$\mathbb{P}(E_1 \,|\, E_2) = \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_2)} \tag{61}$$

and this quantity is called the *conditional probability* of E_1 given E_2. Note that $\mathbb{P}(E_1 \,|\, E_2) \leq 1$ since $E_1 \cap E_2 \subseteq E_2$ and

Example 3.1. Roll a fair die and let $X \in \{1,2,3,4,5,6\}$ be the outcome. Then

$$\mathbb{P}(X = 2) = 1/6, \tag{62}$$

$$\mathbb{P}(X = 2 \,|\, X \text{ is even}) = \frac{\mathbb{P}(X = 2)}{\mathbb{P}(X \in \{2,4,6\})} = 1/3, \tag{63}$$

$$\mathbb{P}(X = 2 \,|\, X \in \{3,4\}) = \frac{\mathbb{P}(\emptyset)}{\mathbb{P}(\{3,4\})} = 0. \tag{64}$$

▲

In the following statement, we show that conditioning on a fixed event induces a valid probability measure.

Proposition 3.2. *Let $(\Omega, \mathbb{P})$ be a probability space, and let $B \subseteq \Omega$ be an event such that $\mathbb{P}(B) > 0$. Then the conditional probability $\mathbb{P}(\cdot \,|\, B)$ is indeed a probability measure on Ω.*

PROOF. We need to verify the two axioms of probability measure. First, note that

$$\mathbb{P}(\emptyset \,|\, B) = \frac{\mathbb{P}(\emptyset \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\emptyset)}{\mathbb{P}(B)} = 0, \tag{65}$$

$$\mathbb{P}(\Omega \mid B) = \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1. \tag{66}$$

This verifies axiom (i). For axiom (ii)', let $A_1, A_2, \cdots$ be disjoint events. Then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i \,\middle|\, B\right) = \frac{\mathbb{P}\left[\left(\bigcup_{i=1}^{\infty} A_i\right) \cap B\right]}{\mathbb{P}(B)} \tag{67}$$

$$= \frac{\mathbb{P}\left(\bigcup_{i=1}^{\infty} (A_i \cap B)\right)}{\mathbb{P}(B)} \tag{68}$$

$$= \frac{1}{\mathbb{P}(B)} \sum_{i=1}^{\infty} \mathbb{P}(A_i \cap B) \tag{69}$$

$$= \sum_{i=1}^{\infty} \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} \tag{70}$$

$$= \sum_{i=1}^{\infty} \mathbb{P}(A_i \mid B). \tag{71}$$

Note that the second equality uses Exercise 1.5; The third one uses the fact that $A_i \cap B$'s are disjoint (since A_i's are) and axiom (ii) for the original probability measure $\mathbb{P}$. The last equality uses definition of conditional probability. □

Example 3.3. Consider rolling two four-sided dice and let (X, Y) be the outcome in the sample space $\Omega = \{0,1,2,3\}^2$. Suppose the two dice somehow affect each other according to the probability distribution on Ω is depicted in Figure 1.

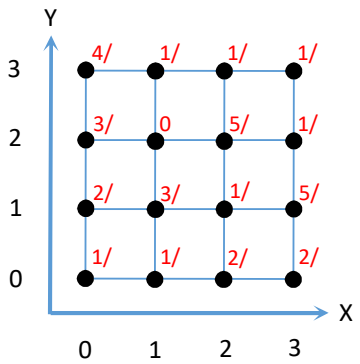

FIGURE 2. Joint distribution on Ω shown in red. Common denominator of 33 is omitted in the figure.

Then we can compute the conditional probability $\mathbb{P}(X \geq 2 \mid Y = 2)$ as below.

$$\mathbb{P}(X \geq 2 \mid Y = 2) = \mathbb{P}(X = 2 \mid Y = 2) + \mathbb{P}(X \geq 3 \mid Y = 2) \tag{72}$$

$$= \frac{\mathbb{P}(X = 2 \text{ and } Y = 2)}{\mathbb{P}(Y = 2)} + \frac{\mathbb{P}(X = 3 \text{ and } Y = 2)}{\mathbb{P}(Y = 2)} \tag{73}$$

$$= \frac{5/33}{(3+0+5+1)/33} + \frac{1/33}{(3+0+5+1)/33} = 2/3. \tag{74}$$

▲

Example 3.4. Consider tossing a fair coin three times. Identify the sample space as $\Omega = \{H, T\}^3$, where we suppose that all possible outcomes are equally likely with probability $1/8$. Consider the following two events:

$$A = \{\text{more heads than tails come up}\}, \qquad B = \{\text{second toss is a tail}\}. \tag{75}$$

Then note that

$$A = \{\text{two heads}\} \cup \{\text{three heads}\} \tag{76}$$
$$= \{(H,H,T),(H,T,H),(T,H,H)\} \cup \{(H,H,H)\}, \tag{77}$$

so by Exercise 2.4,

$$\mathbb{P}(A) = \frac{4}{8} = 1/2. \tag{78}$$

Now suppose we know that the event B occurs. How does this knowledge affects the probability of A? The new probability of A is given by the conditional probability $\mathbb{P}(A|B)$. To compute this, note that

$$B = \{(T,H,T),(T,H,H),(H,H,T),(H,H,H)\}, \quad A \cap B = \{(T,H,H),(H,H,T),(H,H,H)\}. \tag{79}$$

Hence

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{3/8}{4/8} = 3/4. \tag{80}$$

Furthermore, we can determine the full conditional probability measure $\mathbb{P}(\cdot \mid B)$. Recall that conditioning on the event B changes the sample space Ω to B. In order to describe a probability measure, we only need to specify its distribution (see Exercise 2.5). In this case, the distribution for $\mathbb{P}(\cdot \mid B)$ is given by

$$\mathbb{P}(\{(T,H,T)\} \mid B) = \frac{\mathbb{P}(\{(T,H,T)\})}{\mathbb{P}(B)} = 1/4, \tag{81}$$
$$\mathbb{P}(\{(T,H,H)\} \mid B) = \frac{\mathbb{P}(\{(T,H,T)\})}{\mathbb{P}(B)} = 1/4, \tag{82}$$
$$\mathbb{P}(\{(H,H,T)\} \mid B) = \frac{\mathbb{P}(\{(H,H,T)\})}{\mathbb{P}(B)} = 1/4, \tag{83}$$
$$\mathbb{P}(\{(H,H,H)\} \mid B) = \frac{\mathbb{P}(\{(H,H,H)\})}{\mathbb{P}(B)} = 1/4. \tag{84}$$

Hence $\mathbb{P}(\cdot \mid B)$ is the uniform probability measure on B. ▲

4. Partitioning the sample space and Bayes' theorem

4.1. Partitioning the sample space. Sometimes when we try to compute the probability of a certain event A, a 'divide and conquer' approach could be very useful. Namely, we divide the sample space into smaller and disjoint events A_i, and compute the probability of A conditioned on A_i, and then we add up the results. This technique of computing probability is called *partitioning*.

Proposition 4.1 (Partitioning)**.** *Let $(\Omega, \mathbb{P})$ be a probability space and let $A_1, A_2, \cdots, A_k \subseteq \Omega$ be a partition of Ω. That is, A_i's are disjoint and $\bigcup_{i=1}^k A_i = \Omega$. Then for any event $A \subseteq \Omega$,*

$$\mathbb{P}(A) = \mathbb{P}(A \mid A_1)\,\mathbb{P}(A_1) + \mathbb{P}(A \mid A_2)\,\mathbb{P}(A_2) + \cdots + \mathbb{P}(A \mid A_k)\,\mathbb{P}(A_k). \tag{85}$$

PROOF. Since $\bigcup_{i=1}^k A_i = \Omega$, we have

$$A = A \cap \Omega = A \cap \left(\bigcup_{i=1}^k A_i \right) = \bigcup_{i=1}^k A \cap A_i. \tag{86}$$

By definition of conditional probability,

$$\mathbb{P}(A \cap A_i) = \mathbb{P}(A \mid A_i)\mathbb{P}(A_i) \tag{87}$$

for each $1 \le i \le k$. Moreover, $A \cap A_i$'s are disjoint since A_i's are. Hence

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{i=1}^k A \cap A_i \right) = \sum_{i=1}^k \mathbb{P}(A \cap A_i) = \sum_{i=1}^k \mathbb{P}(A \mid A_i)\mathbb{P}(A_i). \tag{88}$$

□

Here are some examples.

Example 4.2. Suppose you role a fair die repeatedly until the first time you see an odd number. What is the probability that the total sum of outcomes is exactly 5?

Before we proceed, let us first identify the sample space Ω and probability measure $\mathbb{P}$. Note that we are not sure how many times we will have to roll the die. Here are some outcomes $\omega \in \Omega$ of the experiment to get started:

$$(1) \tag{89}$$

$$(2,4,5) \tag{90}$$

$$(2,6,6,2,4,3). \tag{91}$$

Hence Ω consists of the sequence of numbers from $\{1,2,3,4,5,6\}$ which consists of even numbers for all but the last coordinate. On the other hand, Observe that

$$\mathbb{P}(\{(1)\}) = 1/6, \tag{92}$$

$$\mathbb{P}(\{(2,4,5)\}) = \frac{1}{6}\cdot\frac{1}{6}\cdot\frac{1}{6} = \frac{1}{6^3}, \tag{93}$$

$$\mathbb{P}(\{(2,6,6,2,4,3)\}) = \frac{1}{6}\cdot\frac{1}{6}\cdot\frac{1}{6}\cdot\frac{1}{6}\cdot\frac{1}{6}\cdot\frac{1}{6} = \frac{1}{6^6}. \tag{94}$$

Hence, we observe that

$$\mathbb{P}(\omega) = \frac{1}{6^{|\omega|}}, \tag{95}$$

where $|\omega|$ denotes the number of coordinates in ω.

In order to use partitioning, we first need to decide how we partition the entire sample space Ω. Let us partition the sample space according to the first outcome; let A_i be the event that first roll gives number i. Then $\mathbb{P}(A_i) = 1/6$ for all $1 \le i \le 6$. On the other hand,

$$\mathbb{P}(A \mid A_1) = 0, \tag{96}$$

$$\mathbb{P}(A \mid A_3) = 0, \tag{97}$$

$$\mathbb{P}(A \mid A_5) = 1 \tag{98}$$

since the roll ends after the first step if i is odd. For the other cases, note that

$$\mathbb{P}(A \mid A_2) = \mathbb{P}(\{(2,2,1) \mid A_2\}) = \frac{1}{6^2} \tag{99}$$

$$\mathbb{P}(A \mid A_4) = \mathbb{P}(\{(4,1)\} \mid A_4) = \frac{1}{6} \tag{100}$$

$$\mathbb{P}(A \mid A_6) = 0. \tag{101}$$

Thus we have

$$\mathbb{P}(A) = \sum_{i=1}^{6} \mathbb{P}(A \mid A_i)\mathbb{P}(A_i) \tag{102}$$

$$= 0\cdot\frac{1}{6} + 0\cdot\frac{1}{6} + 1\cdot\frac{1}{6} + \frac{1}{6^2}\cdot\frac{1}{6} + \frac{1}{6}\cdot\frac{1}{6} + 0\cdot\frac{1}{6} = \frac{6^2+6+1}{6^3} = \frac{43}{6^3}. \tag{103}$$

▲

Exercise 4.3 (Laplace's rule of succession)**.** Laplace 'computed' the probability that the sun will rise tomorrow, given that it has risen for the preceeding 5000 years. The combinatorial model is as follow. Suppose we have N different coins, where the kth coin has probability k/N of coming up heads. We choose one of the N coin uniformly at random and flip it n times. For each $1 \le i \le n$, let R_i be the event that the ith flip comes up heads. We are interested in the following conditional probability

$$\mathbb{P}(R_{n+1} \mid R_n \cap R_{n-1} \cap \cdots \cap R_1). \tag{104}$$

If we think of coin coming up heads as the event of sun rising, then this is a model for the probability that the sun will rise tomorrow given that it has risen for the past n days.

(i) Write $R'_n = R_n \cap R_{n-1} \cap \cdots \cap R_1$. Show that

$$\mathbb{P}(R_{n+1} \mid R'_n) = \frac{\mathbb{P}(R_{n+1} \cap R'_n)}{\mathbb{P}(R'_n)} = \frac{\mathbb{P}(R'_{n+1})}{\mathbb{P}(R'_n)}. \tag{105}$$

(ii) For each $1 \le i \le n+1$, use partitioning to show that

$$\mathbb{P}(R'_i) = \sum_{k=1}^{N} \mathbb{P}(R'_i \mid \text{prob. of heads is } k/N)\mathbb{P}(\text{prob. of heads is } k/N) = \sum_{k=1}^{N} \left(\frac{k}{N}\right)^i \frac{1}{N}. \tag{106}$$

(iii) By considering the upper and lower Riemann sums we have

$$\sum_{k=0}^{N-1} \left(\frac{k}{N}\right)^i \frac{1}{N} \le \int_0^1 t^i \, dt \le \sum_{k=1}^{N} \left(\frac{k}{N}\right)^i \frac{1}{N}. \tag{107}$$

Using (ii) and (iii), show that

$$\mathbb{P}(R'_i) \approx \int_0^1 t^i \, dt = \frac{1}{i+1}, \tag{108}$$

where the approximation becomes exact as $N \to \infty$.

(iv) From (i), conclude that

$$\mathbb{P}(R_{n+1} \mid R'_n) \approx \frac{n+1}{n+2}. \tag{109}$$

For $n = 5000 * 365$ (days), we have $\mathbb{P}(R_{n+1} \mid R'_n) \approx 0.9999994520$. So it's pretty likely that the sun will rise tomorrow as well.

4.2. Inference and Bayes' Theorem. If we have a model for bitcoin price, then we can tune the parameters accordingly and attempt to predict its future prices. But how do we tune the parameters? Say the bitcoin price suddenly drops by 20% overnight (which is no surprise). Which factor is the most likely to have caused this drop? In general, Bayes' Theorem can be used to infer likely factors when we are given the effect or outcome. This is all based on conditional probability and partitioning.

We begin by the following trivial but important observation:

$$\mathbb{P}(B|A)\mathbb{P}(A) = \mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B). \tag{110}$$

And this is all we need.

Theorem 4.4 (Bayes' Theorem)**.** *Let* $(\Omega, \mathbb{P})$ *be a probability space, and let* $A_1, \cdots, A_k \subseteq \Omega$ *be events of positive probability that form a partition of* Ω*. Then*

$$\mathbb{P}(A_1 \mid B) = \frac{\mathbb{P}(B \mid A_1)\mathbb{P}(A_1)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \mid A_1)\mathbb{P}(A_1)}{\mathbb{P}(B \mid A_1)\mathbb{P}(A_1) + \mathbb{P}(B \mid A_2)\mathbb{P}(A_2) + \cdots + \mathbb{P}(B \mid A_k)\mathbb{P}(A_k)}. \tag{111}$$

PROOF. Note that (347) yields the first equality. Then partitioning and rewrite $\mathbb{P}(B)$ gives the second equality. □

What's so important about the Bayes' theorem is its interpretation as a means of *inference*, which is one of the fundamental tool in modern machine learning.

Example 4.5. Suppose Bob has a coin with unknown probability of heads, which we denote by Θ. This is called the *parameter*. Suppose Alice knows that Bob has one of the three kinds of coins A, B, and C, with probability of heads being $p_A = 0.2$, $p_B = 0.5$, and $p_C = 0.8$, respectively. This piece of information is called the *model*. Since Alice has no information, she initially assumes that Bob has one of the three coins equally likely. Namely, she assumes the uniform distribution over the sample space $\Omega = \{0.2, 0.5, 0.8\}$. This knowledge is called *prior*.

Now Bob flips his coin 10 times and got 7 heads, and reports this information, which we call Data, to Alice. Now that Alice has more information, she needs to update her *prior* to *posterior*, which is the probability distribution on Ω that best explains the Data. Namely, it is the conditional probability distribution $\mathbb{P}(\Theta = \theta \,|\, \texttt{Data})$.

First, using our prior, we compute $\mathbb{P}(\texttt{Data})$, the probability of seeing this particular data at hand. This can be done by partitioning:

$$\mathbb{P}(\texttt{Data}) = \mathbb{P}(\texttt{Data}\,|\,\Theta = 0.2)\mathbb{P}(\Theta = 0.2) + \mathbb{P}(\texttt{Data}\,|\,\Theta = 0.5)\mathbb{P}(\Theta = 0.5) \tag{112}$$

$$+ \mathbb{P}(\texttt{Data}\,|\,\Theta = 0.8)\mathbb{P}(\Theta = 0.8) \tag{113}$$

$$= \mathbb{P}(7/10 \text{ heads}\,|\,\Theta = 0.2)\frac{1}{3} + \mathbb{P}(7/10 \text{ heads}\,|\,\Theta = 0.5)\frac{1}{3} + \mathbb{P}(7/10 \text{ heads}\,|\,\Theta = 0.8)\frac{1}{3} \tag{114}$$

$$= \frac{1}{3}\left(\binom{10}{7}(0.2)^7(0.8)^3 + \binom{10}{7}(0.5)^7(0.5)^3 + \binom{10}{7}(0.8)^7(0.2)^3\right) \approx 0.1064, \tag{115}$$

where $\binom{10}{7} = 120$ is the number of ways to choose 7 out of 10 objects.

Second, we reformulate the first equality of Theorem 6.1 as

$$\mathbb{P}(\Theta\,|\,\texttt{Data}) = \frac{\mathbb{P}(\texttt{Data}\,|\,\Theta)\mathbb{P}(\Theta)}{\mathbb{P}(\texttt{Data})}. \tag{116}$$

Hence we can compute the posterior distribution by

$$\mathbb{P}(\Theta = 0.2\,|\,\texttt{Data}) = \frac{\mathbb{P}(\texttt{Data}\,|\,\Theta = 0.2)\mathbb{P}(\Theta = 0.2)}{\mathbb{P}(\texttt{Data})} = \frac{\binom{10}{7}(0.2)^7(0.8)^3\frac{1}{3}}{0.1064} \approx 0.0025 \tag{117}$$

$$\mathbb{P}(\Theta = 0.5\,|\,\texttt{Data}) = \frac{\mathbb{P}(\texttt{Data}\,|\,\Theta = 0.5)\mathbb{P}(\Theta = 0.5)}{\mathbb{P}(\texttt{Data})} = \frac{\binom{10}{7}(0.5)^7(0.5)^3\frac{1}{3}}{0.1064} \approx 0.3670 \tag{118}$$

$$\mathbb{P}(\Theta = 0.8\,|\,\texttt{Data}) = \frac{\mathbb{P}(\texttt{Data}\,|\,\Theta = 0.8)\mathbb{P}(\Theta = 0.8)}{\mathbb{P}(\texttt{Data})} = \frac{\binom{10}{7}(0.8)^7(0.2)^3\frac{1}{3}}{0.1064} \approx 0.6305. \tag{119}$$

Note that according to the posterior distribution, $\Theta = 0.8$ is the most likely value, which is natural given that we have 7 heads out of 10 flips. However, our knowledge is always incomplete so our posterior knowledge is still a probability distribution on the sample space.

What if Bob flips his coin another 10 times and reports only 3 heads to Alice? Then she will have to use her current prior $\pi = [0.0025, 0.3060, 0.6305]$ (which was obtained as the posterior in the previous round) to compute yet another posterior using the new data. This will be likely to give higher weight to $\Theta = 0.2$. ▲

Exercise 4.6. Suppose we have a prior distribution $\pi = [0.0025, 0.3680, 0.6305]$ on the sample space $\Omega = \{0.2, 0.5, 0.8\}$ for the inference problem of unknown parameter Θ. Suppose we are given the data that ten independent flips of probability Θ coin comes up heads twice. Compute the posterior distribution using this data and Bayesian inference.

Exercise 4.7. A test for pancreatic cancer is assumed to be correct %95 of the time: if a person has the cancer, the test results in positive with probability 0.95, and if the person does not have the cancer, then the test results in negative with probability 0.95. From a recent medical research, it is known that only %0.05 of the population have pancreatic cancer. Given that the person just tested positive, what is the probability of having the cancer?

5. Independence

When knowing something about one event does not yield any information of the other, we say the two events are independent. To make this statement a bit more precise, suppose Bob is betting \$5 for whether an event E_1 occurs or not. Suppose Alice tells him that some other event E_2 holds true. If Bob

can somehow leverage on this knowledge to increase his chance of winning, then we should say the two events are not independent.

Formally, we say two events E_1 and E_2 are *independent* if

$$\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2). \tag{120}$$

We say two events or RVs are *dependent* if they are not independent.

Example 5.1. Flip two fair coins at the same time and let X and Y be their outcome, labeled by H and T. Clearly knowing about one coin does not give any information of the other. For instance, the first coin lands on heads with probability $1/2$. Whether the first coin lands on heads or not, the second coin will land on heads with probability $1/2$. So

$$\mathbb{P}(X = H \text{ and } Y = H) = \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(X = H)\mathbb{P}(Y = H). \tag{121}$$

▲

Exercise 5.2. Suppose two events A_1 and A_2 are independent and suppose that $\mathbb{P}(A_2) > 0$. Show that

$$\mathbb{P}(A_1 \mid A_2) = \mathbb{P}(A_1). \tag{122}$$

In terms of the inference problem, the above equation can be written as

$$\mathbb{P}(\Theta \mid \texttt{Data}) = \mathbb{P}(\Theta). \tag{123}$$

That is, after observing `Data`, our prior distribution does not change. This is because the new data does not contain any new information.

Example 5.3. Consider rolling two 4-sided dice, where all 16 outcomes are equally likely. Let $\Omega = \{1,2,3,4\}^2$ be the sample space. Let X and Y be the outcome of the two dice.

(i) Note that for any $(i, j) \in \Omega$,

$$\mathbb{P}(X = i \text{ and } Y = j) = \mathbb{P}(\{(i, j)\}) = \frac{1}{16} \tag{124}$$

$$\mathbb{P}(X = i) = \mathbb{P}(\{(i,1),(i,2),(i,3),(i,4)\}) = \frac{1}{4} \tag{125}$$

$$\mathbb{P}(X = j) = \mathbb{P}(\{(1, j),(2, j),(3, j),(4, j)\}) = \frac{1}{4}. \tag{126}$$

Hence $\mathbb{P}(X = i \text{ and } Y = j) = \mathbb{P}(X = i)\mathbb{P}(Y = j)$ and the events $\{X = i\}$ and $\{Y = j\}$ are independent.

(ii) Let $A = \{X = 1\}$ and $B = \{X + Y = 5\}$. Are these events independent? It seems like knowing that A occurs yields some information regarding B. However, it turns out that conditioning on A does not change the probability of B occuring. Namely,

$$\mathbb{P}(A \cap B) = \mathbb{P}(\{(1,4)\}) = \frac{1}{16} \tag{127}$$

$$\mathbb{P}(A) = \mathbb{P}(\{(1,1),(1,2),(1,3),(1,4)\}) = \frac{1}{4} \tag{128}$$

$$\mathbb{P}(B) = \mathbb{P}(\{(1,4),(2,3),(3,2),(4,1)\}) = \frac{1}{4}. \tag{129}$$

Hence $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, so they are independent. In other words,

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(\{(1,4)\})}{\mathbb{P}(\{(1,1),(1,2),(1,3),(1,4)\})} = \frac{1}{4} = \mathbb{P}(\{(1,4),(2,3),(3,2),(4,1)\}) = \mathbb{P}(B). \tag{130}$$

(iii) Let $A = \{\min(X, Y) = 2\}$ and $B = \{\max(X, Y) = 2\}$. Then

$$\mathbb{P}(A \cap B) = \mathbb{P}(\{(2,2)\}) = \frac{1}{16} \tag{131}$$

$$\mathbb{P}(A) = \mathbb{P}(\{(2,2),(2,3),(2,4),(3,2),(4,2)\}) = \frac{5}{16} \tag{132}$$

$$\mathbb{P}(B) = \mathbb{P}(\{(1,2),(2,2),(2,1)\}) = \frac{3}{16}, \tag{133}$$

so $\mathbb{P}(A \cap B) \neq \mathbb{P}(A)\mathbb{P}(B)$. Hence they are not independent.

▲

There is a conditional version of the notion of independence. Let A, B, C be events such that C has positive probability. We say A and B are *independent conditional on* C if

$$\mathbb{P}(A \cap B \,|\, C) = \mathbb{P}(A \,|\, C)\mathbb{P}(B \,|\, C). \tag{134}$$

Below is an alternative characterization of conditional independence.

Proposition 5.4. *A and B ard independent conditional on C if and only if*

$$\mathbb{P}(A \,|\, C) = \mathbb{P}(A \,|\, B \cap C). \tag{135}$$

PROOF. Note that

$$\mathbb{P}(A \cap B \,|\, C) = \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(C)} = \frac{\mathbb{P}(C)\mathbb{P}(B \,|\, C)\mathbb{P}(A \,|\, B \cap C)}{\mathbb{P}(C)} = \mathbb{P}(B \,|\, C)\mathbb{P}(A \,|\, B \cap C). \tag{136}$$

Hence by comparing with (134), we obtain the assertion. □

In other words, when we know that C have occurred, A and B are conditionally independent when the extra knowledge that B also have occurred does not change the probability of A.

Example 5.5. Flip two independent fair coins. Define three events A_1, A_2, and B by

$$A_1 = \{\text{1st toss is a head}\} \tag{137}$$

$$A_2 = \{\text{2nd toss is a head}\} \tag{138}$$

$$B = \{\text{the two tosses give different results}\} \tag{139}$$

Clearly A_1 and A_2 are independent. However, note that

$$\mathbb{P}(A_1 \cap A_2 \,|\, B) = \mathbb{P}(\emptyset) = 0, \tag{140}$$

$$\mathbb{P}(A_1 \,|\, B) = \frac{\mathbb{P}(\{(H,T)\})}{\mathbb{P}(\{(H,T),(T,H)\})} = \frac{1}{2}, \tag{141}$$

$$\mathbb{P}(A_2 \,|\, B) = \frac{\mathbb{P}(\{(T,H)\})}{\mathbb{P}(\{(H,T),(T,H)\})} = \frac{1}{2}. \tag{142}$$

Hence A_1 and A_2 are not conditionally independent given B. ▲

Exercise 5.6. Suppose Bob has a coin of unknown probability of heads, Θ, from the sample space $\Omega = \{0.2, 0.9\}$. Alice believes that the two probabilities are equally likely. Bob tosses his coin twice independently. Let H_i be the event that the ith toss comes up heads, for $i = 1, 2$. Let $B = \{\Theta = 0.2\}$.

(i) Show that the events H_1 and H_2 are independent conditional on B, regardless of Alice's belief.

(ii) Show that the events H_1 and H_2 are *not* independent under Alice's belief. Is there any prior for Alice such that H_i's are independent?

CHAPTER 2

Random variables

In this note, we introduce the notion of random variables along with a number of important examples. We also introduce expectation and variance as summary statistics for random variables, and learn how to compute them.

1. Discrete random variables

Given a finite probability space $(\Omega, \mathbb{P})$, a (discrete) *random variable* (RV) is any real-valued function $X : \Omega \to \mathbb{R}$. We can think of it as the outcome of some experiment on Ω (e.g., height of a randomly selected friend). We often forget the original probability space and specify a RV by *probability mass function* (PMF) $f_X : \mathbb{R} \to [0,1]$,

$$f_X(x) = \mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) = x\}). \tag{143}$$

Namely, $\mathbb{P}(X = x)$ is the likelihood that the RV X takes value x.

Example 1.1. Say you win \$1 if a fair coin lands heads and lose \$1 if lands tails. We can set up our probability space $(\Omega, \mathbb{P})$ by $\Omega = \{H, T\}$ and $\mathbb{P}$ = uniform probability measure on Ω. The RV $X : \Omega \to \mathbb{R}$ for this game is $X(H) = 1$ and $X(T) = -1$. The PMF of X is given by $f_X(1) = \mathbb{P}(X = 1) = \mathbb{P}(\{H\}) = 1/2$ and likewise $f_X(-1) = 1/2$.

Exercise 1.2. Let $(\Omega, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ be a RV. Let f_X be the PMF of X, that is, $f_X(x) = \mathbb{P}(X = x)$ for all x. Show that f_X adds up to 1, that is,

$$\sum_x f_X(x) = 1, \tag{144}$$

where the summation runs over all numerical values x that X can take.

There are two useful statistics of a RV to summarize its two most important properties: Its average and uncertainty. First, if one has to guess the value of a RV X, what would be the best choice? It is the *expectation* (or mean) of X, defined as below:

$$\mathbb{E}(X) = \sum_x x\,\mathbb{P}(X = x). \tag{145}$$

Example 1.3 (Doubling strategy)**.** Suppose we bet \$$x$ on a game where we have to predict whether a fair coin clip comes up heads. We win \$$x$ on heads and lose \$$x$ on tails. Suppose we are playing the 'doubling strategy'. Namely, we start betting \$1, and until the first time we win, we double our bet; upon the first win, we quit the game. Let X be the random variable giving the net gain of the overall game. How can we evaluate this strategy?

Let N be the random number of coin flips we have to encounter until we see the first head. For instance,

$$\mathbb{P}(N = 1) = \mathbb{P}(\{H\}) = 1/2, \tag{146}$$

$$\mathbb{P}(N = 2) = \mathbb{P}(\{(T, H)\}) = 1/2^2 \tag{147}$$

$$\mathbb{P}(N = 3) = \mathbb{P}(\{(T, T, H)\}) = 1/2^3. \tag{148}$$

In general, we have

$$\mathbb{P}(N = k) = 1/2^k. \tag{149}$$

Note that on the event that $N = k$ (i.e., we bet k times), the net gain $X|\{N = k\}$ is

$$X|\{N = k\} = (-1) + (-2) + (-2^2) + (-2^3) + \cdots + (-2^{k-1}) + 2^k \tag{150}$$

$$= -(2^k - 1) + 2^k = 1. \tag{151}$$

Since the last expression do not depend on k, we conclude that

$$\mathbb{P}(X = 1) = \sum_{k=1}^{\infty} \mathbb{P}(X = 1 \mid N = k)\mathbb{P}(N = k) \tag{152}$$

$$= \sum_{k=1}^{\infty} \mathbb{P}(N = k) = 1. \tag{153}$$

Hence our net gain is \$1 with probability 1. In particular, the expected net gain is also 1:

$$\mathbb{E}[X] = 1 \cdot \mathbb{P}(X = 1) = 1. \tag{154}$$

What if we use tripling strategy? ▲

Exercise 1.4. For any RV X and real numbers $a, b \in \mathbb{R}$, show that

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b. \tag{155}$$

Exercise 1.5 (Tail sum formula for expectation)**.** Let X be a RV taking values on positive integers.

(i) For any x, show that

$$\mathbb{P}(X \geq x) = \sum_{y=x}^{\infty} \mathbb{P}(X = y). \tag{156}$$

(ii) Use (i) and Fubini's theorem to show

$$\sum_{x=1}^{\infty} \mathbb{P}(X \geq x) = \sum_{y=1}^{\infty} \sum_{x=1}^{y} \mathbb{P}(X = y) \tag{157}$$

(iii) From (ii), conclude that

$$\mathbb{E}(X) = \sum_{x=1}^{\infty} \mathbb{P}(X \geq x). \tag{158}$$

On the other hand, say you play two different games where in the first game, you win or lose \$1 depending on a fair coin flip, and in the second game, you win or lose \$10. In both games, your expected winning is 0. But the two games are different in how much the outcome fluctuates around the mean. This notion if fluctuation is captured by the following quantity called *variance*:

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]. \tag{159}$$

Namely, it is the expected squared difference between X and its expectation $\mathbb{E}(X)$.

Here are some of the simplest and yet most important RVs.

Exercise 1.6 (Bernoulli RV)**.** A RV X is a *Bernoulli* variable with (success) probability $p \in [0, 1]$ if it takes value 1 with probability p and 0 with probability $1 - p$. In this case we write $X \sim \mathrm{Bernoulli}(p)$. Show that $\mathbb{E}(X) = p$ and $\mathrm{Var}(X) = p(1 - p)$.

Exercise 1.7 (Indicator variables)**.** Let $(\Omega, \mathbb{P})$ be a probability space and let $E \subseteq \Omega$ be an event. The *indicator variable* of the event E, which is denoted by $\mathbf{1}_E$, is the RV such that $\mathbf{1}_E(\omega) = 1$ if $\omega \in E$ and $\mathbf{1}_E(\omega) = 0$ if $\omega \in E^c$. Show that $\mathbf{1}_E$ is a Bernoulli variable with success probability $p = \mathbb{P}(E)$.

The following exercise ties the expectation and the variance of a RV into a problem of finding a point estimator that minimizes the mean squared error.

Exercise 1.8 (Variance as minimum MSE)**.** Let X be a RV. Let $\hat{x} \in \mathbb{R}$ be a number, which we consider as a 'guess' (or 'estimator' in Statistics) of X. Let $\mathbb{E}[(X - \hat{x})^2]$ be the *mean squared error* (MSE) of this estimation.

(i) Show that

$$\mathbb{E}_Y[(X-\hat{x})^2] = \mathbb{E}_Y[X^2] - 2\hat{x}\mathbb{E}[X] + \hat{x}^2 \tag{160}$$
$$= (\hat{x}-\mathbb{E}[X])^2 + \mathbb{E}[X^2] - \mathbb{E}[X]^2 \tag{161}$$
$$= (\hat{x}-\mathbb{E}[X])^2 + \mathrm{Var}(X). \tag{162}$$

(ii) Conclude that the MSE is minimized when $\hat{x} = \mathbb{E}[X]$ and the global minimum is $\mathrm{Var}(X)$. In this sense, $\mathbb{E}[X]$ is the 'best guess' for X and $\mathrm{Var}(X)$ is the corresponding MSE.

2. Expectation and variance of sums of RVs

In this section, we learn how we can compute expectation and variance for sums of RVs. For expectation, we will see that we can swap the order to summation and expectation. This is called the *linearity of expectation*, whose importance cannot be overemphasized.

Exercise 2.1 (Linearity of expectation)**.** In this exercise, we will show that the expectation of sum of RVs is the sum of expectation of individual RVs.

(i) Let X and Y be RVs. Show that

$$\sum_y \mathbb{P}(X=x, Y=y) = \mathbb{P}(X=x). \tag{163}$$

(ii) Verify the following steps:

$$\mathbb{E}(X+Y) = \sum_z z\,\mathbb{P}(X+Y=z) \tag{164}$$
$$= \sum_z \sum_{\substack{x,y \\ x+y=z}} (x+y)\mathbb{P}(X=x,\, Y=y) \tag{165}$$
$$= \sum_{x,y} (x+y)\,\mathbb{P}(X=x,\, Y=y) \tag{166}$$
$$= \sum_{x,y} x\,\mathbb{P}(X=x,\, Y=y) + \sum_{x,y} y\,\mathbb{P}(X=x,\, Y=y) \tag{167}$$
$$= \sum_x x\left(\sum_y \mathbb{P}(X=x, y=y)\right) + \sum_y y\left(\sum_x \mathbb{P}(X=x,\, Y=y)\right) \tag{168}$$
$$= \sum_x x\mathbb{P}(X=x) + \sum_y y\mathbb{P}(Y=y) \tag{169}$$
$$= \mathbb{E}(X) + \mathbb{E}(Y). \tag{170}$$

(iii) Use induction to show that for any RVs $X_1, X_2, \cdots, X_n$, we have

$$\mathbb{E}(X_1 + X_2 + \cdots + X_n) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_n). \tag{171}$$

Our first application of linearity of expectation is the following useful formula for variance.

Exercise 2.2. For any RV X, show that

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2. \tag{172}$$

Next, we will see a nice application of linearity of expectation.

Exercise 2.3 (Inclusion-exclusion)**.** Let $(\Omega, \mathbb{P})$ be a probability space. Let $A_1, A_2, \cdots, A_k \subseteq \Omega$. We will show the following inclusion-exclusion principle:

$$\mathbb{P}\left(\bigcup_{i=1}^k A_i\right) = \sum_{i_1=1}^k \mathbb{P}(A_{i_1}) - \sum_{1\le i_1<i_2\le k} \mathbb{P}(A_{i_1} \cap A_{i_2}) \tag{173}$$

$$+\sum_{1\le i_1<i_2<i_3\le k}\mathbb{P}(A_{i_1}\cap A_{i_2}\cap A_{i_3})-\cdots+(-1)^k\mathbb{P}\left(\bigcap_{i=1}^k A_i\right). \tag{174}$$

The method is so-called the 'indicator trick'.

For each $1\le i\le k$, let $X_i=\mathbf{1}(A_i)$ be the indicator variable for the event A_i. Consider the following RV

$$Y=(1-X_1)(1-X_2)\cdots(1-X_k). \tag{175}$$

(i) By expanding the product and using linearity of expectation, show that

$$\mathbb{E}[Y]=1-\sum_{i_1=1}^k\mathbb{E}[X_{i_1}]+\sum_{1\le i_1<i_2\le k}\mathbb{E}[X_{i_1}X_{i_2}] \tag{176}$$

$$-\sum_{1\le i_1<i_2<i_3\le k}\mathbb{E}[X_{i_1}X_{i_2}X_{i_3}]+\cdots-(-1)^k\mathbb{E}[X_1X_2\cdots X_k]. \tag{177}$$

(ii) Show that Y is the indicator variable of the event $\bigcap_{i=1}^k A_i^c$. Conclude that

$$\mathbb{E}[Y]=\mathbb{P}\left(\bigcap_{i=1}^k A_i^c\right)=1-\mathbb{P}\left(\bigcup_{i=1}^k A_i\right). \tag{178}$$

(iii) From (i) and (ii), deduce the inclusion-exclusion principle.

Exercise 2.4 (Finite second moment implies finite first moment)**.** Let X be a RV with $\mathbb{E}[X^2]<\infty$.

(i) Show that

$$\mathbb{E}[|X|]=\mathbb{E}[|X|\mathbf{1}(|X|\le 1)]+\mathbb{E}[|X|\mathbf{1}(|X|\le 1)]\le 1+\mathbb{E}[X^2\mathbf{1}(|X|\le 1)]\le 1+\mathbb{E}[X^2]<\infty. \tag{179}$$

Deduce that $\mathbb{E}[X]\in(-\infty,\infty)$.

(ii) Deduce that $\mathrm{Var}(X)<\infty$.

Next, we study how we can simplify the variance of sums of RVs. Unlike expectation, we cannot exchange the order of summation and variance in general.

Example 2.5. Flip two coins at the same time. The sample space for this experiment is $\Omega=\{H,T\}^2$ and the probability measure $\mathbb{P}$ on Ω is given by

$$\mathbb{P}(\{(H,H)\})=\mathbb{P}(\{(T,T)\})=0 \tag{180}$$

$$\mathbb{P}(\{(H,T)\})=\mathbb{P}(\{(T,H)\})=1/2. \tag{181}$$

Namely somehow we know for sure that the two coins will come up with the opposite sides with equal probability.

Define a RV $X:\Omega\to\{-1,1\}$ by

$$X=\begin{cases}1 & \text{if the first coin comes up a head}\\ -1 & \text{if the first coin comes up a tail}\end{cases} \tag{182}$$

Define another RV $Y:\Omega\to\{-1,1\}$ similarly for the second coin. Then

$$\mathbb{E}[X]=1\cdot\mathbb{P}(X=1)+(-1)\mathbb{P}(X=-1) \tag{183}$$

$$=\mathbb{P}(\{(H,H),(H,T)\})-\mathbb{P}(\{(T,H),(T,T)\})=0, \tag{184}$$

$$\mathbb{E}[X^2]=1^2\cdot\mathbb{P}(X=1)+(-1)^2\mathbb{P}(X=-1)=1, \tag{185}$$

$$\mathrm{Var}(X)=\mathbb{E}[X^2]-\mathbb{E}[X]^2=1. \tag{186}$$

A similar computation shows $\mathrm{Var}(Y)=1$. However, note that $X+Y$ takes value 0 with probability 1. Hence $\mathrm{Var}(X+Y)=0$. Thus $0=\mathrm{Var}(X+Y)\neq\mathrm{Var}(X)+\mathrm{Var}(Y)=2$. ▲

In general, when we try to write down the variance of a sum of RVs, in addition to the variance of each RV, we have extra contribution from each pairs of RVs called the covariance.

Exercise 2.6. In this exercise, we will see how we can express the variance of sums of RVs. For two RVs X and Y, define their *covariance* $\mathrm{Cov}(X,Y)$ by

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y). \tag{187}$$

(i) Use Exercises 2.2 and 2.1 to show that

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y). \tag{188}$$

(ii) Use induction to show that for RVs $X_1, X_2 \cdots, X_n$

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{1\le i,j\le n} \mathrm{Cov}(X_i, X_j) \tag{189}$$

When knowing something about one RV does not yield any information of the other, we say the two RVs are independent. Formally, we say two RVs X and Y are *independent* if for any two subsets $A_1, A_2 \subseteq \mathbb{R}$,

$$\mathbb{P}(X \in A_1 \text{ and } Y \in A_2) = \mathbb{P}(X \in A_1)\mathbb{P}(Y \in A_2). \tag{190}$$

We say two events or RVs are *dependent* if they are not independent.

Exercise 2.7. Suppose two RVs X and Y are independent. Then for any subsets $A_1, A_2 \subseteq \mathbb{R}$ such that $\mathbb{P}(Y \in A_2) > 0$, show that

$$\mathbb{P}(X \in A_1 \mid Y \in A_2) = \mathbb{P}(X \in A_1). \tag{191}$$

Example 2.8. Flip two fair coins at the same time, and let $X = 1$ if the first coin lands heads and $X = -1$ if it lands tails. Let Y be a similar RV for the second coin. Suppose each of the four outcomes in $\Omega = \{H,T\}^2$ are equality likely. Clearly knowing about one coin does not give any information of the other. For instance, the first coin lands on heads with probability $1/2$. Whether the first coin lands on heads or not, the second coin will land on heads with probability $1/2$. So

$$\mathbb{P}(X = 1 \text{ and } Y = 1) = \frac{1}{2}\cdot\frac{1}{2} = \mathbb{P}(X=1)\mathbb{P}(Y=1). \tag{192}$$

One can verify the above equation for possible outcomes. Hence X and Y are independent. ▲

Exercise 2.9. Let X, Y be two indepenent RVs. Show that

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]. \tag{193}$$

In the following exercise, we will see that we can swap summation and variance as long as the RVs we are adding up are independent.

Exercise 2.10. Recall the definition of covariance given in Exercise 2.6.

(i) Show that if two RVs X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$

(ii) Use Exercise 2.6 to conclude that if $X_1, \cdots, X_n$ are independent RVs, then

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n). \tag{194}$$

3. Binomial, geometric, and Poisson RVs

Example 3.1 (Binomial RV)**.** Let $X_1, X_2, \cdots, X_n$ be independent and identically distributed Bernoulli p variables. Let $X = X_1 + \cdots + X_n$. One can think of flipping the same probability p coin n times. Then X is the total number of heads. Note that X has the following PMF

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^k \tag{195}$$

for k nonnegative integer, and $\mathbb{P}(X = k) = 0$ otherwise. We say X follows the Binomial distribution with parameters n and p, and write $X \sim \mathrm{Binomial}(n,p)$.

We can compute the mean and variance of X using the above PMF directly, but it is much easier to break it up into Bernoulli variables and use linearity. Recall that $X_i \sim \text{Bernoulli}(p)$ and we have $\mathbb{E}(X_i) = p$ and $\text{Var}(X_i) = p(1-p)$ for each $1 \le i \le n$ (from Exercise 1.6). So by linearity of expectation (Exercise 2.1),

$$\mathbb{E}(X) = \mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n) = np. \tag{196}$$

On the other hand, since X_i's are independent, variance of X is the sum of variance of X_i's (Exercise 2.10) so

$$\text{Var}(X) = \text{Var}(X_1 + \cdots + X_n) = \text{Var}(X_1) + \cdots + \text{Var}(X_n) = np(1-p). \tag{197}$$

Example 3.2 (Geometric RV)**.** Suppose we flip a probability p coin until it lands heads. Let X be the total number of trials until the first time we see heads. Then in order for $X = k$, the first $k-1$ flips must land on tails and the kth flip should land on heads. Since the flips are independent with each other,

$$\mathbb{P}(X = k) = \mathbb{P}(\{T, T, \cdots, T, H\}) = (1-p)^{k-1} p. \tag{198}$$

This is valid for k positive integer, and $\mathbb{P}(X = k) = 0$ otherwise. Such a RV is called a *Geometric RV* with (success) parameter p, and we write $X \sim \text{Geom}(p)$.

The mean and variance of X can be easily computed using its moment generating function, which we will learn soon in this course. For their direct computation, note that

$$\begin{aligned}
\mathbb{E}(X) - (1-p)\mathbb{E}(X) &= (1-p)^0 p + 2(1-p)^1 p + 3(1-p)^2 p + 4(1-p)^3 p \cdots && (199)\\
&\quad - \left[(1-p)^1 p + 2(1-p)^2 p + 3(1-p)^3 p + \cdots\right] && (200)\\
&= (1-p)^0 p + (1-p)^1 p + (1-p)^2 p + (1-p)^3 p \cdots && (201)\\
&= \frac{p}{1-(1-p)} = 1, && (202)
\end{aligned}$$

where we recognized the series after the second equality as a geometric series. This gives

$$\mathbb{E}(X) = 1/p. \tag{203}$$

(In fact, one can apply Exercise 1.5 and quickly compute the expection of a Geometric RV.)

Exercise 3.3. Let $X \sim \text{Geom}(p)$. Use a similar computation as we had in Example 3.2 to show $\mathbb{E}(X^2) = (2-p)/p^2$. Using the fact that $\mathbb{E}(X) = 1/p$, conclude that $\text{Var}(X) = (1-p)/p^2$.

Example 3.4 (Poisson RV)**.** A RV X is a *Poisson RV* with rate $\lambda > 0$ if

$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{204}$$

for all nonnegative integers $k \ge 0$. We write $X \sim \text{Poisson}(\lambda)$.

Poisson distribution is obtained as a limit of the Binomial distribution as the number n of trials tend to infinity while the mean np is kept at constant λ. Namely, let $Y \sim \text{Binomial}(n, p)$ and suppose $np = \lambda$. This means that we expect to see λ successes out of n trials. Then what is the probability that we see, say, k successes out of n trials, when n is large? Since the mean is λ, this probability should be very small when k is large compared to λ. Indeed, we can rewrite the Binomial PMF as

$$\begin{aligned}
\mathbb{P}(Y = k) &= \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} p^k (1-p)^{n-k} && (205)\\
&= \frac{n}{n}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{k-1}{n}\right)\frac{(np)^k}{k!}(1-p)^{n-k} && (206)\\
&= \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{k-1}{n}\right)\frac{\lambda^k}{k!}\left(1 - \frac{\lambda}{n}\right)^{n-k}. && (207)
\end{aligned}$$

As n tends to infinity, the limit of the last expression is precisely the right hand side of (204). [1]

Exercise 3.5. 1.21 Let $X \sim \text{Poisson}(\lambda)$. Show that $\mathbb{E}(X) = \text{Var}(X) = \lambda$.

4. Continuous Random Variables

So far we have only considered discrete RVs, which takes either finitely many or countably many values. While there are many examples of discrete RVs, there are also many instances of RVs which various continuously (e.g., temperature, height, weight, price, etc.). To define a discrete RV, it was enough to specify its PMF. For a continuous RV, *probability distribution function* (PDF) plays an analogous role of PMF. We also need to replace summation $\sum$ with an integral $\int dx$.

Namely, X is a *continuous RV* if there is a function $f_X : \mathbb{R} \to [0, \infty)$ such that for any interval $[a, b]$, the probability that X takes a value from an interval $(a, b]$ is given by integrating f_X over the interval $(a, b]$:

$$\mathbb{P}(X \in (a, b]) = \int_a^b f_X(x)\, dx. \tag{208}$$

The *cumulative distribution function* (CDF) of a RV X (either discrete or continuous), denoted by F_X, is defined by

$$F_X(x) = \mathbb{P}(X \le x). \tag{209}$$

By definition of PDF, we get

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt. \tag{210}$$

Conversely, PDFs can be obtained by differentiating corresponding CDFs.

Exercise 4.1. Let X be a continuous RV with PDF f_X. Let a be a continuity point of f_X, that is, f_X is continuous at a. Show that $F_X(x)$ is differentiable at $x = a$ and

$$\frac{dF_X}{dx}\Big|_{x=a} = f_X(a). \tag{211}$$

The expectation of a continuous RV X with pdf f_X is defined by

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx, \tag{212}$$

and its variance $\text{Var}(X)$ is defined by the same formula (159).

Exercise 4.2 (Tail sum formula for expectation)**.** Let X be a continuous RV with PDF f_X and suppose $f_X(x) = 0$ for all $x < 0$. Use Fubini's theorem to show that

$$\mathbb{E}(X) = \int_0^{\infty} \mathbb{P}(X \ge t)\, dt. \tag{213}$$

Example 4.3 (Higher moments)**.** Let X be a continuous RV with PDF f_X and suppose $f_X(x) = 0$ for all $x < 0$. We will show that for any real number $\alpha > 0$,

$$\mathbb{E}[X^\alpha] = \int_0^{\infty} x^\alpha f_X(x)\, dx. \tag{214}$$

First, Use Exercise 4.2 and to write

$$\mathbb{E}[X^\alpha] = \int_0^{\infty} \mathbb{P}(X^\alpha \ge x)\, dx \tag{215}$$

$$= \int_0^{\infty} \mathbb{P}(X \ge x^{1/\alpha})\, dx \tag{216}$$

[1] Later, we will interpret the value of a Poisson variable $X \sim \text{Poisson}(\lambda)$ as the number of customers arriving during a unit time interval, where the waiting time between consecutive customers is distributed as an independent exponential distribution with mean $1/\lambda$. Such an arrival process is called the Poisson process.

$$= \int_0^\infty \int_{x^{1/\alpha}}^\infty f_X(t)\,dt\,dx. \tag{217}$$

We then use Fubini's theorem to change the order of integral. This gives

$$\mathbb{E}(X) = \int_0^\infty \int_{x^{1/\alpha}}^\infty f_X(t)\,dt\,dx = \int_0^\infty \int_0^{t^\alpha} f_X(t)\,dx\,dt = \int_0^\infty t^\alpha f_X(t)\,dt, \tag{218}$$

as desired. ▲

Exercise 4.4. Let X be a continuous RV with PDF f_X and suppose $f_X(x) = 0$ for all $x < 0$. Let $g : \mathbb{R} \to \mathbb{R}$ be a strictly increasing function. Use Fubini's theorem and tail sum formula for expectation to show

$$\mathbb{E}[g(X)] = \int_0^\infty g(x) f_X(x)\,dx. \tag{219}$$

5. Uniform, exponential, and normal RVs

In this section, we introduce three important continuous RVs.

Example 5.1 (Uniform RV)**.** X is a *uniform RV* on the interval $[a, b]$ (denoted by $X \sim \text{Uniform}([a, b])$) if it has PDF

$$f_X(x) = \frac{1}{b-a}\mathbf{1}(a \le x \le b). \tag{220}$$

An easy computation gives its CDF:

$$\mathbb{P}(X \le x) = \begin{cases} 0 & x < a \\ (x-a)/(b-a) & a \le x \le b \\ 1 & x > b. \end{cases} \tag{221}$$

▲

Exercise 5.2. Let $X \sim \text{Uniform}([a, b])$. Show that

$$\mathbb{E}(X) = \frac{a+b}{2}, \qquad \text{Var}(E) = \frac{(b-a)^2}{12}. \tag{222}$$

Example 5.3 (Exponential RV)**.** X is an *exponential* RV with rate λ (denoted by $X \sim \text{Exp}(\lambda)$) if it has PDF

$$f_X(x) = \lambda e^{-\lambda x}\mathbf{1}(x \ge 0). \tag{223}$$

Integrating the PDF gives its CDF

$$\mathbb{P}(X \le x) = (1 - e^{-\lambda x})\mathbf{1}(x \ge 0). \tag{224}$$

Using Exercise 4.2, we can compute

$$\mathbb{E}(X) = \int_0^\infty e^{-\lambda t}\,dt = \left[-\frac{e^{-\lambda t}}{\lambda}\right]_0^\infty = 1/\lambda. \tag{225}$$

▲

Exercise 5.4. Let $X \sim \text{Exp}(\lambda)$. Show that $\mathbb{E}(X) = 1/\lambda$ directly using definition (212). Also show that $\text{Var}(X) = 1/\lambda^2$.

Example 5.5 (Normal RV)**.** X is a *normal* RV with mean μ and variance σ^2 (denoted by $X \sim N(\mu, \sigma^2)$) if it has PDF

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{226}$$

If $\mu = 0$ and $\sigma^2 = 1$, then X is called a standard normal RV. Note that if $X \sim N(\mu, \sigma^2)$, then $Y := X - \mu$ has PDF

$$f_Y(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}. \tag{227}$$

Since this is an even function, it follows that $\mathbb{E}(Y) = 0$. Hence $\mathbb{E}(X) = \mu$. ▲

Exercise 5.6 (Gaussian integral)**.** In this exercise, we will show $\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}$.

(i) Show that

$$\int x e^{-x^2}\,dx = -\frac{1}{2}e^{-x^2} + C. \tag{228}$$

(ii) Let $I = \int_{-\infty}^{\infty} e^{-x^2}\,dx$. Show that

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)}\,dx\,dy. \tag{229}$$

(iii) Use polar coordinate (r,θ) to rewrite

$$I^2 = \int_0^{2\pi}\int_0^{\infty} e^{-r^2}\,r\,dr\,d\theta = 2\pi\int_0^{\infty} re^{-r^2}\,dr. \tag{230}$$

Then use (i) to deduce $I^2 = \pi$. Conclude $I = \sqrt{\pi}$.

Exercise 5.7. Let $X \sim N(\mu,\sigma^2)$. In this exercise, we will show $\mathrm{Var}(X) = \sigma^2$.

(i) Show that $\mathrm{Var}(X) = \mathrm{Var}(X-\mu)$.

(ii) Use integration by parts and Exercise 5.6 to show that

$$\int_0^{\infty} x^2 e^{-x^2}\,dx = \left[x\left(-\frac{1}{2}e^{-x^2}\right)\right]_0^{\infty} + \int_0^{\infty}\frac{1}{2}e^{-x^2}\,dx = \frac{\sqrt{\pi}}{4}. \tag{231}$$

(iii) Use change of variable $x = \sqrt{2}\sigma t$ and (ii) to show

$$\int_0^{\infty}\frac{x^2}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2}{2\sigma^2}}\,dx = \frac{2\sigma^2}{\sqrt{\pi}}\int_0^{\infty} t^2 e^{-t^2}\,dt = \frac{\sigma^2}{2}. \tag{232}$$

Use (i) to conclude $\mathrm{Var}(X) = \sigma^2$.

Proposition 5.8 (Linear transform)**.** *Let X be a RV with PDF f_X. Fix constants $a,b \in \mathbb{R}$ with $a > 0$, and define a new RV $Y = aX + b$. Then*

$$f_{aX+b}(y) = \frac{1}{|a|}f_X((y-b)/a). \tag{233}$$

PROOF. First suppose $a > 0$. Then

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(aX+b \le y) = \mathbb{P}(X \le (y-b)/a) = \int_{-\infty}^{(y-b)/a} f_X(t)\,dt. \tag{234}$$

By differentiating the last integral by y, we get

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{1}{a}f_X((y-b)/a). \tag{235}$$

For $a < 0$, a similar calculation shows

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(aX+b \le y) = \mathbb{P}(X \ge (y-b)/a) = \int_{(y-b)/a}^{\infty} f_X(t)\,dt, \tag{236}$$

so we get

$$f_Y(y) = \frac{dF_Y(y)}{dy} = -\frac{1}{a}f_X((y-b)/a). \tag{237}$$

This shows the assertion. □

Example 5.9 (Linear transform of normal RV is normal)**.** Let $X \sim N(\mu, \sigma^2)$ and fix constants $a, b \in \mathbb{R}$ with $a \neq 0$. Define a new RV $Y = aX + b$. Then since

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{238}$$

by Proposition 1.4, we have

$$f_Y(y) = \frac{1}{\sqrt{2\pi(a\sigma)^2}} \exp\left(-\frac{(y-b-a\mu)^2}{2(a\sigma)^2}\right). \tag{239}$$

Notice that this is the PDF of a normal RV with mean $a\mu + b$ and variance $(a\sigma)^2$. In particular, if we take $a = 1/\sigma$ and $b = \mu/\sigma$, then $Y = (X-\mu)/\sigma \sim N(0,1)$, the standard normal RV. This is called *standardization* of normal RV.

Example 5.10 (Using the standard normal table)**.** Let $X \sim N(3,4)$. We are going to compute the following probabilities

$$\mathbb{P}(X \geq 5) \qquad \text{and} \qquad \mathbb{P}(X^2 + 1 \geq 5) \tag{240}$$

using the standard normal table in Table 1. Let $Z = (X-3)/2$. Then by Example 1.5, we have $Z \sim N(0,1)$. Our strategy is to write these probabilities only in terms of the probabilities of the form $\mathbb{P}(0 \leq Z \leq a)$.

First note that

$$\mathbb{P}(X \geq 5) = \mathbb{P}\left(\frac{X-3}{2} \geq 1\right) = \mathbb{P}(Z \geq 1). \tag{241}$$

Then since the PDF of standard normal RV is symmetric about the y-axis, we can write

$$\mathbb{P}(Z \geq 1) = \mathbb{P}(Z \geq 0) - \mathbb{P}(0 \leq Z \leq 1) = 1/2 - \mathbb{P}(0 \leq Z \leq 1). \tag{242}$$

According to the standard normal table in Table 1, we have

$$\mathbb{P}(0 \leq Z \leq 1) = 0.3413. \tag{243}$$

This gives

$$\mathbb{P}(X \geq 5) = 0.5 - 0.3413 = 0.1587. \tag{244}$$

For the second example, note that

$$\mathbb{P}(X^2 + 1 \geq 5) = \mathbb{P}(X^2 \geq 4) = \mathbb{P}(X \leq -2) + \mathbb{P}(X \geq 2). \tag{245}$$

Then using the symmetry of the PDF of Z,

$$\mathbb{P}(X \leq -2) = \mathbb{P}(Z \leq -5/2) = \mathbb{P}(Z \geq 5/2) = 0.5 - \mathbb{P}(0 \leq Z \leq 5/2). \tag{246}$$

Since $\mathbb{P}(0 \leq Z \leq 5/2) = 0.4938$ from Table 1, we have

$$\mathbb{P}(X \leq -2) = 0.0062. \tag{247}$$

On the other hand,

$$\mathbb{P}(X \geq 2) = \mathbb{P}(Z \geq -1/2) = \mathbb{P}(-1/2 \leq Z \leq 0) + \mathbb{P}(0 \leq Z) = \mathbb{P}(0 \leq Z \leq 1/2) + 0.5. \tag{248}$$

Since $\mathbb{P}(0 \leq Z \leq 1/2) = 0.1915$ from Table 1, we have

$$\mathbb{P}(X \geq 2) = 0.6915. \tag{249}$$

This yields

$$\mathbb{P}(X^2 + 1 \geq 5) = 0.0062 + 0.6915 = 0.6977. \tag{250}$$

▲

Proposition 5.11 (Sum of ind. normal RVs is normal)**.** *Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ be independent normal RVs. Then*

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2). \tag{251}$$

PROOF. Details omitted. One can use the convolution formula or moment generating functions, which are subjects in MATH 170B. □

Exercise 5.12. Compute the following probabilities using the standard normal table in Table 1.
(i) $\mathbb{P}(-1 \le X \le 2)$ where $X \sim N(0, 3^2)$.
(ii) $\mathbb{P}(X^2 + X - 1 \ge 0)$ where $X \sim N(1, 1)$.
(iii) $\mathbb{P}(\exp(2X + 2Y) - 3\exp(X + Y) + 2 \le 0)$ where $X \sim N(0, 1)$ and $Y \sim N(-2, 3)$.

CHAPTER 3

Joint distributions and conditioning

In this chapter, we learn how to handle multiple random variables at the same time. In the most general situation, the dependence between random variables are described by their joint distribution. We then learn how to handle 'one randomness at a time' by using conditioning.

1. Joint probability mass functions

Let X and Y be discrete RVs defined on sample spaces Ω_1 and Ω_2, which we think of the outcome of two random experiments. We can think of the pair (X,Y) of the outcomes as a single observable that describes the two experiments jointly. This object is called a *random vector*, being a vector of random variables. Namely, (X,Y) is a vector-valued function $\Omega_1 \times \Omega_2 \to \mathbb{R}^2$ defined on the joint sample space $\Omega_1 \times \Omega_2$. As for the PDF of a random variable, which is the probability that a RV takes a particular value, we can also ask the probability that the random vector (X,Y) takes a particular vector (x,y). This is called the *joint PMF* of X and Y:

$$f_{X,Y}(x,y) = \mathbb{P}((X,Y)=(x,y)) = \mathbb{P}(X=x,\ Y=y). \tag{252}$$

Given a joint PMF of X and Y, we can determine the PMFs of X and Y simply by 'integrating out' the other variable. Namely, note that for any $x \in \mathbb{R}$,

$$\mathbb{P}(X=x) = \sum_{y \in Y[\Omega_2]} \mathbb{P}(X=x,\ Y=y) = \sum_{y \in Y[\Omega_2]} f_{X,Y}(x,y), \tag{253}$$

where $Y[\Omega_2]$ denotes the set of all possible values of Y. Hence, we sum $f_{X,Y}(x,y)$ over all y to get $f_X(x)$; likewise, we sum $f_{X,Y}(x,y)$ over all x to get $f_Y(y)$. We also call the PMFs of each X and Y the *marginal PMFs* of $f_{X,Y}$.

Example 1.1. Consider rolling two four-sided dice where each die has has possible outcomes from $\{0,1,2,3\}$. Let $\Omega = \{0,1,2,3\}^2$ be the joint sample space and let (X,Y) denote the random vector describing the outcome. Suppose the two dice somehow affect each other according to the join PMF on Ω, which is depicted in Figure 1.

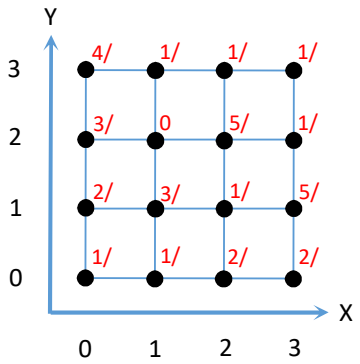

FIGURE 1. Joint PMF of (X,Y) on Ω shown in red. Common denominator of 33 is omitted in the figure.

To compute the PMF of X, note that

$$\mathbb{P}(X=0)=\sum_{y=0}^{3}\mathbb{P}(X=0,\,Y=y)=\frac{1}{33}(1+2+3+4)=\frac{10}{33} \tag{254}$$

$$\mathbb{P}(X=1)=\sum_{y=0}^{3}\mathbb{P}(X=1,\,Y=y)=\frac{1}{33}(1+3+0+1)=\frac{5}{33} \tag{255}$$

$$\mathbb{P}(X=2)=\sum_{y=0}^{3}\mathbb{P}(X=2,\,Y=y)=\frac{1}{33}(2+1+5+1)=\frac{9}{33} \tag{256}$$

$$\mathbb{P}(X=3)=\sum_{y=0}^{3}\mathbb{P}(X=3,\,Y=y)=\frac{1}{33}(2+5+1+1)=\frac{9}{33}. \tag{257}$$

We can compute the PMF of Y similarly. ▲

On the other hand, given only the PMFs of X and Y, we cannot determine the joint PMF for (X,Y), since we do not know how the two RVs depends on each other.

Example 1.2. Let (X,Y) be the outcome of a single roll of two coins. Let $\Omega=\{0,1\}$ be the sample space of each coin. Suppose we have the following PMFs of the two coins:

$$\mathbb{P}(X=0)=1/3,\quad \mathbb{P}(X=1)=2/2, \tag{258}$$

$$\mathbb{P}(Y=0)=\mathbb{P}(Y=1)=1/2. \tag{259}$$

What are the possible joint PMFs for (X,Y) satisfying these constraints? Are there more than one such joint PMFs?

Note that we can describe the joint PMF of X and Y as a 2×2 table, where the rows and columns represent possible values of X and Y, respectively. Namely, let $P=(p_{ij})$ denote the 2×2 matrix where $p_{ij}=\mathbb{P}(X=i,\,Y=j)$ for each $(i,j)\in\{0,1\}^2$. Then

$$p_{00}+p_{01}=\mathbb{P}(X=0)=1/2,\qquad p_{10}+p_{11}=\mathbb{P}(X=1)=1/2, \tag{260}$$

$$p_{00}+p_{10}=\mathbb{P}(Y=0)=1/3,\qquad p_{10}+p_{11}=\mathbb{P}(Y=1)=2/3. \tag{261}$$

Namely, P is a 2×2 matrix with fixed row and column sums, where the total sum equals 1. This is called a *contingency table* in statistics, and *doubly stochastic matrix* in probability. The row and column sums are also called as *margins*.

	$x=0$	$x=1$	
$y=0$	p	$\frac{1}{3}-p$	$1/3$
$y=1$	$\frac{1}{2}-p$	$\frac{1}{6}+p$	$2/3$
	$1/2$	$1/2$	1

FIGURE 2. Population contingency table for joint PMF of (X,Y) with prescribed marginal PMFs.

If we let $p=p_{00}$, then we find P is given as in Figure 2 above. Hence there are infinitely many joint PMFs of X and Y satisfying the prescribed PMF condition, one for each value of $p\in[0,1]$. ▲

Exercise 1.3. Let (X,Y) be the outcome of a single roll of a coin and a three-sided die, where the joint sample space is given by $\Omega=\{0,1\}\times\{0,1,2\}$. Suppose we have marginal PMFs

$$\mathbb{P}(X=0)=1/3,\quad \mathbb{P}(X=1)=2/3, \tag{262}$$

$$\mathbb{P}(Y=0)=\mathbb{P}(Y=1)=\mathbb{P}(Y=2)=1/3. \tag{263}$$

Let $P=(p_{ij})$ denote the 2×3 contingency table describing the joint PMF of X and Y.

(i) Let $p_{00}=p$ and $p_{01}=q$. Express all the other four entries in P in terms of p and q.

(ii) Draw the region $\mathscr{D}=\{(p,q)\,|\,p+q\le 1,\,p\ge 0,q\ge 0\}$ in the coordinate plane. Show that there is a 1-1 correspondence between the points in $\mathscr{D}$ and a joint PMF of X and Y.

Proposition 1.4. *Let X,Y be discrete RVs. Then they are independent if and only if there are functions f and g such that*

$$\mathbb{P}(X=x,\,Y=y)=f(x)g(y) \tag{264}$$

for all possible values (x,y) for (X,Y).

PROOF. If X and Y are independent, then

$$\mathbb{P}(X=x,\,Y=y)=\mathbb{P}(X=x)\mathbb{P}(Y=y)=f_X(x)f_Y(y). \tag{265}$$

Hence the conclusion holds. Conversely, suppose (264) holds for some functions f and g. Then

$$\mathbb{P}(X=x)=\sum_y \mathbb{P}(X=x,\,Y=y)=\sum_y f(x)g(y)=f(x)\sum_y g(y). \tag{266}$$

Since sum of the probabilities $\mathbb{P}(X=x)$ is 1, we have

$$\left(\sum_x f(x)\right)\left(\sum_y g(y)\right)=1. \tag{267}$$

Hence if we define $\bar{f}(x)=f(x)/\sum_x f(x)$ and $\bar{g}(y)=\sum_y g(y)$, then

$$\mathbb{P}(X=x)=f(x)\sum_y g(y)=\frac{f(x)\sum_y g(y)}{\left(\sum_x f(x)\right)\left(\sum_y g(y)\right)}=\bar{f}(x)\sum_y \bar{g}(y)=\bar{f}(x). \tag{268}$$

Similarly, we have $\mathbb{P}(Y=y)=\bar{g}(y)$. This yields

$$\mathbb{P}(X=x,\,Y=y)=f(x)g(y)=\frac{f(x)g(y)}{\left(\sum_x f(x)\right)\left(\sum_y g(y)\right)}=\bar{f}(x)\bar{g}(y)=\mathbb{P}(X=x)\mathbb{P}(Y=y). \tag{269}$$

Since (x,y) was arbitrary, this shows X and Y are independent. □

2. Joint probability density functions

Let X and Y be continuous RVs defined on sample spaces Ω_1 and Ω_2. We say X and Y are *jointly continuous* if the random vector (X,Y) has a function $f_{X,Y}:\mathbb{R}^2\to[0,\infty)$ such that for each subset $\mathscr{A}\subseteq\mathbb{R}^2$,

$$\mathbb{P}((X,Y)\in\mathscr{A})=\int\int_{(x,y)\in\mathscr{A}} f_{X,Y}(x,y)\,dx\,dy. \tag{270}$$

For instance, if $\mathscr{A}$ is a rectangle of the form $\mathscr{A}=[a,b]\times[c,d]$, then by Fubini's theorem, (270) reads

$$\mathbb{P}(a\le X\le b,\,c\le Y\le d)=\int_a^b\int_c^d f_{X,Y}(x,y)\,dx\,dy. \tag{271}$$

Such a function $f_{X,Y}$ is called the *joint PDF* of X and Y.

As for the joint PMFs, we can determine the PDF of X and Y simply by 'integrating out' the other variable. We also call the PMFs of each X and Y the *marginal PMFs* of $f_{X,Y}$.

Proposition 2.1. *Let X and Y be continuous RVs with joint PDF $f_{X,Y}$. Then for any $x,y\in\mathbb{R}$, we have*

$$f_X(x)=\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy,\qquad f_Y(y)=\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx. \tag{272}$$

PROOF. Note that for any interval $[a,b] \subseteq \mathbb{R}$, by definition of joint PDF and Fubini's theorem,

$$\mathbb{P}(a \le X \le b) = \int_a^b \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy \tag{273}$$

$$= \int_a^b \left(\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \right) dx. \tag{274}$$

Then by the definition of PDF of a continuous RV, it follows that the PDF f_X of X is given the the first equation in (272). A Similar argument applies to f_Y as well. □

Example 2.2. Let X and Y be continuous RVs with joint PDF given by

$$f_{X,Y}(x,y) = \lambda^2 e^{-\lambda(x+y)} \mathbf{1}(x,y \ge 0). \tag{275}$$

Then for each $x \in \mathbb{R}$,

$$f_X(x) = \int_{-\infty}^{\infty} \lambda^2 e^{-\lambda(x+y)} \mathbf{1}(x,y \ge 0)\,dy \tag{276}$$

$$= \lambda e^{-\lambda x} \mathbf{1}(x \ge 0) \int_{-\infty}^{\infty} \lambda e^{-\lambda y} \mathbf{1}(y \ge 0)\,dy \tag{277}$$

$$= \lambda e^{-x} \mathbf{1}(x \ge 0). \tag{278}$$

Note that for the last equality, we have recognized the function $y \mapsto \lambda e^{-\lambda y} \mathbf{1}(y \ge 0)$ as the PDF of $\mathrm{Exp}(\lambda)$ RV. This shows

$$f_X(x) = \lambda e^{-\lambda x} \mathbf{1}(x \ge 0). \tag{279}$$

Hence $X \sim \mathrm{Exp}(\lambda)$. Similarly, we obtain $Y \sim \mathrm{Exp}(\lambda)$. ▲

Example 2.3. Let (X,Y) be uniformly distributed over the unit disk $\Omega = \{(x,y) \mid x^2 + y^2 \le 1\}$. Then we have

$$1 = \int\int_{\Omega} f_{X,Y}(x,y)\,dx\,dy = C \int\int_{\Omega} 1\,dx\,dy \tag{280}$$

for some constant $C > 0$, and

$$\int\int_{\Omega} 1\,dx\,dy = \mathrm{Area}(\Omega) = \pi. \tag{281}$$

This yields

$$f_{X,Y}(x,y) = \frac{1}{\pi} \mathbf{1}(x^2 + y^2 \le 1). \tag{282}$$

Now in order to compute the marginal distribution of X, we note that for each $-1 \le x \le 1$,

$$f_X(x) = \int_{-\infty}^{\infty} \frac{1}{\pi} \mathbf{1}(x^2 + y^2 \le 1)\,dy \tag{283}$$

$$= \int_{-\infty}^{\infty} \frac{1}{\pi} \mathbf{1}(-\sqrt{1-x^2} \le y \le \sqrt{1-x^2})\,dy \tag{284}$$

$$= \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi}\,dy = \frac{2\sqrt{1-x^2}}{\pi}. \tag{285}$$

Hence we conclude

$$f_X(x) = \frac{2\sqrt{1-x^2}}{\pi} \mathbf{1}(|x| \le 1). \tag{286}$$

By symmetry, Y has the same PDF as X. ▲

Proposition 2.4. *Let X, Y be continuous RVs with a joint PDF $f_{X,Y}$. Then they are independent if and only if there are functions f and g such that*

$$f_{X,Y}(x,y) = f(x)g(y) \tag{287}$$

for all $x, y \in \mathbb{R}$.

PROOF. Similar to the proof of Proposition 1.4. □

Example 2.5. If (X,Y) has joint PDF in (275) in Exercise 2.2, then by Proposition 2.4, X and Y are independent. On the other hand, if their joint PDF is as in (282), which is not a factor form, then X and Y are not independent by Proposition 2.4. ▲

3. Conditional expectation

Let X, Y be discrete RVs. Recall that the expectation $\mathbb{E}(X)$ is the 'best guess' on the value of X when we do not have any prior knowledge on X. But suppose we have observed that some possibly related RV Y takes value y. What should be our best guess on X, leveraging this added information? This is called the *conditional expectation of X given $Y = y$*, which is defined by

$$\mathbb{E}[X|Y=y] = \sum_x x\,\mathbb{P}(X=x|Y=y). \tag{288}$$

This best guess on X given $Y = y$, of course, depends on y. So it is a function in y. Now if we do not know what value Y might take, then we omit y and $\mathbb{E}[X|Y]$ becomes a RV, which is called the *conditional expectation of X given Y*.

Example 3.1. Suppose we have a biased coin whose probability of heads is itself random and is distributed as $Y \sim \text{Uniform}([0,1])$. Let's flip this coin n times and let X be the total number of heads. Given that $Y = y \in [0,1]$, we know that X follows Binomial(n,y) (in this case we write $X|U \sim \text{Binomial}(n,Y)$). So $\mathbb{E}[X|Y=y] = ny$. Hence as a random variable, $\mathbb{E}[X|Y] = nY \sim \text{Uniform}([0,n])$. So the expectation of $\mathbb{E}[X|Y]$ is the mean of Uniform$([0,n])$, which is $n/2$. This value should be the true expectation of X. ▲

The above example suggests that if we first compute the conditional expectation of X given $Y = y$, and then average this value over all choice of y, then we should get the actual expectation of X. Justification of this observation is based on the following fact

$$\mathbb{P}(Y=y\,|\,X=x)\mathbb{P}(X=x) = \mathbb{P}(X=x,\,Y=y) = \mathbb{P}(X=x\,|\,Y=y)\mathbb{P}(Y=y). \tag{289}$$

That is, if we are interested in the event that $(X,Y) = (x,y)$, then we can either first observe the value of X and then Y, or the other way around.

Proposition 3.2 (Iterated expectation)**.** *Let X, Y be discrete RVs. Then $\mathbb{E}(X) = \mathbb{E}[\mathbb{E}[X|Y]]$.*

PROOF. We are going to write the iterated expectation $\mathbb{E}[\mathbb{E}[X|Y]]$ as a double sum and swap the order of summation (Fubini's theorem, as always).

$$\mathbb{E}[\mathbb{E}[X|Y]] = \sum_y \mathbb{E}[X|Y=y]\mathbb{P}(Y=y) \tag{290}$$

$$= \sum_y \left(\sum_x x\,\mathbb{P}(X=x|Y=y)\right)\mathbb{P}(Y=y) \tag{291}$$

$$= \sum_y \sum_x x\,\mathbb{P}(X=x|Y=y)\mathbb{P}(Y=y) \tag{292}$$

$$= \sum_y \sum_x x\,\mathbb{P}(X=x,\,Y=y) \tag{293}$$

$$= \sum_x \sum_y x\,\mathbb{P}(Y=y|X=x)\mathbb{P}(X=x) \tag{294}$$

$$= \sum_x x \left(\sum_y \mathbb{P}(Y = y | X = x) \right) \mathbb{P}(X = x) \tag{295}$$

$$= \sum_x x \, \mathbb{P}(X = x) = \mathbb{E}(X). \tag{296}$$

□

Remark 3.3. Here is an intuitive reason why the iterated expectation works. Suppose you want to make the best guess $\mathbb{E}(X)$. Pretending you know Y, you can imporove your guess to be $E(X \mid Y)$. Then you admit that you didn't know anything about Y and average over all values of Y. The result is $\mathbb{E}[\mathbb{E}[X \mid Y]]$, and this should be the same best guess on X when we don't know anything about Y.

All our discussions above hold for continuous RVs as well: We simply replace the sum by integral and PMF by PDF. To summarize how we compute the iterated expectations when we condition on discrete and continuous RV:

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \begin{cases} \sum_y \mathbb{E}[X \mid Y = y] \mathbb{P}(Y = y) & \text{if } Y \text{ is discrete} \\ \int_{-\infty}^{\infty} \mathbb{E}[X \mid Y = y] f_Y(y) \, dy & \text{if } Y \text{ is continuous.} \end{cases} \tag{297}$$

Exercise 3.4 (Iterated expectation for probability)**.** Let X, Y be RVs.

(i) For any $x \in \mathbb{R}$, show that $\mathbb{P}(X \le x) = \mathbb{E}[\mathbf{1}(X \le x)]$.

(ii) By using iterated expectation, show that

$$\mathbb{P}(X \le x) = \mathbb{E}[\mathbb{P}(X \le x \mid Y)], \tag{298}$$

where the expectation is taken over for all possible values of Y.

Example 3.5 (Example 3.1 revisited)**.** Let $Y \sim \text{Uniform}([0,1])$ and $X \sim \text{Binomial}(n, Y)$. Then $X|Y = y \sim \text{Binomial}(n, y)$ so $\mathbb{E}[X|Y = y] = ny$. Hence

$$\mathbb{E}[X] = \int_0^1 \mathbb{E}[X \mid Y = y] f_Y(y) \, dy = \int_0^1 ny \, dy = n/2. \tag{299}$$

▲

Example 3.6. Let $X_1 \sim \text{Exp}(\lambda_1)$ and $X_2 \sim \text{Exp}(\lambda_2)$ be independent exponential RVs. We will show that

$$\mathbb{P}(X_1 < X_2) = \frac{\lambda_1}{\lambda_1 + \lambda_2} \tag{300}$$

using the iterated expectation. Using iterated expectation for probability,

$$\mathbb{P}(X_1 < X_2) = \int_0^\infty \mathbb{P}(X_1 < X_2 \mid X_1 = x_1) \lambda_1 e^{-\lambda_1 x_1} \, dx_1 \tag{301}$$

$$= \int_0^\infty \mathbb{P}(X_2 > x_1) \lambda_1 e^{-\lambda_1 x_1} \, dx_1 \tag{302}$$

$$= \lambda_1 \int_0^\infty e^{-\lambda_2 x_1} e^{-\lambda_1 x_1} \, dx_1 \tag{303}$$

$$= \lambda_1 \int_0^\infty e^{-(\lambda_1 + \lambda_2) x_1} \, dx_1 = \frac{\lambda_1}{\lambda_1 + \lambda_2}. \tag{304}$$

▲

Exercise 3.7. Consider a post office with two clerks. Three people, A, B, and C, enter simultaneously. A and B go directly to the clerks, and C waits until either A or B leaves, and then she starts getting serviced. Let X_A be the time that A spends at a register, and define X_B and X_B similarly. Compute the probability $\mathbb{P}(X_A > X_B + X_C)$ that A leaves the post office after B and C do so in the following scenarios:

(a) The service time for each clerk is exactly (nonrandom) ten minutes.

(b) The service times are i, independently with probability $1/3$ for $i \in \{1, 2, 3\}$.

(c) The service times are independent Exp(λ) RVs. You may use the fact that the PDF of $X_B + X_C$ is

$$f_{X_B+Y_C}(z) = \lambda^2 z e^{-\lambda z} \mathbf{1}(z \geq 0). \tag{305}$$

Exercise 3.8. Suppose we have a stick of length L. Break it into two pieces at a uniformly chosen point and let X_1 be the length of the longer piece. Break this longer piece into two pieces at a uniformly chosen point and let X_2 be the length of the longer one. Define $X_3, X_4, \cdots$ in a similar way.

(i) Let $U \sim$ Uniform($[0, L]$). Show that X_1 takes values from $[L/2, L]$, and that $X_1 = \max(U, L-U)$.
(ii) From (i), deduce that for any $L/2 \leq x \leq L$, we have

$$\mathbb{P}(X_1 \geq x) = \mathbb{P}(U \geq x \text{ or } L-U \geq x) = \mathbb{P}(U \geq x) + \mathbb{P}(U \leq L-x) = \frac{2(L-x)}{L}. \tag{306}$$

Conclude that $X_1 \sim$ Uniform($[L/2, L]$). What is $\mathbb{E}[X_1]$?
(iii) Show that $X_2 \sim$ Uniform($[x_1/2, x_1]$) conditional on $X_1 = x_1$. Using iterated expectation, show that $\mathbb{E}[X_2] = (3/4)^2 L$.
(iv) In general, show that $X_{n+1} \mid X_n \sim$ Uniform($[X_n/2, X_n]$). Conclude that $\mathbb{E}[X_n] = (3/4)^n L$.

4. Conditional expectation as an estimator

We introduced the conditional expectation $\mathbb{E}[X \mid Y = y]$ as the best guess on X given that $Y = y$. Such a 'guess' on a RV is called an *estimator*. Let's first take a look at two extremal cases, where observing Y gives absolutely no information on X or gives everything.

Example 4.1. Let X and Y be independent discrete RVs. Then knowing the value of Y should not yield any information on X. In other words, given that $Y = y$, the best guess of X should still be $\mathbb{E}(X)$. Indeed,

$$\mathbb{E}(X|Y=y) = \sum_{x=0}^{n} x\,\mathbb{P}(X=x \mid Y=y) = \sum_{x=0}^{n} x\,\mathbb{P}(X=x) = \mathbb{E}(X). \tag{307}$$

On the other hand, given that $X = x$, the best guess on X is just x, since the RV X has been revealed and there is no further randomness. In other words,

$$\mathbb{E}(X|X=x) = \sum_{z=0}^{n} z\,\mathbb{P}(X=z \mid X=x) = \sum_{x=0}^{n} x\mathbf{1}(z=x) = x. \tag{308}$$

▲

Exercise 4.2. Let X, Y be discrete RVs. Show that for any function $g : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[Xg(Y) \mid Y] = g(Y)\mathbb{E}[X \mid Y]. \tag{309}$$

We now observe some general properties of the conditional expectation as an estimator.

Exercise 4.3. Let X, Y be RVs and denote $\hat{X} = \mathbb{E}[X|Y]$, meaning that $\hat{X}$ is an estimator of X given Y. Let $\tilde{X} = \hat{X} - X$ be the *estimation error*.

(i) Show that $\hat{X}$ is an *unbiased* estimator of X, that is, $\mathbb{E}(\hat{X}) = \mathbb{E}(X)$.
(ii) Show that $\mathbb{E}[\hat{X}|Y] = \hat{X}$. Hence knowing Y does not improve our current best guess $\hat{X}$.
(iii) Show that $\mathbb{E}[\tilde{X}] = 0$.
(iv) Show that $\mathrm{Cov}(\hat{X}, \tilde{X}) = 0$. Conclude that

$$\mathrm{Var}(X) = \mathrm{Var}(\hat{X}) + \mathrm{Var}(\tilde{X}). \tag{310}$$

5. Conditional variance

As we have defined conditional expectation, we could define the variance of a RV X given that anther RV Y takes a particular value. Recall that the (unconditioned) variance of X is defined by

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]. \tag{311}$$

Note that there are two places where we take expectation. Given Y, we should improve both expectations so the *conditional variance of X given Y is defined by*

$$\mathrm{Var}(X \mid Y) = \mathbb{E}[(X - \mathbb{E}[X \mid Y])^2 \mid Y]. \tag{312}$$

Proposition 5.1. *Let X and Y be RVs. Then*

$$\mathrm{Var}(X \mid Y) = \mathbb{E}[X^2 \mid Y] - \mathbb{E}[X \mid Y]^2. \tag{313}$$

PROOF. Using linearity of conditional expectation and the fact that $\mathbb{E}[X \mid Y]$ is not random given Y,

$$\begin{aligned}
\mathrm{Var}(X \mid Y) &= \mathbb{E}[X^2 - 2X\mathbb{E}[X \mid Y] + \mathbb{E}[X \mid Y]^2 \mid Y] && (314)\\
&= \mathbb{E}[X^2 \mid Y] - \mathbb{E}[2X\mathbb{E}[X \mid Y] \mid Y] + \mathbb{E}[\mathbb{E}[X \mid Y]^2 \mid Y] && (315)\\
&= \mathbb{E}[X^2 \mid Y] - \mathbb{E}[X \mid Y]\mathbb{E}[2X \mid Y] + \mathbb{E}[X \mid Y]^2\mathbb{E}[1 \mid Y] && (316)\\
&= \mathbb{E}[X^2 \mid Y] - 2\mathbb{E}[X \mid Y]^2 + \mathbb{E}[X \mid Y]^2 && (317)\\
&= \mathbb{E}[X^2 \mid Y] - \mathbb{E}[X \mid Y]^2. && (318)
\end{aligned}$$

□

The following exercise explains in what sense the conditional expectation $\mathbb{E}[X \mid Y]$ is the best guess on X given Y, and that the minimum possible mean squared error is exactly the conditional variance $\mathrm{Var}(X \mid Y)$.

Exercise 5.2. Let X, Y be RVs. For any function $g : \mathbb{R} \to \mathbb{R}$, consider $g(Y)$ as an estimator of X. Let $\mathbb{E}_Y[(X - g(Y))^2 \mid Y]$ be the *mean squared error*.

(i) Show that

$$\begin{aligned}
\mathbb{E}_Y[(X - g(Y))^2 \mid Y] &= \mathbb{E}_Y[X^2 \mid Y] - 2g(Y)\mathbb{E}_Y[X \mid Y] + g(Y)^2 && (319)\\
&= (g(Y) - \mathbb{E}_Y(X \mid Y))^2 + \mathbb{E}_Y[X^2 \mid Y] - \mathbb{E}_Y[X \mid Y]^2 && (320)\\
&= (g(Y) - \mathbb{E}_Y(X \mid Y))^2 + \mathrm{Var}(X \mid Y). && (321)
\end{aligned}$$

(ii) Conclude that the mean squared error is minimized when $g(Y) = \mathbb{E}_Y[X \mid Y]$ and the global minimum is $\mathrm{Var}(X \mid Y)$.

Next, we study how we can decompose the variance of X by conditioning on Y.

Proposition 5.3 (Law of total variance)**.** *Let X and Y be RVs. Then*

$$\mathrm{Var}(X) = \mathbb{E}(\mathrm{Var}(X \mid Y)) + \mathrm{Var}(\mathbb{E}[X \mid Y]). \tag{322}$$

PROOF. Using previous result, iterated expectation, and linearity of expectation, we have

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 && (323)\\
&= \mathbb{E}_Y(\mathbb{E}(X^2|Y)) - (\mathbb{E}_Y(\mathbb{E}(X|Y)))^2 && (324)\\
&= \mathbb{E}_Y(\mathrm{Var}(X|Y) + (\mathbb{E}(X|Y))^2) - (\mathbb{E}_Y(\mathbb{E}(X|Y)))^2 && (325)\\
&= \mathbb{E}_Y(\mathrm{Var}(X|Y)) + \left[\mathbb{E}_Y(\mathbb{E}(X|Y))^2) - (\mathbb{E}_Y(\mathbb{E}(X|Y)))^2\right] && (326)\\
&= \mathbb{E}_Y(\mathrm{Var}(X|Y)) + \mathrm{Var}_Y(\mathbb{E}(X|Y)). && (327)
\end{aligned}$$

□

Here is a handwavy explanation on why the above is true. Given Y, we should measure the fluctuation of $X \mid Y$ from the conditional expectation $\mathbb{E}[X \mid Y]$, and this is measured as $\mathrm{Var}(X \mid Y)$. Since we don't know Y, we average over all Y, giving $\mathbb{E}(\mathrm{Var}(X \mid Y))$. But the reference point $\mathbb{E}[X \mid Y]$ itself varies with Y, so we should also measure its own fluctuation by $\mathrm{Var}(\mathbb{E}[X \mid Y])$. These fluctuations add up nicely like Pythagorian theorem because $\mathbb{E}[X \mid Y]$ is an optimal estimator so that these two fluctuations are 'orthogonal'.

Exercise 5.4. Let X, Y be RVs. Write $\bar{X} = \mathbb{E}[X \mid Y]$ and $\tilde{X} = X - \mathbb{E}[X \mid Y]$ so that $X = \bar{X} + \tilde{X}$. Here $\bar{X}$ is the estimate of X given Y and $\tilde{X}$ is the estimation error.

(i) Using Exercise 4.3 (iii) and iterated expectation, show that

$$\mathbb{E}[\tilde{X}^2] = \mathrm{Var}(\mathbb{E}[X \mid Y]). \tag{328}$$

(ii) Using Exercise 4.3 (iv), conclude that

$$\mathrm{Var}(X) = \mathbb{E}(\mathrm{Var}(X \mid Y)) + \mathrm{Var}(\mathbb{E}[X \mid Y]). \tag{329}$$

Example 5.5. Let $Y \sim \mathrm{Uniform}([0,1])$ and $X \sim \mathrm{Binomial}(n, Y)$. Since $X \mid Y = y \sim \mathrm{Binomial}(n, y)$, we have $\mathbb{E}[X \mid Y = y] = ny$ and $\mathrm{Var}(X \mid Y = y) = ny(1-y)$. Also, since $Y \sim \mathrm{Uniform}([0,1])$, we have

$$\mathrm{Var}(\mathbb{E}[X \mid Y]) = \mathrm{Var}(nY) = \frac{n^2}{12}. \tag{330}$$

So by iterated expectation, we get

$$\mathbb{E}(X) = \mathbb{E}_Y(\mathbb{E}[X \mid Y]) = \int_0^1 ny\,dy = \frac{n}{2}. \tag{331}$$

On the other hand, by law of total variance,

$$\mathrm{Var}(X) = \mathbb{E}(\mathrm{Var}(X \mid Y)) + \mathrm{Var}(\mathbb{E}(X \mid Y)) \tag{332}$$

$$= \int_0^1 ny(1-y)\,dy + \mathrm{Var}(nY) \tag{333}$$

$$= n\left[\frac{y^2}{2} - \frac{y^3}{3}\right]_0^1 + \frac{n^2}{12} \tag{334}$$

$$= \frac{n^2}{12} + \frac{n}{6}. \tag{335}$$

▲

In fact, we can figure out the entire distribution of the binomial variable with uniform rate using conditioning, not just its mean and variance (credit to our TA Daniel).

Exercise 5.6. Let $Y \sim \mathrm{Uniform}([0,1])$ and $X \sim \mathrm{Binomial}(n, Y)$ as in Exercise 5.5.

(i) Use iterated expectation for probability to write

$$\mathbb{P}(X = k) = \binom{n}{k} \int_0^1 y^k (1-y)^{n-k}\,dy. \tag{336}$$

(ii) Write $A_{n,k} = \int_0^1 y^k (1-y)^{n-k}\,dy$. Use integration by parts and show that

$$A_{n,k} = \frac{k}{n-k+1} A_{n,k-1}. \tag{337}$$

for all $1 \le k \le n$. Conclude that for all $0 \le k \le n$,

$$A_{n,k} = \frac{1}{\binom{n}{k}} \frac{1}{n+1}. \tag{338}$$

(iii) Conclude that $X \sim \mathrm{Uniform}(\{0, 1, \cdots, n\})$.

Exercise 5.7 (Beta distribution)**.** A random varible X taking values from $[0,1]$ has Beta distribution of parameters α and β, which we denote by $\text{Beta}(\alpha,\beta)$, if it has PDF

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \tag{339}$$

where $\Gamma(z)$ is the Euler Gamma function defined by

$$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,dx. \tag{340}$$

(i) Use integration by parts to show the following recursion

$$\Gamma(z+1) = z\Gamma(z). \tag{341}$$

Deduce that $\Gamma(n) = (n-1)!$ for all integers $n \geq 1$.

(ii) Let $X \sim \text{Beta}(k+1, n-k+1)$. Use (i) to show that

$$f_X(x) = \frac{n!(n+1)}{k!(n-k)!} x^k(1-x)^{n-k} = \frac{x^k(1-x)^{n-k}}{1/\binom{n}{k}(n+1)}. \tag{342}$$

Use Exercise 5.6 to verify that the above function is indeed a PDF (i.e., it integrates to 1).

(iii)* Show that if $X \sim \text{Beta}(\alpha,\beta)$, then

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}, \quad \text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{343}$$

Exercise 5.8 (Exercise 3.8 continued)**.** Let $X_1, X_2, \cdots, X_n$ be as in Exercise 3.8.

(i) Show that $\text{Var}(X_1) = L^2/48$.

(ii) Show that $\text{Var}(X_2) = (7/12)\,\text{Var}(X_1) + (1/48)\mathbb{E}(X_1)^2$.

(iii) Show that $\text{Var}(X_{n+1}) = (7/12)\,\text{Var}(X_n) + (1/48)\mathbb{E}(X_n)^2$ for any $n \geq 1$.

(iv) Using Exercise 3.8, show the following recursion on variance holds:

$$\text{Var}(X_{n+1}) = \frac{7}{12}\text{Var}(X_n) + \frac{1}{48}\left(\frac{9}{16}\right)^n L^2. \tag{344}$$

Furthermore, compute $\text{Var}(X_2)$ and $\text{Var}(X_3)$.

(v)* Let $A_n = \left(\frac{16}{9}\right)^n \text{Var}(X_n)$. Show that A_n's satisfy

$$A_{n+1} + L^2 = \left(\frac{28}{27}\right)(A_n + L^2). \tag{345}$$

(vi)* Show that $A_n = \left(\frac{28}{27}\right)^{n-1}(A_1 + L^2) - L^2$ for all $n \geq 1$.

(vii)* Conclude that

$$\text{Var}(X_n) = \left[\left(\frac{7}{12}\right)^n - \left(\frac{9}{16}\right)^n\right] L^2. \tag{346}$$

6. Bayesian inference

We have discussed an elementary form of Bayesian inference in the discrete setting. In this section, we give a more general treatment in the continuous setting.

We begin by recalling the following basic observation, from which Bayes' theorem can be easily derived:

$$\mathbb{P}(B|A)\mathbb{P}(A) = \mathbb{P}(A\cap B) = \mathbb{P}(A|B)\mathbb{P}(B). \tag{347}$$

In case of the events involving continuous random variables taking a certain value, we interpret the above probabilities as probability densities. The following is a random variable version of Bayes' Theorem we have seen in Lecture note 1.

Theorem 6.1 (Bayes' Theorem)**.** *Let X and Θ be random variables. Then*

$$\mathbb{P}(\Theta=\theta\,|\,X=x)=\frac{\mathbb{P}(X=x\,|\,\Theta=\theta)\mathbb{P}(\Theta=\theta)}{\mathbb{P}(X=x)}. \tag{348}$$

PROOF. Follows from (347). □

Remark 6.2. Depending on whether X and Θ are discrete or continuous, we interpret the probabilities in the above theorem as PDF or PMF, accordingly.

Bayesian inference is usually carried out in the following procedure.

Bayesian inference:

(i) To explain an observable $\mathbf{x}$, we choose a *probabilistic model* $p(x\,|\,\theta)$, which is a probability distribution on the possible values x of $\mathbf{x}$, depending on a parameter θ.

(ii) Choose a probability distribution $\pi(\theta)$, called the *prior distribution*, that expresses our beliefs about a parameter θ to the best of our current knowledge.

(iii) After observing data $\mathscr{D}=\{x_1,\cdots,x_n\}$, we update our beliefs and compute the *posterior distribution* $p(\theta\,|\,\mathscr{D})$.

When we generate data $\mathscr{D}=\{x_1,\cdots,x_n\}$, we assume that each x_i is obtained by an independent sample X_i of the observable $\mathbf{x}$ from the true distribution of $\mathbf{x}$. According to Bayes' Theorem, we can compute the posterior distribution as

$$\mathbb{P}(\Theta=\theta\,|\,X_1=x_1,\cdots,X_n=x_n)=\frac{\mathbb{P}(X_1=x_1,\cdots,X_n=x_n\,|\,\Theta=\theta)\mathbb{P}(\Theta=\theta)}{\mathbb{P}(X_1=x_1,\cdots,X_n=x_n)}, \tag{349}$$

or in a more compact form,

$$p(\theta\,|\,x_1,\cdots,x_n)=\frac{p(x_1,\cdots,x_n\,|\,\theta)\pi(\theta)}{p(x_1,\cdots,x_n)}. \tag{350}$$

By a slight abuse of notation, we use lowercase p to denote either PMF or PDF, depending on the context.

The conditional probability $p(x_1,\cdots,x_n\,|\,\theta)$ is called the *likelyhood function*, which is the probability of obtaining independent data sample $\mathscr{D}=\{x_1,\cdots,x_n\}$ according to our probability model $p(x\,|\,\theta)$ assuming model parameter $\Theta=\theta$. By the independence between samples, the likelihood function factors into the product of marginal likelihood of each sample:

$$p(x_1,\cdots,x_n\,|\,\theta)=\mathbb{P}(X_1=x_1,\cdots,X_n=x_n\,|\,\Theta=\theta) \tag{351}$$

$$=\prod_{i=1}^{n}\mathbb{P}(X_i=x_i\,|\,\Theta=\theta)=\prod_{i=1}^{n}p(x_i\,|\,\theta). \tag{352}$$

On the other hand, the joint probability $\mathbb{P}(X_1=x_1,\cdots,X_n=x_n)$ of obtaining the data sample $\mathscr{D}=\{x_1,\cdots,x_n\}$ from our probability model can be computed by conditioning on the values of model parameter Θ:

$$\mathbb{P}(X_1=x_1,\cdots,X_n=x_n)=\mathbb{E}\left[\mathbb{P}(X_1=x_1,\cdots,X_n=x_n\,|\,\Theta)\,\middle|\,\Theta\right] \tag{353}$$

$$=\int_{-\infty}^{\infty}p(x_1,\cdots,x_n\,|\,\theta)\pi(\theta)\,d\theta. \tag{354}$$

Example 6.3 (Signal detection)**.** A binary signal $X\in\{-1,1\}$ is transmitted, and we are given that

$$\mathbb{P}(X=1)=p,\qquad \mathbb{P}(X=-1)=1-p \tag{355}$$

for some $p\in[0,1]$, as the prior distribution of X. There is a white noise in the channel so we receive a perturbed signal

$$Y=X+Z, \tag{356}$$

where $Z\sim N(0,1)$ is a standard normal variable.

Conditional on $Y = y$, what is the probability that the actual signal X is 1? In order to use Bayes theorem, first observe that $Y|X = x \sim N(x,1)$. Hence

$$f_{Y|X=x}(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-x)^2}{2}}. \tag{357}$$

So by Bayes Theorem,

$$\mathbb{P}(X = 1 \,|\, Y = y) = \frac{f_{Y|X=1}(y)\mathbb{P}(X=1)}{f_Y(y)} \tag{358}$$

$$= \frac{\frac{p}{\sqrt{2\pi}} e^{-\frac{(y-1)^2}{2}}}{\frac{p}{\sqrt{2\pi}} e^{-\frac{(y-1)^2}{2}} + \frac{1-p}{\sqrt{2\pi}} e^{-\frac{(y+1)^2}{2}}} \tag{359}$$

Thus we get

$$\mathbb{P}(X = 1 \,|\, Y = y) = \frac{pe^y}{pe^y + (1-p)e^{-y}}, \tag{360}$$

and similarly,

$$\mathbb{P}(X = -1 \,|\, Y = y) = \frac{(1-p)e^{-y}}{pe^y + (1-p)e^{-y}}, \tag{361}$$

This is our posterior distribution of X after observing $Y = y$. ▲

Example 6.4 (Bernoulli model and uniform prior)**.** Bob has a coin with unknown probability Θ of heads. Alice has no information whatsoever, so her prior distribution π for Θ is the uniform distribution Uniform$([0,1])$. Bob flips his coin independently n times, and let $X_1, \cdots, X_n$ be the outcome, X_i's are i.i.d. Bernoulli(Θ) variables. Let x_i be the observed value of X_i, and let $s_n = x_1 + \cdots + x_n$ be the number of heads in the n flips. Given data $\mathscr{D} = \{x_1, \cdots, x_n\}$, Alice wants to compute her posterior distribution on Θ.

The the likelyhood function is given by

$$p(x_1, \cdots, x_n \,|\, \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{x_i} \tag{362}$$

$$= \theta^{s_n}(1-\theta)^{n-s_n}, \tag{363}$$

where we denote $s_n = x_1 + \cdots + x_n$. Since $\pi \equiv 1$, the posterior distribution is given by

$$p(\theta \,|\, x_1, \cdots, x_n) = \frac{\theta^{s_n}(1-\theta)^{n-s_n}}{p(x_1, \cdots, x_n)}, \tag{364}$$

Note that

$$p(x_1, \cdots, x_n) = \int_0^1 \theta^{s_n}(1-\theta)^{n-s_n}\, d\theta = \frac{1}{\binom{n}{s_n}(n+1)}, \tag{365}$$

where the last equality follows from Exercise 5.6. Hence

$$p(\theta \,|\, x_1, \cdots, x_n) = \binom{n}{s_n}(n+1)\theta^{s_n}(1-\theta)^{n-s_n} \tag{366}$$

$$= \frac{(n+1)!}{s_n!(n-s_n)!}\theta^{(s_n+1)-1}(1-\theta)^{(n-s_n+1)-1}. \tag{367}$$

Hence, according to Exercise 5.7, we can write

$$\theta \,|\, x_1, \cdots, x_n \sim \text{Beta}(s_n + 1, n - s_n + 1). \tag{368}$$

▲

Exercise 6.5. Bob has a coin with unknown probability Θ of heads. Alice has the following Beta prior (See Exercise 5.7 for the definition of Beta distribution):

$$\pi = \text{Beta}(\alpha, \beta). \tag{369}$$

Suppose that Bob gives Alice the data $\mathscr{D}_n = \{x_1, \cdots, x_n\}$, which is the outcome of n independent coin flips. Denote $s_n = x_1 + \cdots + x_n$. Show that Alice's posterior distribution is $\text{Beta}(\alpha + S_n, \beta + n - S_n)$. Namely,

$$\Theta | \mathscr{D}_n \sim \text{Beta}(\alpha + S_n, \beta + n - S_n). \tag{370}$$

CHAPTER 4

Random variable as a function of another random variable

We have studied some of the fundamental RVs such as Bernoulli, Binomial, geometric and Poisson for discrete RVs and uniform, exponential, and normal for continuous RVs. Since different RVs can be used to model different situations, it is desirable to enlarge our vocabulary of RVs. A nice way to doing so is to compose a RV with a function to get a new RV.

1. Functions of one or two RVs

Suppose X is a RV with a known CDF. We may define a new random variable $Y = g(X)$ for a function $g : \mathbb{R} \to \mathbb{R}$. Can we derive CDF of Y and also its PDF (or PMF if X is discrete)? If we can solve the CDF of $Y = g(X)$, $\mathbb{P}(g(X) \le x)$, and recognize it as the CDF of some known RV, then we can identify what RV Y is.

Example 1.1 (Exponential from uniform)**.** Let $X \sim \text{Uniform}([0,1))$ and fix a constant $\lambda > 0$. Define a random variable $Y = -\frac{1}{\lambda}\log(1-X)$. Then $Y \sim \text{Exp}(\lambda)$. To see this, we calculate the CDF of Y as below:

$$\mathbb{P}(Y \le y) = \mathbb{P}\left(-\frac{1}{\lambda}\log(1-X) \le y\right) \tag{371}$$

$$= \mathbb{P}\left(\log(1-X) \ge -\lambda y\right) \tag{372}$$

$$= \mathbb{P}\left(1 - X \ge e^{-\lambda y}\right) \tag{373}$$

$$= \mathbb{P}\left(X \le 1 - e^{-\lambda y}\right) \tag{374}$$

$$= (1 - e^{-\lambda y})\mathbf{1}(y \ge 0). \tag{375}$$

Hence the CDF of Y is that of an exponential RV with rate λ. Note that the third equality above uses the fact that exponential function is an increasing function. ▲

Remark 1.2. In fact, this is how a computer generates an exponential RV: it first samples a uniform RV X from $[0,1)$, and then outputs $-\frac{1}{\lambda}\log(1-X)$. So the computer does not need to know the exponential distribution in order to generate exponential RVs.

In general, we can at least describe the PDF of $Y = g(X)$ if we know the PDF of X. See the following example for an illustration.

Example 1.3. Let X be a RV with PDF f_X. Define $Y = X^2$. Then

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(X^2 \le y) = \mathbb{P}(-\sqrt{y} \le X \le \sqrt{y})\mathbf{1}(y \ge 0) \tag{376}$$

$$= \mathbf{1}(y \ge 0)\int_{-\sqrt{y}}^{\sqrt{y}} f_X(t)\,dt. \tag{377}$$

By fundamental theorem of calculus and chain rule, we can differentiate the last expression by y and get

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \mathbf{1}(y \ge 0)\left(f_X(\sqrt{y})\frac{1}{2\sqrt{y}} + f_X(-\sqrt{y})\frac{1}{2\sqrt{y}}\right). \tag{378}$$

▲

A particularly simple but useful instance is when g is a linear function. We first record a general observation.

Proposition 1.4 (Linear transform)**.** *Let X be a RV with PDF f_X. Fix constants $a, b \in \mathbb{R}$ with $a > 0$, and define a new RV $Y = aX + b$. Then*

$$f_{aX+b}(y) = \frac{1}{|a|} f_X((y-b)/a). \tag{379}$$

PROOF. First suppose $a > 0$. Then

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(aX + b \le y) = \mathbb{P}(X \le (y-b)/a) = \int_{-\infty}^{(y-b)/a} f_X(t)\, dt. \tag{380}$$

By differentiating the last integral by y, we get

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{1}{a} f_X((y-b)/a). \tag{381}$$

For $a < 0$, a similar calculation shows

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(aX + b \le y) = \mathbb{P}(X \ge (y-b)/a) = \int_{(y-b)/a}^{\infty} f_X(t)\, dt, \tag{382}$$

so we get

$$f_Y(y) = \frac{dF_Y(y)}{dy} = -\frac{1}{a} f_X((y-b)/a). \tag{383}$$

This shows the assertion. □

Example 1.5 (Linear transform of normal RV is normal)**.** Let $X \sim N(\mu, \sigma^2)$ and fix constants $a, b \in \mathbb{R}$ with $a \neq 0$. Define a new RV $Y = aX + b$. Then since

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{384}$$

by Proposition 1.4, we have

$$f_Y(y) = \frac{1}{\sqrt{2\pi(a\sigma)^2}} \exp\left(-\frac{(y-b-a\mu)^2}{2(a\sigma)^2}\right). \tag{385}$$

Notice that this is the PDF of a normal RV with mean $a\mu + b$ and variance $(a\sigma)^2$. In particular, if we take $a = 1/\sigma$ and $b = \mu/\sigma$, then $Y = (X-\mu)/\sigma \sim N(0,1)$, the standard normal RV. This is called *standardization* of normal RV. ▲

Exercise 1.6 (Linear transform of exponential RV)**.** Let $X \sim \mathrm{Exp}(\lambda)$ and fix constants $a, b \in \mathbb{R}$ with $a \neq 0$. Show that

$$f_{aX+b}(x) = \frac{\lambda}{|a|} e^{-\lambda(x-b)/a} \mathbf{1}((x-b)/a > 0). \tag{386}$$

Is $aX + b$ always an exponential RV?

If we compare Example 1.1 and Proposition 1.4 against Example 1.3, we see the invertibility of the function g makes the computation of $F_{g(X)}$ much cleaner. Moreover, if we inspect the formula (379) more closely, we see that the constant factor $1/|a|$ is in fact $1/|g'(x)|$ and $(y-b)/a$ is the inverse function of g, where $g(x) = ax + b$. This leads us to the following observation.

Proposition 1.7 (invertible transform)**.** *Let X be a RV with PDF f_X. Let $g : \mathbb{R} \to \mathbb{R}$ be a differentiable and invertible function. Then we have*

$$f_{g(X)}(y) = (g^{-1})'(y) f_X(g^{-1}(y)) = \frac{1}{|g'(g^{-1}(y))|} f_X(g^{-1}(y)). \tag{387}$$

PROOF. Since g is invertible, g is either strictly increasing or strictly decreasing. First suppose the former, so $g' > 0$ everywhere. Then

$$F_{g(X)}(y) = \mathbb{P}(g(X) \le y) = \mathbb{P}(X \le g^{-1}(y)) = \int_{-\infty}^{g^{-1}(y)} f_X(t)\,dt. \tag{388}$$

Recall that since $g^{-1}(g(x)) = x$, by chain rule we have $(g^{-1})'(g(x)) \cdot g'(x) = 1$. If we write $y = g(x)$, then

$$(g^{-1})'(y) = 1/g'(g^{-1}(y)). \tag{389}$$

Hence differentiating (390) gives (387).

Second, suppose g is strictly decreasing so $g' < 0$ everywhere. Then

$$F_{g(X)}(y) = \mathbb{P}(g(X) \le y) = \mathbb{P}(X \ge g^{-1}(y)) = \int_{g^{-1}(y)}^{\infty} f_X(t)\,dt, \tag{390}$$

so differentiating by y and using (389) also gives (387), as desired. □

Exercise 1.8 (Cauchy from uniform)**.** Let $X \sim \text{Uniform}((-\pi/2, \pi/2))$. Define $Y = \tan(X)$.

(i) Show that $d\tan(y)/dy = \sec^2(y)$.

(ii) Show that $1 + \tan^2(y) = \sec^2(y)$. (Hint: draw a right triangle with angle y)

(iii) Recall that arctan is the inverse function of tan. Show that $\arctan(t)$ is strictly increasing and differentiable. Furthermore, show that

$$\frac{d}{dt}\arctan(t) = \frac{1}{1+t^2}. \tag{391}$$

(iv) Show that Y is a standard Cauchy random variable, that is,

$$f_Y(y) = \frac{1}{\pi(1+y^2)}. \tag{392}$$

Remark 1.9. The expectation of Cauchy random variables are not well-defined. We say the expectation of a continuous RV with PDF $f_X(x)$ is well-defined if

$$\int_{-\infty}^{\infty} |x| f_X(x)\,dx < \infty. \tag{393}$$

But for the standard Cauchy distribution,

$$\int_{-\infty}^{\infty} |x| f_X(x)\,dx = 2\int_0^{\infty} |x| f_X(x)\,dx = \frac{2}{\pi}\int_0^{\infty} \frac{x}{1+x^2}\,dx \approx \frac{2}{\pi}\int_0^{\infty} \frac{x}{x^2}\,dx = \frac{2}{\pi}\int_0^{\infty} \frac{1}{x}\,dx = \infty. \tag{394}$$

In such situation, we say the distribution is 'heavy-tailed'.

We can also cook up a RV from two random variables. Namely, if X, Y are RVs and g is a two variable function $g : \mathbb{R}^2 \to \mathbb{R}$, then $Z = g(X, Y)$ is a random variable.

Example 1.10. Let $X_1 \sim \text{Exp}(\lambda_1)$ and $X_2 \sim \text{Exp}(\lambda_2)$ and suppose they are independent. Define $Y = \max(X_1, X_2)$. To calculate its CDF, note that

$$\mathbb{P}(Y \le y) = \mathbb{P}(\max(X_1, X_2) \le y) = \mathbb{P}(X_1 \le y \text{ and } X_2 \le y) = \mathbb{P}(X_1 \le y)\mathbb{P}(X_2 \le y), \tag{395}$$

where the last equality uses the independence between X_1 and X_2. Using the CDF of exponential RV,

$$\mathbb{P}(Y \le y) = (1 - e^{-\lambda_1 y})(1 - e^{-\lambda_2 y})\mathbf{1}(y \ge 0). \tag{396}$$

Differentiating by y, we get the PDF of Y

$$f_Y(y) = \lambda_1 e^{-\lambda_1 y}(1 - e^{-\lambda_2 y}) + (1 - e^{-\lambda_1 y})\lambda_2 e^{-\lambda_2 y} \tag{397}$$

$$= \lambda_1 e^{-\lambda_1 y} + \lambda_2 e^{-\lambda_2 y} - (\lambda_1 + \lambda_2) e^{-(\lambda_1+\lambda_2)y}. \tag{398}$$

▲

Exercise 1.11. Let $X_1 \sim \text{Exp}(\lambda_1)$ and $X_2 \sim \text{Exp}(\lambda_2)$ and suppose they are independent. Define $Y = \min(X_1, X_2)$. Show that $Y \sim \text{Exp}(\lambda_1 + \lambda_2)$. (Hint: Compute $\mathbb{P}(Y \ge y)$.)

Exercise 1.12. Let $X, Y \sim \text{Uniform}([0,1])$ be independent uniform RVs. Define $Z = X + Y$. Observe that the pair (X, Y) is uniformly distributed over the unit square $[0,1]^2$. So

$$\mathbb{P}(Z \le z) = \mathbb{P}(X + Y \le z) = \text{Area of the region } \{(x,y) \in [0,1]^2 \mid x + y \le z\}. \tag{399}$$

(i) Draw a picture shows that

$$\mathbb{P}(Z \le z) = \begin{cases} z^2/2 & \text{if } 0 \le z \le 1 \\ 1 - (2-z)^2/2 & \text{if } 1 \le z \le 2 \\ 0 & \text{otherwise.} \end{cases} \tag{400}$$

(ii) Conclude that

$$f_Z(z) = \begin{cases} z & \text{if } 0 \le z \le 1 \\ 2 - z & \text{if } 1 \le z \le 2 \\ 0 & \text{otherwise.} \end{cases} \tag{401}$$

2. Sums of independent RVs – Convolution

When two RVs X and Y are independent and if the new random variable Z is their sum $X + Y$, then the distribution Z is given by the *convolution* of PMFs (or PDFs) of each RV. The idea should be clear from the following baby example.

Example 2.1 (Two dice)**.** Roll two dice independently and let their outcome be recorded by RVs X and Y. Note that both X and Y are uniformly distributed over $\{1,2,3,4,5,6\}$. So the pair (X, Y) is uniformly distributed over the (6×6) integer grid $\{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$. In other words,

$$\mathbb{P}((X,Y) = (x,y)) = \mathbb{P}(X = x)\mathbb{P}(Y = y) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}. \tag{402}$$

Now, what is the distribution of the sum $Z = X + Y$? Since each point (x, y) in the grid is equality probable, we just need to count the number of such points on the line $x + y = z$, for each value of z. In other words,

$$\mathbb{P}(X + Y = z) = \sum_{x=1}^{6} \mathbb{P}(X = x)\mathbb{P}(Y = z - x). \tag{403}$$

This is easy to compute from the following picture: For example, $\mathbb{P}(X + Y = 7) = 6/36 = 1/6$. ▲

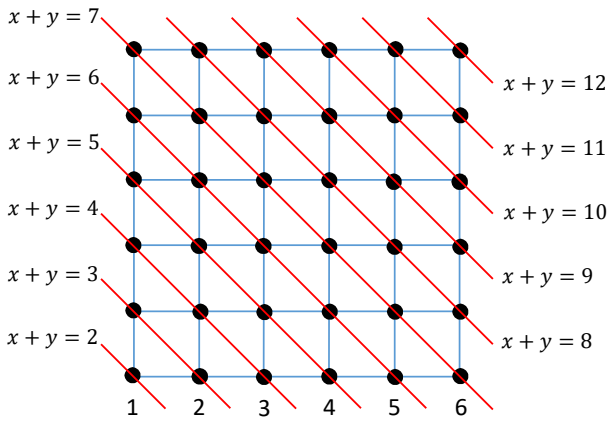

FIGURE 1. Probability space for two dice and lines on which sum of the two are constant.

Proposition 2.2 (Convolution of PMFs)**.** *Let X, Y be two independent integer-valued RVs. Let $Z = X + Y$. Then*

$$\mathbb{P}(Z = z) = \sum_x \mathbb{P}(X = x)\mathbb{P}(Y = z - x). \tag{404}$$

PROOF. Note that the pair (X, Y) is distributed over $\mathbb{Z}^2$ according to the distribution

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y), \tag{405}$$

since X and Y are independent. Hence in order to get $\mathbb{P}(Z = z) = \mathbb{P}(X + Y = z)$, we need to add up all probabilities of the pairs (x, y) over the line $x + y = z$. If we first fix the values of x, then y should take value $z - x$. Varying the range of x, we get (404). □

Exercise 2.3 (Sum of ind. Poisson RVs is Poisson)**.** Let $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2)$ be independent Poisson RVs. Show that $X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

For the continuous case, a similar observation should hold as well. Namely, we should be integrating all the probabilities of the pair (X, Y) at points (x, y) along the line $x + y = z$ in order to get the probability density $f_{X+Y}(z)$. We will show this in the following proposition using Fubini's theorem and change of variables.

Proposition 2.4 (Convolution of PDFs)**.** *Let X, Y be two independent RVs with PDFs f_X and f_Y, respectively. Then the RV $Z := X + Y$ has PDF*

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\, dx. \tag{406}$$

PROOF. As usual, we begin with computing the CDF of Z. Note that since X, Y are independent, the pair (X, Y) is distributed over the plane $\mathbb{R}^2$ according to the distribution

$$f_{X,Y}(x, y) = f_X(x) f_Y(y). \tag{407}$$

So we can write the probability $\mathbb{P}(Z \le z)$ as the following double integral

$$\mathbb{P}(Z \le z) = \int_{x+y \le z} f_X(x) f_Y(y)\, dy\, dx \tag{408}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f_X(x) f_Y(y)\, dy\, dx, \tag{409}$$

where for the second inequality we have used Fubini's theorem. Next, make a change of variable $t = x + y$. Then $y = t - x$ and $dy = dt$, so

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{z} f_X(x) f_Y(t - x)\, dt\, dx. \tag{410}$$

Swapping the order of dt and dx by using Fubini one more time,

$$= \int_{-\infty}^{z} \int_{-\infty}^{\infty} f_X(x) f_Y(t - x)\, dx\, dt = \int_{-\infty}^{z} g(t)\, dt, \tag{411}$$

where we have written the inner integral as a function of t. By differentiating with respect to z, we get

$$f_Z(z) = g(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\, dx. \tag{412}$$

□

Example 2.5. Let $X, Y \sim N(0, 1)$ be independent standard normal RVs. Let $Z = X + Y$. We will show that $Z \sim N(0, 2)$ using the convolution formula. Recall that X and Y have the following PDFs:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \tag{413}$$

By taking convolution of the above PDFs, we have

$$f_Z(z) = \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)\right)\left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-x)^2}{2}\right)\right) dx \tag{414}$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2} - \frac{(z-x)^2}{2}\right) dx \tag{415}$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left(-x^2 + xz - \frac{z^2}{2}\right) dx \tag{416}$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left(-\left(x - \frac{z}{2}\right)^2 - \frac{z^2}{4}\right) dx \tag{417}$$

$$= \frac{1}{\sqrt{4\pi}} e^{-z^2/4} \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} \exp\left(-\left(x - \frac{z}{2}\right)^2\right) dx = \frac{1}{\sqrt{4\pi}} e^{-z^2/4}, \tag{418}$$

where we have recognized the integrand in the line as the PDF of $N(-z/2, 1/2)$ so that the integral is 1. Since the last expression is the PDF of $N(0,2)$, it follows that $Z \sim N(0,2)$. ▲.

The following example generalizes the observation we made in the previous example.

Example 2.6 (Sum of ind. normal RVs is normal)**.** Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ be independent normal RVs. We will see that $Z = X+Y$ is again a normal random variable with distribution $N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$. The usual convolution computation for this is pretty messy (c.f., wikipedia article). Instead let's save some work by using the fact that normal distributions are preserved under linear transform (Exercise 1.5). So instead of X and Y, we may consider $X' := (X-\mu_1)/\sigma_1$ and $Y' := (Y-\mu_1)/\sigma_1$ (It is important to note that we must use the same linear transform here for X and Y). Then $X' \sim N(0,1)$, and $Y' \sim N(\mu, \sigma^2)$ where $\mu = (\mu_2-\mu_1)/\sigma_1$ and $\sigma = \sigma_2/\sigma_1$. Now it suffices to show that $Z' := X' + Y' \sim N(\mu, 1+\sigma^2)$ (see the following exercise for details).

To compute the convolution of the corresponding normal PDFs:

$$f_Z(z) = \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)\right)\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z-x-\mu)^2}{2\sigma^2}\right)\right) dx \tag{419}$$

$$= \frac{1}{2\pi\sigma} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2} - \frac{(z-x-\mu)^2}{2\sigma^2}\right) dx. \tag{420}$$

At this point, we need to 'complete the square' for x for the bracket inside the exponential as below:

$$\frac{x^2}{2} + \frac{(z-x-\mu)^2}{2\sigma^2} = \frac{1}{2\sigma^2}\left(\sigma^2 x^2 + (x+\mu-z)^2\right) \tag{421}$$

$$= \frac{1+\sigma^2}{2\sigma^2}\left(x^2 + \frac{2(\mu-z)x}{1+\sigma^2} + \frac{(\mu-z)^2}{1+\sigma^2}\right) \tag{422}$$

$$= \frac{1+\sigma^2}{2\sigma^2}\left[\left(x + \frac{(\mu-z)}{1+\sigma^2}\right)^2 + \frac{(\mu-z)^2}{1+\sigma^2} - \frac{(\mu-z)^2}{(1+\sigma^2)^2}\right] \tag{423}$$

$$= \frac{1+\sigma^2}{2\sigma^2}\left[\left(x + \frac{(\mu-z)}{1+\sigma^2}\right)^2 + \frac{(z-\mu)^2}{1+\sigma^2}\frac{\sigma^2}{1+\sigma^2}\right] \tag{424}$$

$$= \frac{1+\sigma^2}{2\sigma^2}\left(x + \frac{(\mu-z)}{1+\sigma^2}\right)^2 + \frac{(z-\mu)^2}{2(1+\sigma^2)}. \tag{425}$$

Now rewriting (420),

$$f_Z(z) = \frac{1}{2\pi\sigma} \int_{-\infty}^{\infty} \exp\left(-\frac{(z-\mu)^2}{2(1+\sigma^2)}\right) \exp\left(-\frac{1+\sigma^2}{2\sigma^2}\left(x + \frac{(\mu-z)}{1+\sigma^2}\right)^2\right) dx \tag{426}$$

$$= \frac{1}{\sqrt{2\pi(1+\sigma^2)}} \exp\left(-\frac{(z-\mu)^2}{2(1+\sigma^2)}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\frac{\sigma^2}{1+\sigma^2}}} \exp\left(-\frac{\left(x + \frac{(\mu-z)}{1+\sigma^2}\right)^2}{\frac{2\sigma^2}{1+\sigma^2}}\right) dx \tag{427}$$

$$= \frac{1}{\sqrt{2\pi(1+\sigma^2)}} \exp\left(-\frac{(z-\mu)^2}{2(1+\sigma^2)}\right), \tag{428}$$

where we have recognized the integral after second equality as that of the PDF of a normal RV with mean $\frac{z-\mu}{1+\sigma^2}$ and variance $\frac{\sigma^2}{1+\sigma^2}$. Hence $Z' \sim N(\mu, 1+\sigma^2)$, as desired. ▲

Exercise 2.7. Let X, Y be independent RVs and fix constants $a > 0$ and $b \in \mathbb{R}$.

(i) Show that $X+Y$ is a normal RV if and only if $(aX+b)+(aY+b)$ is so.

(ii) Show that $X+Y$ is a normal RV, then $X+Y \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$, where $\mu_1 = \mathbb{E}(X)$, $\mu_2 = \mathbb{E}(Y)$, $\sigma_1 = \mathrm{Var}(X)$, and $\sigma_1^2 = \mathrm{Var}(Y)$.

Exercise 2.8 (Sum of i.i.d. Exp is Erlang)**.** Let $X_1, X_2, \cdots, X_n \sim \mathrm{Exp}(\lambda)$ be independent exponential RVs.

(i) Show that $f_{X_1+X_2}(z) = \lambda^2 z e^{-\lambda z}\mathbf{1}(z \geq 0)$.

(ii) Show that $f_{X_1+X_2+X_3}(z) = 2^{-1}\lambda^3 z^2 e^{-\lambda z}\mathbf{1}(z \geq 0)$.

(iii) Let $S_n = X_1 + X_2 + \cdots + X_n$. Use induction to show that $S_n \sim \mathrm{Erlang}(n, \lambda)$, that is,

$$f_{S_n}(z) = \frac{\lambda^n z^{n-1} e^{-\lambda z}}{(n-1)!}. \tag{429}$$

3. Covariance and Correlation

When two RVs X and Y are independent, we know that the pair (X, Y) is distributed according to the product distribution $\mathbb{P}((X,Y) = (x,y)) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$ and we can say a lot of things about their sum, difference, product, maximum, etc. For instance, the expectation of their product is the product of their expectations:

Exercise 3.1. Let X and Y be two independent RVs. Show that $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

But what if they are not independent? Then their joint distribution $\mathbb{P}((X,Y) = (x,y))$ can be very much different from the product distribution $\mathbb{P}(X = x)\mathbb{P}(Y = y)$. Covariance is the quantity that measures the 'average disparity' between the true joint distribution $\mathbb{P}((X,Y) = (x,y))$ and the product distribution $\mathbb{P}(X = x)\mathbb{P}(Y = y)$.

Definition 3.2 (Covariance)**.** Given two RVs X and Y, their *covariance* is denoted by $\mathrm{Cov}(X,Y)$ and is defined by

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y). \tag{430}$$

We say X and Y are *correlated* (resp., *uncorrelated*) if $\mathrm{Cov}(X,Y) \neq 0$ (resp., $\mathrm{Cov}(X,Y) = 0$).

Exercise 3.3. Show the following.

(i) $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.

(ii) $\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$.

Exercise 3.4. Show that two RVs X and Y are uncorrelated if they are independent.

Example 3.5 (Uncorrelated but dependent)**.** Two random variables can be uncorrelated but still be dependent. Let (X,Y) be a uniformly sampled point from the unit circle in the 2-dimensional plane. Parameterize the unit circle by $S^1 = \{(\cos\theta, \sin\theta) \mid 0 \leq \theta < 2\pi\}$. Then we can first sample a uniform angle $\Theta \sim \mathrm{Uniform}([0, 2\pi))$, and then define $(X,Y) = (\cos\Theta, \sin\Theta)$. Recall from your old memory that

$$\sin 2t = 2\cos t \sin t. \tag{431}$$

Now

$$\mathbb{E}(XY) = \mathbb{E}(\cos\Theta \sin\Theta) \tag{432}$$

$$= \frac{1}{2}\mathbb{E}(\sin 2\Theta) \tag{433}$$

$$= \frac{1}{2}\int_0^{2\pi} \sin 2t \, dt \tag{434}$$

$$= \frac{1}{2}\left[-\frac{1}{2}\cos 2t\right]_0^{2\pi} = 0. \tag{435}$$

On the other hand,

$$\mathbb{E}(X) = \mathbb{E}(\cos\Theta) = \int_0^{2\pi} \cos t \, dt = 0 \tag{436}$$

and likewise $\mathbb{E}(Y) = 0$. This shows $\mathrm{Cov}(X,Y) = 0$, so X and Y are uncorrelated. However, they satisfy the following deterministic relation

$$X^2 + Y^2 = 1, \tag{437}$$

so clearly they cannot be independent. ▲

So if uncorrelated RVs can be dependent, what does the covariance really measure? It turns out, $\mathrm{Cov}(X,Y)$ measures the 'linear tendency' between X and Y.

Example 3.6 (Linear transform)**.** Let X be a RV, and define another RV Y by $Y = aX+b$ for some constants $a, b \in \mathbb{R}$. Let's compute their covariance using linearity of expectation.

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(X, aX+b) \tag{438}$$

$$= \mathbb{E}(aX^2 + bX) - \mathbb{E}(X)\mathbb{E}(aX+b) \tag{439}$$

$$= a\mathbb{E}(X^2) + b\mathbb{E}(X) - \mathbb{E}(X)(a\mathbb{E}(X)+b) \tag{440}$$

$$= a[\mathbb{E}(X^2) - \mathbb{E}(X)^2] \tag{441}$$

$$= a\mathrm{Var}(X). \tag{442}$$

Thus, $\mathrm{Cov}(X, aX+b) > 0$ if $a > 0$ and $\mathrm{Cov}(X, aX+b) < 0$ if $a < 0$. In other words, if $\mathrm{Cov}(X,Y) > 0$, then X and Y tend to be large at the same time; if $\mathrm{Cov}(X,Y) > 0$, then Y tends to be small if X tends to be large. ▲

From the above example, it is clear that why the x- and y-coordinates of a uniformly sampled point from the unit circle are uncorrelated – they have no linear relation!

Exercise 3.7 (Covariance is symmetric and bilinear)**.** Let X and Y be RVs and fix constants $a, b \in \mathbb{R}$. Show the following.

(i) $\mathrm{Cov}(aX+b, Y) = a\mathrm{Cov}(X,Y)$.
(ii) $\mathrm{Cov}(X+Z, Y) = \mathrm{Cov}(X,Y) + \mathrm{Cov}(Z,Y)$.
(iii) $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$.

Next, let's say four RVs X, Y, Z, and W are given. Suppose that $\mathrm{Cov}(X,Y) > \mathrm{Cov}(Z,W) > 0$. Can we say that 'the positive linear relation' between X and Y is stronger than that between Z and W? Not quite.

Example 3.8. Suppose X is a RV. Let $Y = 2X$, $Z = 2X$, and $W = 4X$. Then

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(X, 2X) = 2\mathrm{Var}(X), \tag{443}$$

and

$$\mathrm{Cov}(Z,W) = \mathrm{Cov}(2X, 4X) = 8\mathrm{Var}(X). \tag{444}$$

But $Y = 2X$ and $W = 2Z$, so the linear relation between the two pairs should be same. ▲

So to compare the magnitude of covariance, we first need to properly normalize covariance so that the effect of fluctuation (variance) of each coordinate is not counted: then only the correlation between the two coordinates will contribute. This is captured by the following quantity.

Definition 3.9 (Correlation coefficient)**.** Given two RVs X and Y, their *correlation coefficient* $\rho(X,Y)$ is defined by

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}\,Y}}. \tag{445}$$

Example 3.10. Suppose X is a RV and fix constants $a,b \in \mathbb{R}$. Then

$$\rho(X,aX+b) = \frac{a\mathrm{Cov}(X,X)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(aX+b)}} = \frac{a\mathrm{Var}(X)}{\sqrt{\mathrm{Var}(X)}\sqrt{a^2\,\mathrm{Var}(X)}} = \frac{a}{|a|} = \mathrm{sign}(a). \tag{446}$$

▲

Exercise 3.11 (Cauchy-Schwarz inequality)**.** Let X,Y are RVs. Suppose $\mathbb{E}(Y^2) > 0$. We will show that the 'inner product' of X and Y is at most the product of their 'magnitudes'

(i) For any $t \in \mathbb{R}$, show that

$$\mathbb{E}[(X-tY)^2] = t^2\mathbb{E}(Y^2) - 2t\mathbb{E}(XY) + \mathbb{E}(X^2) \tag{447}$$

$$= \mathbb{E}(Y^2)\left(t - \frac{\mathbb{E}(XY)}{\mathbb{E}(Y^2)}\right)^2 + \frac{\mathbb{E}(X^2)\mathbb{E}(Y^2) - \mathbb{E}(XY)^2}{\mathbb{E}(Y^2)}. \tag{448}$$

Conclude that

$$0 \le \mathbb{E}\left[\left(X - \frac{\mathbb{E}(XY)}{\mathbb{E}(Y^2)}Y\right)^2\right] = \frac{\mathbb{E}(X^2)\mathbb{E}(Y^2) - \mathbb{E}(XY)^2}{\mathbb{E}(Y^2)}. \tag{449}$$

(ii) Show that a RV Z satisfies $\mathbb{E}(Z^2) = 0$ if and only if $\mathbb{P}(Z=0) = 1$.
(iii) Show that

$$\mathbb{E}(XY) \le \sqrt{\mathbb{E}(X^2)}\sqrt{\mathbb{E}(Y^2)}, \tag{450}$$

where the equality holds if and only if

$$\mathbb{P}\left(X = \frac{\mathbb{E}(XY)}{\mathbb{E}(Y^2)}Y\right) = 1. \tag{451}$$

Exercise 3.12. Let X,Y are RVs such that $\mathrm{Var}(Y) > 0$. Let $\bar{X} = X - \mathbb{E}(X)$ and $\bar{Y} = Y - \mathbb{E}(Y)$.

(i) Use (449) to show that

$$0 \le \mathbb{E}\left[\left(\bar{X} - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}\bar{Y}\right)^2\right] = \mathrm{Var}(X)\left(1 - \rho(X,Y)^2\right). \tag{452}$$

(ii) Show that $|\rho(X,Y)| \le 1$.
(iii) Show that $|\rho(X,Y)| = 1$ if and only if $\bar{X} = a\bar{Y}$ for some constant $a \neq 0$.

4. Variance of sum of RVs

Let X,Y be RVs. If they are not necessarily independent, what is the variance of their sum? Using linearity of expectation, we compute

$$\mathrm{Var}(X+Y) = \mathbb{E}[(X+Y)^2] - \mathbb{E}(X+Y)^2 \tag{453}$$

$$= \mathbb{E}[X^2 + Y^2 + 2XY] - (\mathbb{E}(X) + \mathbb{E}(Y))^2 \tag{454}$$

$$= [\mathbb{E}(X^2) - \mathbb{E}(X)^2] + [\mathbb{E}(Y^2) - \mathbb{E}(Y)^2] + 2[\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)] \tag{455}$$

$$= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y). \tag{456}$$

Note that $\mathrm{Cov}(X,Y)$ shows up in this calculation. We can push this computation for sum of more than just two RVs.

Proposition 4.1. *For RVs $X_1, X_2 \cdots, X_n$, we have*

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}(X_i) + 2 \sum_{1\le i,j\le n} \operatorname{Cov}(X_i, X_j). \tag{457}$$

PROOF. By linearity of expectation, we have

$$\begin{aligned}
\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) &= \mathbb{E}\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] - \left(\mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right)^2 && (458)\\
&= \mathbb{E}\left[\sum_{1\le i,j\le n} X_i X_j\right] - \sum_{1\le i,j\le n} \mathbb{E}(X_i)\mathbb{E}(X_j) && (459)\\
&= \left[\sum_{1\le i,j\le n} \mathbb{E}(X_i X_j)\right] - \sum_{1\le i,j\le n} \mathbb{E}(X_i)\mathbb{E}(X_j) && (460)\\
&= \sum_{1\le i,j\le n} \left[\mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j)\right] && (461)\\
&= \sum_{1\le i\le n} [\mathbb{E}(X_i X_i) - \mathbb{E}(X_i)\mathbb{E}(X_i)] + \sum_{1\le i\ne j\le n} \left[\mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j)\right] && (462)\\
&= \sum_{1\le i\le n} \operatorname{Var}(X_i) + 2 \sum_{1\le i<j\le n} \left[\mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j)\right] && (463)\\
&= \sum_{1\le i\le n} \operatorname{Var}(X_i) + 2 \sum_{1\le i<j\le n} \operatorname{Cov}(X_i, X_j). && (464)
\end{aligned}$$

□

Exercise 4.2. Show that for independent RVs $X_1, X_2, \cdots, X_n$, we have

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}(X_i). \tag{465}$$

Example 4.3 (Number of fixed point in a random permutation)**.** Suppose n people came to a party and somehow the host mixed up their car keys and gave them back completely randomly at the end of the party. Let X_i be a RV, which takes value 1 if person i got the right key and 0 otherwise. Let $N_n = X_1 + X_2 + \cdots + X_n$ be the total number of people who got their own keys back. We will show that $\mathbb{E}(N_n) = \operatorname{Var}(N_n) = 1$.

First, we observe that each $X_i \sim \text{Bernoulli}(1/n)$. So we know that $\mathbb{E}(X_i) = 1/n$ and $\operatorname{Var}(X_i) = \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2 = \mathbb{E}(X_i) - \mathbb{E}(X_i)^2 = n^{-1} - n^{-2} = (n-1)/n^2$. Clearly X_i's are not independent: If the first person got the key number 2, then the second person will never get the right key.

A very important fact is that the linearity of expectation holds regardless of dependence (c.f. Exercise 1.8 in Note 0), so

$$\mathbb{E}[N_n] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}(X_i) = \sum_{i=1}^{n} \frac{1}{n} = 1. \tag{466}$$

On the other hand, to compute the covariance, let's take a look at $\mathbb{E}(X_1 X_2)$. Note that if the first person got her key, then the second person gets his key with probability $1/(n-1)$. So

$$\mathbb{E}(X_1 X_2) = 1 \cdot \mathbb{P}(X_1 = 1,\, X_2 = 1) = \mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 1 \mid X_1 = 1) = \frac{1}{n} \cdot \frac{1}{n-1}. \tag{467}$$

Hence we can compute their covariance:

$$\operatorname{Cov}(X_1, X_2) = \mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2) = \frac{1}{n(n-1)} - \frac{1}{n^2} = \frac{n-(n-1)}{n^2(n-1)} = \frac{1}{n^2(n-1)}. \tag{468}$$

Since there is nothing special about the pair (X_1, X_2), we get

$$\mathrm{Var}(N_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{1\le i,j\le n} \mathrm{Cov}(X_i, X_j) \tag{469}$$

$$= \sum_{i=1}^{n} \frac{n-1}{n^2} + 2 \sum_{1\le i,j\le n} \frac{1}{n^2(n-1)} \tag{470}$$

$$= \frac{n-1}{n} + 2\binom{n}{2} \frac{1}{n^2(n-1)} \tag{471}$$

$$= \frac{n-1}{n} + 2\frac{n(n-1)}{2!} \frac{1}{n^2(n-1)} \tag{472}$$

$$= \frac{n-1}{n} + \frac{1}{n} = 1. \tag{473}$$

So in the above example, we have shown $\mathbb{E}(N) = \mathrm{Var}(N) = 1$. Does this ring a bell? If $X \sim \text{Poisson}(1)$, then $\mathbb{E}(X) = \mathrm{Var}(N) = 1$ (c.f. Exercise 1.21 in Note 0). So is N_n somehow related to the Poisson RV with rate 1? In the following two exercises, we will show that N_n approximately follows Poisson(1) if n is large. ▲

Exercise 4.4 (Derangements)**.** In reference to Example 4.3, let D_n be the total number of arrangements of n keys so that no one gets the correct key.

(i) Show that the total number of arrangements of n keys is $n!$.
(ii) Show that there are $(n-1)!$ arrangements where person 1 got the right key.
(iii) Show that there are $(n-2)!$ arrangements where person 1 and 2 got the right key.
(iv) Show that there are $(n-k)!$ arrangements where person $i_1, i_2, \cdots, i_k$ got the right key.
(v) By using inclusion-exclusion, show that

$$D_n = n! - \binom{n}{1}(n-1)! + \binom{n}{2}(n-2)! - \binom{n}{3}(n-3)! + \cdots + (-1)^n \binom{n}{n}(n-n)! \tag{474}$$

$$= n!\left(1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n \frac{1}{n!}\right) \to \frac{1}{e} \quad \text{as } n \to \infty. \tag{475}$$

Exercise 4.5. Let $N_n = X_1 + X_2 + \cdots + X_n$ be as in Example 4.3.

(i) Use Exercise 4.4 to show that for each $1 \le k \le n$,

$$\mathbb{P}(N_n = k) = \binom{n}{k} \frac{D_{n-k}}{n!} \tag{476}$$

$$= \frac{n!}{k!(n-k)!} \frac{(n-k)!}{n!} \left(1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^{n-k} \frac{1}{(n-k)!}\right) \tag{477}$$

$$= \frac{1}{k!}\left(1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^{n-k} \frac{1}{(n-k)!}\right). \tag{478}$$

(ii) Conclude that

$$\lim_{n\to\infty} \mathbb{P}(N_n = k) = \frac{e^{-1}}{k!} = \mathbb{P}(\text{Poisson}(1) = k). \tag{479}$$

Remark 4.6. Recall that Poisson(1) can be obtained from Binomial(n, p) where $p = 1/n$, for large n (c.f. Example 1.20 in Note 0). In other words, the sum of n independent Bernoulli$(1/n)$ RVs is distributed approximately as Poisson(1). In the key arrangement problem in Example 4.3, note that the correlation coefficient between X_i and X_j is very small:

$$\rho(X_i, X_j) = \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{Var}(X_j)}\sqrt{\mathrm{Var}(X_j)}} = \frac{n^2}{n^2(n-1)^2} = \frac{1}{(n-1)^2}. \tag{480}$$

So it's kind of make sense that X_i's are almost independent for large n, so $N_n \sim \text{Poisson}(1)$ approximately for large n.

CHAPTER 5

Transforms of RVs

In this chapter, we will see how we associate a function $M_X(t)$ to each RV X and how we can understand X by looking at $M_X(t)$ instead. The advantage is that now we can use powerful tools from calculus and analysis (e.g., differentiation, integral, power series, Taylor expansion, etc.) to study RVs.

1. Moment generating function

Let X be a RV. Consider a new RV $g(X) = e^{tX}$, where t is a real-valued parameter we inserted for a reason to be clear soon. A classic point of view of studying X is to look at its *moment generating function* (MGF), which is the expectation $\mathbb{E}[e^{tX}]$ of the RV e^{tX}.

Example 1.1. Let X be a discrete RV with PMF

$$\mathbb{P}(X = x) = \begin{cases} 1/2 & \text{if } x = 2 \\ 1/3 & \text{if } x = 3 \\ 1/6 & \text{if } x = 5. \end{cases} \tag{481}$$

Its MGF is

$$\mathbb{E}[e^{tX}] = \frac{e^{2t}}{2} + \frac{e^{3t}}{3} + \frac{e^{5t}}{6}. \tag{482}$$

▲

Here is a heuristic for why we might be interested in the MGF of X. Recall the Taylor expansion of the exponential function e^s:

$$e^s = 1 + \frac{s}{1!} + \frac{s^2}{2!} + \frac{s^3}{3!} + \cdots. \tag{483}$$

Plug in $s = tX$ and get

$$e^{tX} = 1 + \frac{X}{1!}t + \frac{X^2}{2!}t^2 + \frac{X^3}{3!}t^3 + \cdots. \tag{484}$$

Taking expectation and using its 'linearity', this gives us

$$\mathbb{E}[e^{tX}] = 1 + \frac{\mathbb{E}[X]}{1!}t + \frac{\mathbb{E}[X^2]}{2!}t^2 + \frac{\mathbb{E}[X^3]}{3!}t^3 + \cdots. \tag{485}$$

Notice that the right hand side is a power series in variable t, and the kth moment $\mathbb{E}[X^k]$ of X shows up in the coefficient of the kth order term t^k. In other words, by simply taking the expectation of e^{tX}, we can get all higher moments of X. In this sense, the MGF $\mathbb{E}[e^{tX}]$ generates all moments of X, hence we call its name 'moment generating function'.

As you might have noticed, the equation (485) needs more justification. For example, what if $\mathbb{E}[X^3]$ is infinity? Also, can we really use linearity of expectation for a sum of infinitely many RVs as in the right hand side of (484)? We will get to this theoretical point later, and for now let's get ourselves more familiar to MGF computation.

Example 1.2 (Bernoulli RV)**.** Let $X \sim \text{Bernoulli}(p)$. Then

$$\mathbb{E}[e^{tX}] = e^t p + e^0(1-p) = 1 - p + e^t p. \tag{486}$$

▲

Example 1.3 (Poisson RV)**.** Let $X \sim \text{Poisson}(\lambda)$. Then using the Taylor expansion of the exponential function,

$$\mathbb{E}[e^{tX}] = \sum_{k=0}^{\infty} e^{kt} \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^t\lambda)^k}{k!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t-1)}. \tag{487}$$

Exercise 1.4 (Geometric RV)**.** Let $X \sim \text{Geom}(p)$. Show that

$$\mathbb{E}[e^{tX}] = \frac{pe^t}{1-(1-p)e^t}. \tag{488}$$

Example 1.5 (Uniform RV)**.** Let $X \sim \text{Uniform}([a,b])$. Then

$$\mathbb{E}[e^{tX}] = \int_a^b e^{tx} \frac{1}{b-a}\, dx = \frac{1}{b-a}\left[\frac{e^{tx}}{t}\right]_a^b = \frac{e^{bt}-e^{at}}{t(b-a)}. \tag{489}$$

▲

Example 1.6 (Exponential RV)**.** Let $X \sim \text{Exp}(\lambda)$. Then

$$\mathbb{E}[e^{tX}] = \int_0^\infty e^{tx} \lambda e^{-\lambda x}\, dx = \lambda \int_0^\infty e^{(t-\lambda)x}\, dx. \tag{490}$$

Considering two cases when $t < \lambda$ and $t \geq \lambda$, we get

$$\mathbb{E}[e^{tX}] = \begin{cases} \frac{\lambda}{\lambda - t} & \text{if } t < \lambda \\ \infty & \text{if } t \geq \lambda. \end{cases} \tag{491}$$

▲

Example 1.7 (Standard normal RV)**.** Let $X \sim N(0,1)$. Then

$$\mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2+tx}\, dx. \tag{492}$$

By completing square, we can write

$$-\frac{x^2}{2} + tx = -\frac{1}{2}(x^2 - 2tx) = \frac{1}{2}(x-t)^2 + \frac{t^2}{2}. \tag{493}$$

So we get

$$\mathbb{E}[e^{tX}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} e^{t^2/2}\, dx = e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(x-t)^2/2}\, dx. \tag{494}$$

Notice that the integrand in the last expression is the PDF of a normal RV with distribution $N(t,1)$. Hence the last integral equals 1, so we conclude

$$\mathbb{E}[e^{tX}] = e^{t^2/2}. \tag{495}$$

▲

Exercise 1.8 (MGF of linear transform)**.** Let X be a RV and a, b be constants. Let $M_X(t)$ be the MGF of X. Then show that

$$\mathbb{E}[e^{t(aX+b)}] = e^{bt} M_X(at). \tag{496}$$

Exercise 1.9 (Standard normal)**.** Let $X \sim N(\mu, \sigma^2)$ and $Z \sim N(0,1)$. Using the fact that $\mathbb{E}[e^{tZ}] = e^{t^2/2}$ and Exercise 1.9, show that

$$\mathbb{E}[e^{tY}] = e^{\sigma^2 t^2/2 + t\mu}. \tag{497}$$

2. Two important theorems about MGFs

The power series expansion (485) of MGF may not be valid in general. The following theorem gives a sufficient condition for which such an expansion is true. We omit its proof in this lecture.

Theorem 2.1. *Let X be a RV. Suppose there exists a constant $h > 0$ such that $\mathbb{E}[e^{tX}] < \infty$ for all $x \in (-h, h)$. Then the kth moment $\mathbb{E}[X^k]$ exists for all $k \geq 0$ and there exists a constant $\varepsilon > 0$ such that for all $t \in (-\varepsilon, \varepsilon)$,*

$$\mathbb{E}[e^{tX}] = \sum_{k=0}^{\infty} \frac{\mathbb{E}[X^k]}{k!} t^k. \tag{498}$$

For each RV X, we say its MGF *exists* whenever the hypothesis of the above theorem holds. One of the consequence of the above theorem is that we can access its kth moment by taking kth derivative of its MGF and evaluating at $t = 0$.

Exercise 2.2. Suppose the MGF of a RV X exists. Then show that for each integer $k \geq 0$,

$$\frac{d^k}{dt^k} \mathbb{E}[e^{tX}]\Big|_{t=0} = \mathbb{E}[X^k]. \tag{499}$$

Example 2.3 (Poisson RV)**.** Let $X \sim \text{Poisson}(\lambda)$. In Example 1.3, we have computed

$$\mathbb{E}[e^{tX}] = e^{\lambda(e^t - 1)} \qquad \forall\, t \in \mathbb{R}. \tag{500}$$

Differentiating by t and evaluating at $t = 0$, we get

$$\mathbb{E}[X] = \frac{d}{dt} e^{\lambda(e^t-1)}\Big|_{t=0} = e^{\lambda(e^t-1)} \lambda e^t \Big|_{t=0} = \lambda. \tag{501}$$

We can also compute its second moment as

$$\mathbb{E}[X^2] = \frac{d^2}{dt^2} e^{\lambda(e^t-1)}\Big|_{t=0} = \frac{d}{dt} \lambda e^{\lambda(e^t-1)+t}\Big|_{t=0} = \lambda e^{\lambda(e^t-1)+t}(\lambda e^t + 1)\Big|_{t=0} = \lambda(\lambda+1). \tag{502}$$

This also implies that

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \lambda(\lambda+1) - \lambda^2 = \lambda. \tag{503}$$

▲

Example 2.4 (Exponential RV)**.** Let $X \sim \text{Exp}(\lambda)$. Our calculation in Example 1.6 implies that

$$\mathbb{E}[e^{tX}] = \frac{\lambda}{\lambda - t} \qquad t \in (-\lambda, \lambda). \tag{504}$$

We can compute the first and second moment of X:

$$\mathbb{E}[X] = \frac{d}{dt} \frac{\lambda}{\lambda - t}\Big|_{t=0} = \frac{\lambda}{(\lambda - t)^2}\Big|_{t=0} = \frac{1}{\lambda} \tag{505}$$

$$\mathbb{E}[X^2] = \frac{d^2}{dt^2} \frac{\lambda}{\lambda - t}\Big|_{t=0} = \frac{d}{dt} \frac{\lambda}{(\lambda - t)^2}\Big|_{t=0} = \frac{2\lambda}{(\lambda - t)^3}\Big|_{t=0} = \frac{2}{\lambda^2}. \tag{506}$$

In fact, by recognizing $\lambda/(\lambda - t)$ as a geometric series,

$$\mathbb{E}[e^{tX}] = \frac{1}{1 - t/\lambda} = 1 + (t/\lambda) + (t/\lambda)^2 + (t/\lambda)^3 + \cdots \tag{507}$$

$$= 1 + \frac{1!/\lambda}{1!} t + \frac{2!/\lambda^2}{2!} t^2 + \frac{3!/\lambda^3}{3!} t^3 + \cdots. \tag{508}$$

Hence by comparing with (498), we conclude that $\mathbb{E}[X^k] = k!/\lambda^k$ for all $k \geq 0$. ▲

The second theorem for MGFs is that they determine the distribution of RVs. This will be critically used later in the proof of the central limit theorem.

Theorem 2.5. *Let X, Y, and X_n for $n \geq 1$ be RVs whose MGFs exist.*

(i) *(Uniqueness) Suppose* $\mathbb{E}[e^{tX}] = \mathbb{E}[e^{tY}]$ *for all sufficiently small* t*. Then* $\mathbb{P}(X \le s) = \mathbb{P}(Y \le s)$ *for all* $s \in \mathbb{R}$*.*

(ii) *(Continuity) Suppose* $\lim_{n\to\infty} \mathbb{E}[e^{tX_n}] = \mathbb{E}[e^{tX}]$ *for all sufficiently small* t *and that* $\mathbb{E}[e^{tX}]$ *is continuous at* $t = 0$*. Then* $\mathbb{P}(X_n \le s) \to \mathbb{P}(X \le s)$ *for all* s *such that* $\mathbb{P}(X \le x)$ *is continuous at* $x = s$*.*

3. MGF of sum of independent RVs

One of the nice properties of MGFs is the following factorization for sums of independent RVs.

Proposition 3.1. *Let* X, Y *be independent RVs. Then*

$$\mathbb{E}[e^{t(X+Y)}] = \mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}]. \tag{509}$$

If you believe that the RVs e^{tX} and e^{tY} are independent, then the proof of the above result is one-line:

$$\mathbb{E}[e^{t(X+Y)}] = \mathbb{E}[e^{tX}e^{tY}] = \mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}]. \tag{510}$$

In general, it is a special case of the following result.

Proposition 3.2. *Let* X, Y *be independent RVs. Then for any integrable functions* $g_1, g_2 : \mathbb{R} \to \mathbb{R}$*, we have*

$$\mathbb{E}[g_1(X)g_2(Y)] = \mathbb{E}[g_1(X)]\mathbb{E}[g_2(Y)]. \tag{511}$$

PROOF. If X, Y are continuous RVs,

$$\begin{aligned}
\mathbb{E}[g_1(X)g_2(Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_1(x)g_2(y) f_{X,Y}(x,y)\,dx\,dy && (512)\\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_1(x)g_2(y) f_X(x) f_Y(y)\,dx\,dy && (513)\\
&= \int_{-\infty}^{\infty} g_1(x) f_X(x) \left(\int_{-\infty}^{\infty} g_2(y) f_Y(y) \right) dy && (514)\\
&= \mathbb{E}[g_2(Y)] \int_{-\infty}^{\infty} g_1(x) f_X(x)\,dx && (515)\\
&= \mathbb{E}[g_1(X)]\mathbb{E}[g_2(Y)]. && (516)
\end{aligned}$$

For discrete RVs, use summation and PMF instead of integral and PDF. □

Exercise 3.3 (Binomial RV)**.** Let $X \sim \text{Binomial}(n, p)$. Use the MGF of Bernoulli RV and Proposition 3.1 to show that

$$\mathbb{E}[e^{tX}] = (1 - p + e^t p)^n. \tag{517}$$

Example 3.4 (Sum of independent Poisson RVs)**.** Let $X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \sim \text{Poisson}(\lambda_2)$ be independent Poisson RVs. Let $Y = X_1 + X_2$. Using Exercise 1.3, we have

$$\mathbb{E}[e^{tY}] = \mathbb{E}[e^{tX_1}]\mathbb{E}[e^{tX_2}] = e^{(\lambda_1+\lambda_2)(e^t-1)}. \tag{518}$$

Notice that the last expression is the MGF of a Poisson RV with rate $\lambda_1 + \lambda_2$. By the Uniqueness of MGF (Theorem 2.5 (i)), we conclude that $Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$. ▲

Exercise 3.5 (Sum of independent normal RVs)**.** Let $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ be independent normal RVs.

(i) Show that $\mathbb{E}[e^{t(X_1+X_2)}] = \exp[(\sigma_1^2 + \sigma_2^2)t^2/2 + t(\mu_1 + \mu_2)]$.

(ii) Conclude that $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

4. Sum of random number of independent RVs

Suppose $X_1, X_2, \cdots$ are independent and identically distributed (i.i.d.) RVs and let N be another independent RV taking values in nonnegative integers (e.g., Binomial). For a new RV Y by

$$Y = X_1 + X_2 + \cdots + X_N. \tag{519}$$

Note that we are summing a random number of X_i's, so there are two sources of randomness that determines Y. As usual, we use conditioning to study such RVs. For instance,

$$\mathbb{E}[Y \mid N = n] = \mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = n\mathbb{E}X_1 \tag{520}$$

$$\mathrm{Var}(Y \mid N = n) = \mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = n\,\mathrm{Var}(X_1). \tag{521}$$

Hence iterated expectation gives

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|N]] = \mathbb{E}[N\mathbb{E}[X_1]] = \mathbb{E}[X_1]\mathbb{E}[N]. \tag{522}$$

On other other hand, law of total variance gives

$$\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y \mid N)] + \mathrm{Var}(\mathbb{E}[Y \mid N]) \tag{523}$$

$$= \mathbb{E}[N\,\mathrm{Var}(X_1)] + \mathrm{Var}(N\mathbb{E}[X_1]) \tag{524}$$

$$= \mathrm{Var}(X_1)\mathbb{E}[N] + \mathbb{E}[X_1]^2\,\mathrm{Var}(N). \tag{525}$$

Furthermore, can we also figure out the MGF of Y? After all, MGF is an expectation so we can also get it by iterated expectation. First we compute the conditional version. Denoting $M_X(t) = \mathbb{E}[e^{tX}]$,

$$\mathbb{E}[e^{tY} \mid N = n] = \mathbb{E}[e^{t(X_1+\cdots+X_n)}] = \mathbb{E}[e^{tX_1} \cdots e^{tX_n}] \tag{526}$$

$$= \mathbb{E}[e^{tX_1}] \cdots \mathbb{E}[e^{tX_n}] = \mathbb{E}[e^{tX_1}]^n = M_{X_1}(t)^n \tag{527}$$

$$= e^{n \log M_{X_1}(t)}. \tag{528}$$

The last line is the trick here. Now the iterated expectation gives

$$\mathbb{E}[e^{tY}] = \mathbb{E}[\mathbb{E}[e^{tY} \mid N]] = \mathbb{E}[e^{(\log M_{X_1}(t))N}]. \tag{529}$$

Note that the last expression is nothing but the MFG of N evaluated at $\log M_{X_1}(t)$ instead of t. Hence

$$\mathbb{E}[e^{tY}] = M_N(\log M_{X_1}(t)). \tag{530}$$

Let us summarize what have obtained so far.

Proposition 4.1. *Let $X_1, X_2, \cdots$ be i.i.d. RVs and let N be another independent RV which takes values from nonnegative integers. Let $Y = \sum_{k=0}^{N} X_k$. Denote the MGF of any RV Z by $M_Z(t)$. Then we have*

$$\mathbb{E}[Y] = \mathbb{E}[N]\mathbb{E}[X_[1]] \tag{531}$$

$$\mathrm{Var}[Y] = \mathrm{Var}(X_1)\mathbb{E}[N] + \mathbb{E}[X_1]^2\,\mathrm{Var}(N) \tag{532}$$

$$M_Y(t) = M_N(\log M_{X_1}(t)). \tag{533}$$

Example 4.2. Let $X_i \sim \mathrm{Exp}(\lambda)$ for $i \geq 0$ and let $N \sim \mathrm{Poisson}(\lambda)$. Suppose all RVs are independent. Define $Y = \sum_{k=1}^{N} X_i$. Then

$$\mathbb{E}[Y] = \mathbb{E}[N]\mathbb{E}[X_1] = \lambda/\lambda = 1, \tag{534}$$

$$\mathrm{Var}(Y) = \mathrm{Var}(X_1)\mathbb{E}[N] + \mathbb{E}[X_1]^2\,\mathrm{Var}(N) = \frac{\lambda}{\lambda^2} + \frac{\lambda}{\lambda^2} = \frac{2}{\lambda}. \tag{535}$$

On other hand, recall that $M_{X_1}(t) = \frac{\lambda}{\lambda - t}$ and $M_N(t) = e^{\lambda(e^t - 1)}$. Hence

$$\mathbb{E}[e^{tY}] = e^{\lambda(\exp(\log \frac{\lambda}{\lambda - t}) - 1)} = e^{\lambda(\frac{\lambda}{\lambda - t} - 1)} = e^{\frac{\lambda t}{\lambda - t}}. \tag{536}$$

So we know everything about Y. Knowing the MGF of Y, we could get all the moments of Y. For instance,

$$\mathbb{E}[Y] = \frac{d}{dt} e^{\frac{\lambda t}{\lambda - t}}\Big|_{t=0} = e^{\frac{\lambda t}{\lambda - t}} \frac{\lambda(\lambda - t) + \lambda t}{(\lambda - t)^2}\Big|_{t=0} = 1. \tag{537}$$

▲

Exercise 4.3. Let $X_1, X_2, \cdots$ be i.i.d. RVs and let N be another independent RV which takes values from nonnegative integers. Let $Y = \sum_{k=0}^{N} X_k$. Denote the MGF of any RV Z by $M_Z(t)$. Using the fact that $M_Y(t) = M_N(\log M_{X_1}(t))$, derive

$$\mathbb{E}[Y] = \mathbb{E}[N]\mathbb{E}[X_1], \tag{538}$$

$$\mathrm{Var}[Y] = \mathrm{Var}(X_1)\mathbb{E}[N] + \mathbb{E}[X_1]^2 \mathrm{Var}(N). \tag{539}$$

Example 4.4. Let $X_i \sim \mathrm{Exp}(\lambda)$ for $i \geq 0$ and let $N \sim \mathrm{Geom}(p)$. Let $Y = \sum_{k=1}^{N} X_k$. Suppose all RVs are independent. Recall that

$$M_{X_1}(t) = \frac{\lambda}{\lambda - t}, \qquad M_N(t) = \frac{pe^t}{1-(1-p)e^t}. \tag{540}$$

Hence

$$M_Y(t) = \frac{p\frac{\lambda}{\lambda - t}}{1-(1-p)\frac{\lambda}{\lambda - t}} = \frac{p\lambda}{(\lambda - t) - \lambda(1-p)} = \frac{p\lambda}{p\lambda - t}. \tag{541}$$

Notice that this is the MGF of an $\mathrm{Exp}(p\lambda)$ variable. Thus by uniqueness, we conclude that $Y \sim \mathrm{Exp}(p\lambda)$. If you remember, sum of k independent $\mathrm{Exp}(\lambda)$ RVs were not an exponential RV (its distribution is $\mathrm{Erlang}(k, \lambda)$. See Exercise 1.19 in Note 1). But as we have seen in this example, if you sum a random number of independent exponentials, they could by exponential again. ▲

Exercise 4.5. Let $X_i \sim \mathrm{Geom}(q)$ for $i \geq 0$ and let $N \sim \mathrm{Geom}(p)$. Suppose all RVs are independent. Let $Y = \sum_{k=0}^{N} X_k$.

(i) Show that the MGF of Y is given by

$$\mathbb{E}[e^{tY}] = \frac{pqe^t}{1-(1-pq)e^t}. \tag{542}$$

(ii) Conclude that $Y \sim \mathrm{Geom}(pq)$.

CHAPTER 6

Elementary limit theorems

1. Overview of limit theorems

The primary subject in this note is the sequence of i.i.d. RVs and their partial sums. Namely, let $X_1, X_2, \cdots$ be an (infinite) sequence of i.i.d. RVs, and define their nth partial sum $S_n = X_1 + X_2 + \cdots + X_n$ for all $n \geq 1$. If we call X_i the ith step size or *increment*, then the sequence of RVs $(S_n)_{n\geq 1}$ is called a *random walk*, where we usually set $S_0 = 0$. Think of X_i as the gain or loss after betting once in a casino. Then S_n is the net gain of fortune after betting n times. Of course there are ups and downs in the short term, but what we want to analyze using probability theory is the long-term behavior of the random walk $(S_n)_{n\geq 1}$. Results of this type is called limit theorems.

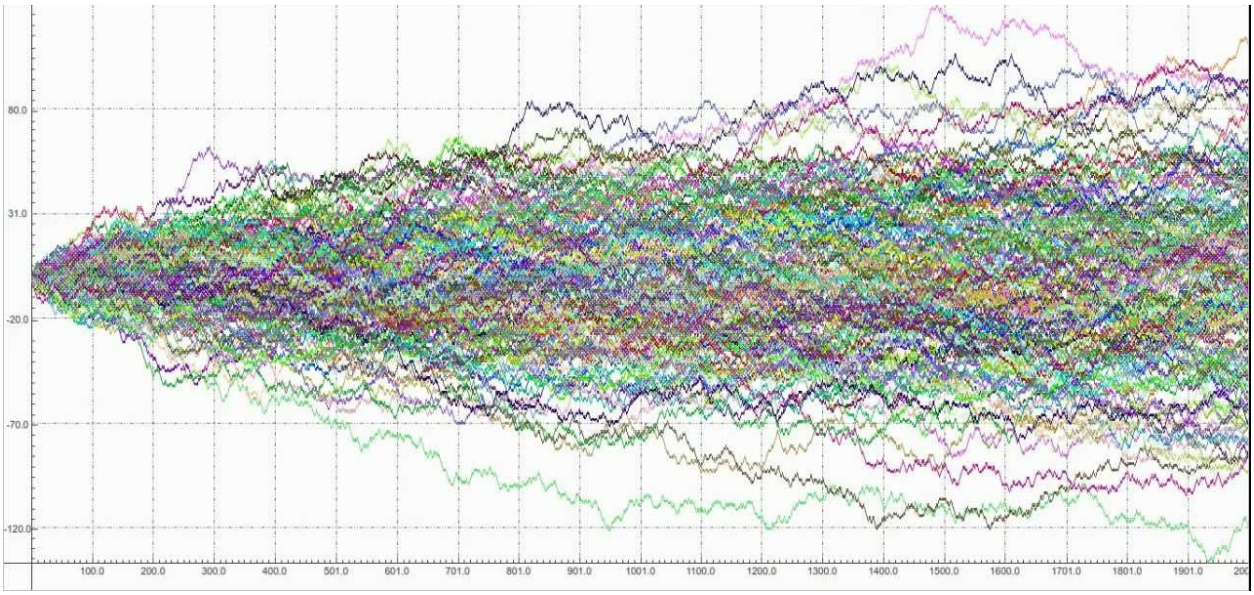

FIGURE 1. Simulation of simple random walks

Suppose each increment X_k has a finite mean μ. Then by linearity of expectation and independence of the increments, we have

$$\mathbb{E}\left(\frac{S_n}{n}\right) = \frac{\mathbb{E}[S_n]}{n} = \mu, \tag{543}$$

$$\mathrm{Var}\left(\frac{S_n}{n}\right) = \frac{\mathrm{Var}(S_n)}{n^2} = \frac{n\,\mathrm{Var}(X_1)}{n^2} = \frac{\mathrm{Var}(X_1)}{n}. \tag{544}$$

So the sample mean S_n/n has constant expectation and shrinking variance. Hence it makes sense to guess that it should behave as the constant μ, without taking the expectation. That is,

$$\lim_{n\to\infty} \frac{S_n}{n} = \mu. \tag{545}$$

But this expression is shaky, since the left hand side is a limit of RVs while the right hand side is a constant. In what sense the random sample means converge to μ? This is the content of the *law of large numbers*, for which we will prove a weak and a strong versions.

The first limit theorem we will encounter is called the Weak Law of Large Numbers (WLLN), which is stated below:

Theorem 1.1 (WLLN)**.** *Let $(X_k)_{k\geq 1}$ be i.i.d. RVs with mean $\mu < \infty$ and let $S_n = \sum_{k=1}^n X_i$, $n \geq 1$ be a random walk. Then for any positive constant $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) = 0. \tag{546}$$

In words, the probability that the sample mean S_n/n is *not* within ε distance from its expectation μ decays to zero as n tends to infinity. In this case, we say the sequence of RVs $(S_n/n)_{n\geq 1}$ converges to μ *in probability*.

The second version of law of large numbers is call the *strong law of large numbers* (SLLN), which is available if the increments have finite variance.

Theorem 1.2 (SLLN)**.** *Let $(X_k)_{k\geq 1}$ be i.i.d. RVs and let $S_n = \sum_{k=1}^n X_i$, $n \geq 1$ be a random walk. Suppose $\mathbb{E}[X_1] = \mu < \infty$ and $\mathbb{E}[X_1^2] < \infty$. Then*

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} = \mu\right) = 1. \tag{547}$$

To make sense out of this, notice that the limit of sample mean $\lim_{n\to} S_n/n$ is itself a RV. Then SLLN says that this RV is well defined and its value is μ with probability 1. In this case, we say the sequence of RVs $(S_n/n)_{n\geq 1}$ converges to μ *with probability 1* or *almost surely*.

Perhaps one of the most celebrated theorems in probability theory is the *central limit theorem* (CLT), which tells about how the sample mean S_n/n "fluctuates" around its mean μ. From 544, if we denote $\sigma^2 = \mathrm{Var}(X_1) < \infty$, we know that $\mathrm{Var}(S_n/n) = \sigma^2/n \to 0$ as $n \to \infty$. So the fluctuation decays as we add up more increments. To see the effect of fluctuation, we first center the sample mean by subtracting its expectation and "zoom in" by dividing by the standard deviation $\sigma/\sqrt{n}$. This is where the name 'central limit' comes from: it describes the limit of centered random walks.

Theorem 1.3 (CLT)**.** *Let $(X_k)_{k\geq 1}$ be i.i.d. RVs and let $S_n = \sum_{k=1}^n X_i$, $n \geq 1$ be a random walk. Suppose $\mathbb{E}[X_1] = \mu < \infty$ and $\mathbb{E}[X_1^2] = \sigma^2 < \infty$. Let $Z \sim N(0,1)$ be a standard normal RV and define*

$$Z_n = \frac{S_n - \mu n}{\sigma\sqrt{n}} = \frac{S_n/n - \mu}{\sigma/\sqrt{n}}. \tag{548}$$

Then for all $z \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}(Z_n \leq z) = \mathbb{P}(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}}\, dx. \tag{549}$$

In words, the centered and rescaled RV Z_n is asymptotically distributed as a standard normal RV $Z \sim N(0,1)$. In this case, we say Z_n converges to Z as $n \to \infty$ *in distribution*. This is a remarkable result since as long as the increments X_k have finite mean and variance, it does not matter which distribution that they follow: the 'central limit' always looks like a standard normal distribution. Later in this section, we will prove this result by using the MGF of S_n and Taylor-expanding it up to the second order term.

In 170A, we will only study the Weak Law of Large Numbers and the Central Limit Theorem in a very special case when X_i's are Binomial RVs.

2. Bounding tail probabilities

In this subsection, we introduce two general inequalities called the Markov's and Chebyshef's inequalities. They are useful in bounding tail probabilities of the form $\mathbb{P}(X \geq x)$ using the expectation $\mathbb{E}[X]$

and variance $\mathrm{Var}(X)$, respectively. Their proofs are quite simple but they have lots of nice applications and implications.

Proposition 2.1 (Markov's inequality)**.** *Let $X \geq 0$ be a nonnegative RV with finite expectation. Then for any $a > 0$, we have*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}. \tag{550}$$

PROOF. Consider an auxiliary RV Y define as follows:

$$Y = \begin{cases} a & \text{if } X \geq a \\ 0 & \text{if } X < a. \end{cases} \tag{551}$$

Note that we always have $Y \leq X$. Hence we should have $\mathbb{E}[Y] \leq \mathbb{E}[X]$. But since $\mathbb{E}[Y] = a\mathbb{P}(X \geq a)$, we have

$$\lambda\mathbb{P}(X \geq a) \leq \mathbb{E}[X]. \tag{552}$$

Dividing both sides by $a > 0$ gives the assertion. □

Example 2.2. We will show that, for any RV Z, $\mathbb{E}[Z^2] = 0$ implies $\mathbb{P}(Z = 0) = 1$. Indeed, Markov's inequality gives that for any $a > 0$,

$$\mathbb{P}(Z^2 \geq a) \leq \frac{\mathbb{E}[Z^2]}{a} = 0. \tag{553}$$

This means that $\mathbb{P}(Z^2 = 0) = 1$, so $\mathbb{P}(Z = 0) = 1$. ▲

Proposition 2.3 (Chebyshef's inequality)**.** *Let X be any RV with $\mathbb{E}[X] = \mu < \infty$ and $\mathrm{Var}(X) < \infty$. Then for any $a > 0$, we have*

$$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}. \tag{554}$$

PROOF. Applying Markov's inequality for the nonnegative RV $(X-\mu)^2$, we get

$$\mathbb{P}(|X - \mu| \geq a) = \mathbb{P}((X-\mu)^2 \geq a^2) \leq \frac{\mathbb{E}[(X-\mu)^2]}{a^2} = \frac{\mathrm{Var}(X)}{a^2}. \tag{555}$$

□

Example 2.4. Let $X \sim \mathrm{Exp}(\lambda)$. Since $\mathbb{E}[X] = 1/\lambda$, for any $a > 0$, the Markov's inequality gives

$$\mathbb{P}(X \geq a) \leq \frac{1}{a\lambda}, \tag{556}$$

while the true probability is

$$\mathbb{P}(X \geq a) = e^{-\lambda a}. \tag{557}$$

On the other hand, $\mathrm{Var}(X) = 1/\lambda^2$ so Chebyshef's inequality gives

$$\mathbb{P}(|X - 1/\lambda| \geq a) = \frac{1}{a^2\lambda^2}. \tag{558}$$

If $1/\lambda \leq a$, the true probability is

$$\mathbb{P}(|X - 1/\lambda| \geq a) = \mathbb{P}(X \geq a + 1/\lambda) + \mathbb{P}(X \leq -a + 1/\lambda) \tag{559}$$

$$= \mathbb{P}(X \geq a + 1/\lambda) = e^{-\lambda(a+1/\lambda)} = e^{-1-\lambda a}. \tag{560}$$

As we can see, both Markov's and Chebyshef's inequalities give loose estimates, but the latter gives a slightly stronger bound. ▲

Example 2.5 (Chebyshef's inequality for bounded RVs)**.** Let X be a RV taking values from the interval $[a,b]$. Suppose we don't know anything else about X. Can we say anything useful about tail probability $\mathbb{P}(X \geq \lambda)$? If we were to use Markov's inequality, then certainly $a \leq \mathbb{E}[X] \leq b$ and in the worst case $\mathbb{E}[X] = b$. Hence we can at least conclude

$$\mathbb{P}(X \geq \lambda) \leq \frac{b}{\lambda}. \tag{561}$$

On the other hand, let's get a bound on $\mathrm{Var}(X)$ and use Chebyshef's inequality instead. We claim that

$$\mathrm{Var}(X) \leq \frac{(b-a)^4}{4}, \tag{562}$$

which would yield by Chebyshef's inequality that

$$\mathbb{P}(|X - \mathbb{E}[X]| \leq \lambda) \leq \frac{(b-a)^2}{4\lambda^2}. \tag{563}$$

Intuitively speaking, $\mathrm{Var}(X)$ is the largest when the value of X is as much spread out as possible at the two extreme values, a and b. Hence the largest variance will be achieved when X takes a and b with equal probabilities. In this case, $\mathbb{E}[X] = (a+b)/2$ so

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{a^2+b^2}{2} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{4}. \tag{564}$$

▲

Exercise 2.6. Let X be a RV taking values from the interval $[a,b]$.

(i) Use the usual 'completing squares' trick for a second moment to show that

$$0 \leq \mathbb{E}[(X-t)^2] = (t - \mathbb{E}[X])^2 + \mathrm{Var}(X) \quad \forall t \in \mathbb{R}. \tag{565}$$

(ii) Conclude that $\mathbb{E}[(X-t)^2]$ is minimized when $t = \mathbb{E}[X]$ and the minimum is $\mathrm{Var}(X)$.

(iii) By plugging in $t = (a+b)/2$ in (565), show that

$$\mathrm{Var}(X) = \mathbb{E}[(X-a)(X-b)] + \frac{(b-a)^2}{4} - \left(\mathbb{E}[X] - \frac{a+b}{2}\right)^2. \tag{566}$$

(iv) Show that $\mathbb{E}[(X-a)(X-b)] \leq 0$.

(v) Conclude that $\mathrm{Var}(X) \leq (b-a)^2/4$, where the equality holds if and only if X takes the extreme values a and b with equal probabilities.

Exercise 2.7 (Paley-Zigmund inequality)**.** Let X be a nonnegative RV with $\mathbb{E}[|X|] < \infty$. Fix a constant $\theta \geq 0$. We prove the Paley-Zigmund inequality, which gives a lower bound on the tail probabilities and also implies the so-called 'second moment method'.

(i) Write $X = X\mathbf{1}(X > \theta\mathbb{E}[X]) + X\mathbf{1}(X \leq \theta\mathbb{E}[X])$. Show that

$$\mathbb{E}[X] = \mathbb{E}[X\mathbf{1}(X \leq \theta\mathbb{E}[X])] + \mathbb{E}[X\mathbf{1}(X > \theta\mathbb{E}[X])] \tag{567}$$

$$\leq \theta\mathbb{E}[X] + \mathbb{E}[X\mathbf{1}(X > \theta\mathbb{E}[X])]. \tag{568}$$

(ii) Use Cauchy-Schwartz inequality (Exc 2.11 in Lecture note 2) to show

$$(\mathbb{E}[X\mathbf{1}(X > \theta\mathbb{E}[X])])^2 \leq \mathbb{E}[X^2]\mathbb{E}[\mathbf{1}(X > \theta\mathbb{E}[X])^2] \tag{569}$$

$$= \mathbb{E}[X^2]\mathbb{E}[\mathbf{1}(X > \theta\mathbb{E}[X])] \tag{570}$$

$$= \mathbb{E}[X^2]\mathbb{P}(X > \theta\mathbb{E}X). \tag{571}$$

(iii) From (i) and (ii), derive

$$\mathbb{E}[X] \leq \theta\mathbb{E}[X] + \sqrt{\mathbb{E}[X^2]\mathbb{P}(X > \theta\mathbb{E}X)}. \tag{572}$$

Conclude that

$$\mathbb{P}(X > \theta \mathbb{E}X) \geq \frac{(1-\theta)^2 \mathbb{E}[X]^2}{\mathbb{E}[X^2]}. \tag{573}$$

(iv) (Second moment method) From (iii), conclude that

$$\mathbb{P}(X > 0) \geq \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]}. \tag{574}$$

Exercise 2.8 (Kolmogorov's maximal inequality)**.** Let $X_1, X_2, \cdots$ be i.i.d. RVs with $\mathbb{E}[X_i] = 0$. Denote $S_n = X_1 + \cdots + X_n$ and $S_0 = 0$. In this exercise, we will show that

$$\mathbb{P}\left(\max_{1 \leq k \leq n} |S_k| \geq t\right) \leq t^{-2} \mathrm{Var}(S_n). \tag{575}$$

(i) Let $\tau = \inf\{k \geq 0 : |S_k| \geq t\}$ denote the first time that $|S_k|$ exceeds t. Show that

$$\mathbb{E}[S_n^2] \geq \sum_{k=1}^{n} \mathbb{E}[S_n^2 \mathbf{1}(\tau = k)]. \tag{576}$$

(ii) For each $1 \leq k \leq n$, note that $S_k \mathbf{1}(\tau = k)$ and $S_n - S_k = X_{k+1} + \cdots + X_n$ are independent. Deduce that

$$\mathbb{E}[S_k \mathbf{1}(\tau = k)(S_n - S_k)] = 0. \tag{577}$$

(iii) For each $1 \leq k \leq n$, write $S_n^2 = S_k^2 + 2S_k(S_n - S_k) + (S_n - S_k)^2$. Using (ii), show that

$$\mathbb{E}[S_n^2 \mathbf{1}(\tau = k)] \geq \mathbb{E}[(S_k^2 + 2S_k(S_n - S_k))\mathbf{1}(\tau = k)] \tag{578}$$

$$= \mathbb{E}[(S_k^2 \mathbf{1}(\tau = k)] + 2\mathbb{E}[S_k \mathbf{1}(\tau = k)(S_n - S_k)] \tag{579}$$

$$= \mathbb{E}[(S_k^2 \mathbf{1}(\tau = k)] \geq t^2 \mathbb{E}[\mathbf{1}(\tau = k)]. \tag{580}$$

(iv) From (i)-(iii), deduce that

$$\mathbb{E}[S_n^2] \geq t^2 \sum_{k=1}^{n} \mathbb{E}[\mathbf{1}(\tau = k)] = t^2 \mathbb{E}\left[\sum_{k=1}^{n} \mathbf{1}(\tau = k)\right] = t^2 \mathbb{P}(\tau \leq n) = t^2 \mathbb{P}\left(\max_{1 \leq k \leq n} |S_k| \geq t\right). \tag{581}$$

Conclude Kolmogorov's maximal inequality (575).

3. The WLLN and convergence in probability

In this subsection, we prove the weak law of large numbers (Theorem 1.1) and study the notion of convergence in probability. Assuming finite variance for each increment, the weak law is an easy consequence of Chebyshef's inequality.

Theorem 3.1 (WLLN with second moment)**.** *Let $(X_k)_{k \geq 1}$ be i.i.d. RVs with finite mean $\mu < \infty$ and finite variance. Let $S_n = \sum_{k=1}^{n} X_i$, $n \geq 1$. Then for any positive constant $\varepsilon > 0$,*

$$\lim_{n \to \infty} \mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) = 0. \tag{582}$$

PROOF. By Chebyshef's inequality, for any $\varepsilon > 0$ we have

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \leq \frac{\mathrm{Var}(S_n/n)}{\varepsilon^2} = \frac{\mathrm{Var}(X_1)}{n\varepsilon^2}, \tag{583}$$

where the last expression converges to 0 as $n \to \infty$. □

The proof of the full WLLN without the finite second moment assumption needs another technique called 'truncation'. We won't cover this technicality in this course and take Theorem 1.1 for granted.

The weak law of large numbers is the first time that we encounter the notion of 'convergence in probability'. We say a sequence of RVs converge to a constant in probability if the the probability of staying away from that constant goes to zero:

Definition 3.2. Let $(X_n)_{n\geq 1}$ be a sequence of RVs and let $\mu \in \mathbb{R}$ be a constant. We say X_n converges to μ *in probability* if for each $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|X_n - \mu\right| > \varepsilon\right) = 0. \tag{584}$$

Before we proceed further, let us take a moment and think about the definition of convergence in probability. Recall that a sequence of real numbers $(x_n)_{n\geq 0}$ *converges* to x if for each 'error level' $\varepsilon > 0$, there exists a large integer $N(\varepsilon) > 0$ such that

$$|x_n - x| < \varepsilon \quad \forall n \geq N(\varepsilon). \tag{585}$$

If we would like to say that a sequence of RVs $(X_n)_{\geq 0}$ 'converges' to some real number x, how should we formulate this? Since X_n is an RV, $\{|X_n - x| < \varepsilon\}$ is an event. On the other hand, we can also view each x_n as an RV, even though it is a real number. Then we can rewrite (587) as

$$\mathbb{P}(|x_n - x| < \varepsilon) = 1 \quad \forall n \geq N(\varepsilon). \tag{586}$$

For general RVs, requiring $\mathbb{P}(|X_n - x| < \varepsilon) = 1$ for any large n might not be possible. But we can fix any desired level of 'confidence', $\delta > 0$, and require

$$\mathbb{P}(|x_n - x| < \varepsilon) \geq 1 - \delta \tag{587}$$

for sufficiently large n. This is precisely (584).

Example 3.3 (Empirical frequency)**.** Let A be an event of interest. We would like to estimate the unknown probability $p = \mathbb{P}(A)$ by observing a sequence of indpendent experiments. namely, let $(X_k)_{k\geq 0}$ be a sequence of i.i.d. RVs where $X_k = \mathbf{1}(A)$ is the indicator variable of the event A for each $k \geq 1$. Let $\hat{p}_n := (X_1 + \cdots + X_n)/n$. Since $\mathbb{E}[X_1] = \mathbb{P}(A) = p$, by WLLN we conclude that, for any $\varepsilon > 0$,

$$\mathbb{P}\left(|\hat{p}_n - p| > \varepsilon\right) \to 0 \quad \text{as } n \to \infty. \tag{588}$$

▲

Example 3.4 (Polling)**.** Let E_A be the event that a randomly select voter supports candidate A. Using a poll, we would like to estimate $p = \mathbb{P}(E_A)$, which can be understood as the proportion of supporters of candidate A. As before, we observe a sequence of i.i.d. indicator variables $X_k = \mathbf{1}(E_A)$. Let $\hat{p}_n := S_n/n$ be the empirical proportion of supporters of A out of n samples. We know by WLLN that $\hat{p}_n$ converges to p in probability. But if we want to guarantee a certain confidence level α for an error bound ε, how many samples should be take?

By Chebyshef's inequality, we get the following estimate:

$$\mathbb{P}\left(|\hat{p}_n - p| > \varepsilon\right) \leq \frac{\mathrm{Var}(\hat{p}_n)}{\varepsilon^2} = \frac{\mathrm{Var}[X_1]}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}. \tag{589}$$

Note that for the last inequality, we noticed that $X_1 \in [0,1]$ and used Exercise 2.6 (or you can use that for $Y \sim \text{Bernoulli}(p)$, $\mathrm{Var}(Y) = p(1-p) \leq 1/4$). Hence, for instance, if $\varepsilon = 0.01$ and $\alpha = 0.95$, then we would need to set n large enough so that

$$\mathbb{P}\left(|\hat{p}_n - p| > 0.01\right) \leq \frac{10000}{4n} \leq 0.05. \tag{590}$$

This yields $n \geq 50{,}000$. In other words, if we survey at least $n = 50{,}000$ independent voters, then the empirical frequency $\hat{p}_n$ is between $p - 0.01$ and $p + 0.01$ with probability at least 0.95. Still in other words, the true frequency p is between $\hat{p}_n - 0.01$ and $\hat{p}_n + 0.01$ with probability at least 0.95 if $n \geq 50{,}000$. We don't actually need this many samples. We will improve this result later using central limit theorem. ▲

Exercise 3.5 (Monte Carlo integration)**.** Let $(X_k)_{k\geq 1}$ be i.i.d. Uniform($[0,1]$) RVs and let $f : [0,1] \to \mathbb{R}$ be a continuous function. For each $n \geq 1$, let

$$I_n = \frac{1}{n}\left(f(X_1) + f(X_2) + \cdots + f(X_n)\right). \tag{591}$$

(i) Suppose $\int_0^1 |f(x)|\,dx < \infty$. Show that $I_n \to I := \int_0^1 f(x)\,dx$ in probability.

(ii) Further assume that $\int_0^1 |f(x)|^2\,dx < \infty$. Use Chebyshef's inequality to show that

$$\mathbb{P}\left(|I_n - I| \geq a/\sqrt{n}\right) \leq \frac{\mathrm{Var}(f(X_1))}{a^2} = \frac{1}{a^2}\left(\int_0^1 f(x)^2\,dx - I^2\right). \tag{592}$$

Exercise 3.6. Let $(X_n)_{n\geq 0}$ be a sequence of i.i.d. Exp(λ) RVs. Define $Y_n = \min(X_1, X_2, \cdots, X_n)$.
(i) For each $\varepsilon > 0$, show that $\mathbb{P}(|Y_n - 0| > \varepsilon) = e^{-\lambda\varepsilon n}$.
(ii) Conclude that $Y_n \to 0$ in probability as $n \to \infty$.

Example 3.7. For each integer $n \geq 1$, define a RV X_n by

$$X_n = \begin{cases} n & \text{with prob. } 1/n \\ 1/n & \text{with prob. } 1 - 1/n. \end{cases} \tag{593}$$

Then $X_n \to 0$ in probability as $n \to \infty$. Indeed, for each $\varepsilon > 0$,

$$\mathbb{P}(|X_n - 0| > \varepsilon) = \mathbb{P}(X_n > \varepsilon) = 1/n \tag{594}$$

for all $n > 1/\varepsilon$. Hence $\lim_{n\to\infty} \mathbb{P}(|X_n - 0| > \varepsilon) = 0$. However, note that

$$\mathbb{E}[X_n] = 1 + n^{-1} - n^{-2} \to 1 \quad \text{as } n \to \infty. \tag{595}$$

This example indicates that convergence in probability only cares about probability of the event $\mathbb{P}(|X_n - \mathbb{E}[X_n]| > \varepsilon)$ but not the actual value of X_n when that 'bad' event occurs. ▲

Exercise 3.8. Let $X_n \to x$ and $Y_n \to y$ in probability as $n \to \infty$.
(i) Show that for every $\varepsilon > 0$,

$$\mathbb{P}\left(|X_n + Y_n - x - y| > \varepsilon\right) \leq \mathbb{P}\left(|X_n - x| > \varepsilon/2\right) + \mathbb{P}\left(|Y_n - y| > \varepsilon/2\right). \tag{596}$$

Conclude that $X_n + Y_n \to x + y$ in probability as $n \to \infty$.
(ii) Show that for every $\varepsilon > 0$,

$$\begin{aligned} \mathbb{P}\left(|X_nY_n - xy| > \varepsilon\right) &= \mathbb{P}\left(|X_nY_n - X_ny + X_ny - xy| > \varepsilon\right) && (597)\\ &\leq \mathbb{P}\left(|X_n||Y_n - y\| + |y|\|X_n - x| > \varepsilon\right) && (598)\\ &\leq \mathbb{P}\left(|X_n||Y_n - y\| > \varepsilon/2\right) + \mathbb{P}\left(|y|\|X_n - x| > \varepsilon/2\right). && (599) \end{aligned}$$

Conclude that $X_nY_n \to xy$ in probability as $n \to \infty$.
(ii) Suppose $x \neq 0$ and $\mathbb{P}(X_n \neq 0) = 1$ for all $n \geq 1$. Show that for every $\varepsilon > 0$,

$$\begin{aligned} \mathbb{P}\left(\left|\frac{Y_n}{X_n} - \frac{y}{x}\right| > \varepsilon\right) &= \mathbb{P}\left(\left|\frac{Y_n}{X_n} - \frac{y}{X_n} + \frac{y}{X_n} - \frac{y}{x}\right| > \varepsilon\right) && (600)\\ &= \mathbb{P}\left(\frac{1}{|X_n|}\left|Y_n - y\right| + |y|\frac{|X_n - x|}{|X_nx|} > \varepsilon\right) && (601)\\ &= \mathbb{P}\left(\frac{1}{|X_n|}\left|Y_n - y\right| > \varepsilon/2\right) + \mathbb{P}\left(|y|\frac{|X_n - x|}{|X_nx|} > \varepsilon/2\right). && (602) \end{aligned}$$

Conclude that $Y_n/X_n \to y/x$ in probability as $n \to \infty$.

Example 3.9 (Winning strategy)**.** Consider gambling in Vegas for a simple game: after each fair coin flip, one gains twice the bet if heads or loses if tails. There is a simple, always-winning strategy that is banned by Vegas: Double the bet every time you lose and stop after you win the first time.

To analyze this strategy, let $(X_n)_{n\geq 0}$ be a sequence of i.i.d. Uniform($\{-1, 1\}$) variables. Let τ be the first time that $X_n = 1$, the first time you win the game. Suppose you start betting \$1 at the first game, and let Y be your net gain after you exit the game. Since you are doubling the bet every time you lose and you gain twice the bet on your first win,

$$Y = -1 - 2^1 - 2^2 - \cdots - 2^{\tau-1} + 2^{\tau+1}. \tag{603}$$

Then

$$\mathbb{E}[Y \mid \tau = t] = -(1 + 2^1 + 2^2 + \cdots + 2^{t-1}) + 2^{t+1} = -(2^t - 1) + 2^{t+1} = 2^t + 1. \tag{604}$$

Hence by iterated expectation,

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid \tau]] = \mathbb{E}[(2^\tau + 1)] = \sum_{t=1}^{\infty} (2^t + 1)\mathbb{P}(\tau = t) = \sum_{t=1}^{\infty} (2^t + 1)\frac{1}{2^{t-1}}\frac{1}{2} = \sum_{t=1}^{\infty} 1 + 2^{-t} = \infty. \tag{605}$$

So your expected gain is infinity! ▲

Example 3.10 (Coupon collector's problem)**.** Let $(X_t)_{\geq t}$ be a sequence of i.i.d. Uniform$(\{1, 2, \cdots, n\})$ variables. Think of the value of X_t as the label of the coupon you collect at tth trial. We are interested in how many times we need to reveal a new random coupon to collect a full set of n distinct coupons. That is, let

$$\tau^n = \min\{r \geq 1 \mid \#\{X_1, X_2, \cdots, X_r\} = n\}. \tag{606}$$

Because of the possible overlap, we expect n reveals should not get us the full set of n coupons. Indeed,

$$\mathbb{P}(\tau^n = n) = \frac{n-1}{n} \cdot \frac{n-2}{n} \cdots \frac{1}{n} = \frac{n!}{n^n}. \tag{607}$$

Certainly this probability rapidly goes to zero as $n \to \infty$. So we need to reveal more than n coupons. But how many? The answer turns out to be $\tau^n \approx n \log n$. More precisely,

$$\frac{\tau^n}{n \log n} \to 1 \quad \text{as } n \to \infty \text{ in probability.} \tag{608}$$

A change of perspective might help us. Instead of waiting to collect all n coupons, let's progressively collect k distinct coupons for $k = 1$ to n. Namely, for each $1 \leq k \leq n$, define

$$\tau^k = \min\{r \geq 1 \mid \#\{X_1, X_2, \cdots, X_r\} = k\}. \tag{609}$$

So τ^k is the first time that we collect k distinct coupons.

Now consider what has to happen to collect $k+1$ distinct coupons from k distinct coupons? Here is an example. Say at time τ^2 we have coupons $\{1, 3\}$. τ^3 is the first time that we pick up a new coupon from except 1 and 3. This happens with probability $(n-2)/n$ and since each draw is i.i.d.,

$$\tau^3 - \tau^2 \sim \text{Geom}\left(\frac{n-2}{n}\right). \tag{610}$$

A similar reasoning shows

$$\tau^{k+1} - \tau^k \sim \text{Geom}\left(\frac{n-k}{n}\right). \tag{611}$$

So starting from the first coupon, we wait a Geom$(1/n)$ time to get a new coupon, and wait a Geom$(2/n)$ time to get another new coupon, and so on. Note that these geometric waiting times are all independent. So we can decompose τ^n into a sum of independent geometric RVs:

$$\tau^n = \sum_{k=1}^{n-1} (\tau^{k+1} - \tau^k). \tag{612}$$

Then using the estimates in Exercise 3.11, it is straightforward to show that

$$\mathbb{E}[\tau^n] \approx n \log n, \quad \text{Var}(\tau^n) \leq n^2. \tag{613}$$

In Exercise 3.12, we will show (608) using Chebyshef's inequality. ▲

Exercise 3.11. In this exercise, we estimate some partial sums using integral comparison.

(i) For any integer $d \geq 1$, show that

$$\sum_{k=2}^{n} \frac{1}{k^d} \leq \int_1^n \frac{1}{x^d}\, dx \leq \sum_{k=1}^{n-1} \frac{1}{k^d} \tag{614}$$

by considering the upper and lower sum for the Riemann integral $\int_1^n x^{-d}\, dx$.

(ii) Show that

$$\log n \leq \sum_{k=1}^{n-1} \frac{1}{k} \leq 1 + \log(n-1). \tag{615}$$

(iii) Show that for all $d \geq 2$,

$$\sum_{k=1}^{n-1} \frac{1}{k^d} \leq \sum_{k=1}^{\infty} \frac{1}{k^d} \leq 1 + \int_1^{\infty} \frac{1}{x^d}\, dx \leq 2. \tag{616}$$

Exercise 3.12. For each $n \geq 1$, let $X_{1,n}, X_{2,n}, \cdots, X_{n,n}$ be a sequence of independent geometric RVs where $X_{k,n} \sim \text{Geom}((n-k)/n)$. Define $\tau^n = X_{1,n} + X_{2,n} + \cdots + X_{n,n}$.

(i) Show that $\mathbb{E}[\tau^n] = n \sum_{k=1}^{n-1} k^{-1}$. Using Exercise 3.11 (ii), deduce that

$$n \log n \leq \mathbb{E}[\tau^n] \leq n \log(n-1) + n. \tag{617}$$

(ii) Using $\text{Var}(\text{Geom}(p)) = (1-p)/p^2 \leq p^{-2}$ and Exercise 3.11 (iii), show that

$$\text{Var}(\tau^n) \leq n^2 \sum_{k=1}^{n-1} k^{-2} \leq 2n^2. \tag{618}$$

(iii) By Chebyshef's inequality, show that for each $\varepsilon > 0$,

$$\mathbb{P}\left(|\tau^n - \mathbb{E}[\tau^n]| > \varepsilon n \log n\right) \leq \frac{\text{Var}(\tau^n)}{\varepsilon^2 n^2 \log^2 n} \leq \frac{2}{\varepsilon^2 \log^2 n}. \tag{619}$$

Conclude that

$$\frac{\tau^n - \mathbb{E}[\tau^n]}{n \log n} \to 0 \quad \text{as } n \to \infty \text{ in probability.} \tag{620}$$

(iv) By using part (i), conclude that

$$\frac{\tau^n}{n \log n} \to 1 \quad \text{as } n \to \infty \text{ in probability.} \tag{621}$$

Exercise 3.13 (WLLN for RVs with infinite variance)**.** (Optional*) In this exercise, we will prove the WLLN for RVs with infinite variance by using a 'truncation argument'. (Note that we cannot use Chebyshef here.)

Let $(X_k)_{k \geq 1}$ be i.i.d. RVs such that $\mathbb{E}[|X_1|] < \infty$ and $\text{Var}(X_1) \in [0, \infty]$. Let $\mu = \mathbb{E}[X_1]$ and $S_n = \sum_{k=1}^n X_i$, $n \geq 1$. We will show that for any positive constant $\varepsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) = 0. \tag{622}$$

(i) Fix $M \geq 1$. Let $S_n^{\leq M} := \sum_{i=1}^n X_i \mathbf{1}(|X_i| \leq M)$ and $\mu^{\leq M} := \mathbb{E}[X_1 \mathbf{1}(|X_1 \leq M|)]$. Show that $n^{-1} S_n^{\leq M} \to \mu^{\leq M}$ as $n \to \infty$ in probability.

(ii) Show that $\mu^{\leq M} = \mathbb{E}[|X_1| \mathbf{1}(X_1 \leq M)] \to \mu$ as $M \to \infty$. (Hint: Use dominated convergence theorem.)

(iii) Let $S_n^{>M} := \sum_{i=1}^n X_i \mathbf{1}(|X_i| > M)$. Use Markov's inequality and (ii) to show that, for any $\delta > 0$,

$$\mathbb{P}\left(\left|\frac{S_n^{\leq M}}{n}\right| \geq \delta\right) \leq \delta^{-1} \mathbb{E}\left[|X_1| \mathbf{1}(|X_1| > M)\right] \to 0 \quad \text{as } M \to \infty. \tag{623}$$

(iv) Fix $\varepsilon, \delta > 0$, and show the following inequality

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) \leq \mathbb{P}\left(\left|\frac{S_n^{\leq M}}{n} - \mu^{\leq M}\right| \geq \varepsilon/3\right) + \mathbb{P}\left(\left|\frac{S_n^{>M}}{n}\right| \geq \varepsilon/3\right) + \mathbf{1}(|\mu^{\leq M} - \mu| \geq \varepsilon/3). \tag{624}$$

Use (ii)-(iii) to deduce that there exists a large $M' \geq 1$ such that

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) \leq \mathbb{P}\left(\left|\frac{S_n^{\leq M'}}{n} - \mu^{\leq M'}\right| \geq \varepsilon/3\right) + \delta/2. \tag{625}$$

Finally, use (i) to show that there exists a large $n' \geq 1$ such that

$$\mathbb{P}\left(\left|\frac{S_{n'}}{n'} - \mu\right| \geq \varepsilon\right) \leq \delta. \tag{626}$$

Conclude (622).

4. Central limit theorem

Let $(X_t)_{t\geq 0}$ be a sequence of i.i.d. RVs with finite mean μ and variance σ^2. Let $S_n = X_1 + \cdots + X_n$ for $n \geq 1$. We have calculated the mean and variance of the sample mean S_n/n:

$$\mathbb{E}[S_n/n] = \mu, \quad \mathrm{Var}(S_n/n) = \sigma^2/n. \tag{627}$$

Since $\mathrm{Var}(S_n/n) \to 0$ as $n \to \infty$, we expect the sequence of RVs S_n/n to converge its mean μ in probability.

Central limit theorem is a limit theorem for the sample mean with different regime, namely, it describes the 'fluctuation' of the sample mean around its expectation, as $n \to \infty$. For this purpose, we need to standardize the sample mean so that the mean is zero and variance is unit. Namely, let

$$Z_n = \frac{S_n/n - \mu}{\sigma/\sqrt{n}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}, \tag{628}$$

so that

$$\mathbb{E}[Z_n] = 0, \quad \mathrm{Var}(Z_n) = 1. \tag{629}$$

Since the variance is kept at 1, we should not expect the sequence of RVs $(Z_n)_{n\geq 0}$ converge to some constant in probability, as in the law of large number situation. Instead, Z_n should converge to some other RV, if it ever converges in some sense. Central limit theorem states that Z_n converges to the standard normal RV $Z \sim N(0,1)$ 'in distribution'.

Let us state the central limit theorem (Theorem 1.3).

Theorem 4.1 (CLT)**.** *Let $(X_k)_{k\geq 1}$ be i.i.d. RVs and let $S_n = \sum_{k=1}^n X_i$, $n \geq 1$. Suppose $\mathbb{E}[X_1] < \infty$ and $\mathbb{E}[X_1^2] = \sigma^2 < \infty$. Let $Z \sim N(0,1)$ be a standard normal RV and define*

$$Z_n = \frac{S_n - \mu n}{\sigma\sqrt{n}} = \frac{S_n/n - \mu}{\sigma/\sqrt{n}}. \tag{630}$$

Then Z_n converges to Z as $n \to \infty$ in distribution, namely,

$$\lim_{n\to} \mathbb{P}(Z_n \leq z) = \mathbb{P}(Z \leq z). \tag{631}$$

PROOF. First notice that we can assume $\mathbb{E}[X_1] = 0$ without loss of generality (why?). Then $\sigma^2 = \mathrm{Var}(X_1) = \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 = \mathbb{E}[X_1^2]$. Our proof is based on computing moment generating function of Z_n, and we will show that this converges to the MGF of the standard normal. (This is why we learned moment generating function.)

Since the increments X_i's are i.i.d., we have

$$\mathbb{E}[e^{tS_n}] = \mathbb{E}[e^{tX_1}]\mathbb{E}[e^{tX_2}]\cdots\mathbb{E}[e^{tX_n}] = \mathbb{E}[e^{tX_1}]^n. \tag{632}$$

Since we are assuming $\mathbb{E}[X_1] = 0$ and $\mathbb{E}(X_1^2) < \infty$, we have

$$\mathbb{E}[e^{tX_1}] = 1 + \frac{\sigma^2}{2}t^2 + O(t^3), \tag{633}$$

where $O(t^3)$ contains the rest of terms of order $t \geq 3$. Hence

$$\mathbb{E}[e^{tS_n}] = \left(1 + \frac{\sigma^2}{2}t^2 + O(t^3)\right)^n. \tag{634}$$

This yields

$$\mathbb{E}[e^{tZ_n}] = \mathbb{E}[e^{tS_n/(\sigma\sqrt{n})}] = \mathbb{E}[e^{(t/\sigma\sqrt{n})S_n}] \tag{635}$$

$$= \left(1 + \frac{t^2}{2n} + O(n^{-3/2})\right)^n. \tag{636}$$

Recall that

$$\lim_{n\to\infty}\left(1 + \frac{t^2}{2n}\right)^n = e^{t^2/2} = \mathbb{E}[e^{tZ}]. \tag{637}$$

Since the $O(n^{-3/2})$ term vanishes as $n \to \infty$, this shows that $\mathbb{E}[e^{tZ_n}] \to \mathbb{E}[e^{tZ}]$ as $n \to \infty$. Since MGFs determine distribution of RVs (see Theorem 4.14 (ii) in Note 4), it follows that the CDF of Z_n converges to that of Z. □

As a typical application of CLT, we can approximate Binomial(n, p) variables by normal RVs.

Exercise 4.2. Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. Poisson(λ) RVs. Let $S_n = X_1 + \cdots + X_n$.

(i) Let $Z_n = (S_n - n\lambda)/\sqrt{n\lambda}$. Show that as $n \to \infty$, Z_n converges to the standard normal RV $Z \sim N(0,1)$ in distribution.

(ii) Conclude that if $Y_n \sim$ Poisson$(n\lambda)$, then

$$\frac{Y_n - n\lambda}{\sqrt{n\lambda}} \Rightarrow Z \sim N(0,1). \tag{638}$$

(iii) From (ii) deduce that we have the following approximation

$$\mathbb{P}(Y_n \leq x) \approx \mathbb{P}\left(Z \leq \frac{x - n\lambda}{\sqrt{n\lambda}}\right), \tag{639}$$

which becomes more accurate as $n \to \infty$.

Exercise 4.3. Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. Bernoulli(p) RVs. Let $S_n = X_1 + \cdots + X_n$.

(i) Let $Z_n = (S_n - np)/\sqrt{np(1-p)}$. Show that as $n \to \infty$, Z_n converges to the standard normal RV $Z \sim N(0,1)$ in distribution.

(ii) Conclude that if $Y_n \sim$ Binomial(n, p), then

$$\frac{Y_n - np}{\sqrt{np(1-p)}} \Rightarrow Z \sim N(0,1). \tag{640}$$

(iii) From (ii), deduce that have the following approximation

$$\mathbb{P}(Y_n \leq x) \approx \mathbb{P}\left(Z \leq \frac{x - np}{\sqrt{np(1-p)}}\right), \tag{641}$$

which becomes more accurate as $n \to \infty$.

Example 4.4 (Polling revisited)**.** Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. Bernoulli(p) RVs. Denote $\hat{p}_n = n^{-1}(X_1 + \cdots + X_n)$. In Exercise 3.4, we used Chebyshef's inequality to deduce that

$$\mathbb{P}(|\hat{p}_n - p| \leq 0.01) \geq 0.95 \tag{642}$$

whenever $n \geq 50{,}000$. In this example, we will use CLT to improve this lower bound on n.

First, from Exercise 4.3, it is immediate to deduce the following convergence in distribution

$$\frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}} \Rightarrow Z \sim N(0,1). \tag{643}$$

Hence for any $\varepsilon > 0$, we have

$$\mathbb{P}\left(|\hat{p}_n - p| \le \varepsilon\right) = \mathbb{P}\left(\left|\frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}}\right| \le \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) \tag{644}$$

$$\ge \mathbb{P}\left(\left|\frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}}\right| \le 2\varepsilon\sqrt{n}\right) \tag{645}$$

$$\approx \mathbb{P}\left(|Z| \le 2\varepsilon\sqrt{n}\right) = 2\mathbb{P}(0 \le Z \le 2\varepsilon\sqrt{n}), \tag{646}$$

where for the inequality we have used the fact that $p(1-p) \le 1/4$ for all $0 \le p \le 1$. The last expression is at least 0.95 if and only if

$$\mathbb{P}(0 \le Z \le 2\varepsilon\sqrt{n}) \ge 0.475. \tag{647}$$

From the table of standard normal distribution, we know that $\mathbb{P}(0 \le Z \le 1.96) = 0.475$. Hence (647) holds if and only if $2\varepsilon\sqrt{n} \ge 1.96$, or equivalently,

$$n \ge \left(\frac{0.98}{\varepsilon}\right)^2. \tag{648}$$

For instance, $\varepsilon = 0.01$ gives $n \ge 9604$. This is a drastic improvement from $n \ge 50,000$ via Chebyshef.

Exercise 4.5. Let $X_1, Y_1, \cdots, X_n, Y_n$ be i.i.d. Uniform($[0,1]$) RVs. Let

$$W_n = \frac{(X_1 + \cdots + X_n) - (Y_1 + \cdots + Y_n)}{n}. \tag{649}$$

Find a numerical approximation to the quantity

$$\mathbb{P}(|W_{20} - \mathbb{E}[W_{20}]| < 0.001). \tag{650}$$

Exercise 4.6. Let $(\lambda_i)_{i\ge1}$ be a sequence of positive numbers. Let $X_i \sim \text{Poisson}(\lambda_i)$ and assume X_i's are independent. For each $n \ge 1$, let $S_n = X_1 + \cdots + X_n$ and

$$Z_n := \frac{S_n - \mathbb{E}[S_n]}{\sqrt{\text{Var}(S_n)}} = \frac{S_n - \sum_{i=1}^n \lambda_i}{\sqrt{\sum_{i=1}^n \lambda_i}}. \tag{651}$$

We will show that

$$Z_n \underset{n\to\infty}{\Longrightarrow} Z \sim N(0,1) \qquad \Longleftrightarrow \qquad \sum_{n=1}^{\infty} \lambda_n = \infty. \tag{652}$$

(i) Show that $\mathbb{E}[\exp(tS_n)] = \exp\left((e^t - 1)\sum_{i=1}^n \lambda_i\right)$. Deduce that

$$\mathbb{E}\left[\exp(tZ_n)\right] = \exp\left[\left(e^{t/\sqrt{\sum_{i=1}^n \lambda_i}} - 1\right)\sum_{i=1}^n \lambda_i\right] \exp\left[-t\sqrt{\sum_{i=1}^n \lambda_i}\right]. \tag{653}$$

(You may use the fact that if $X \sim \text{Poisson}(\lambda)$, then $\mathbb{E}[\exp(tX)] = \exp(\lambda(e^t - 1))$.)

(ii) For each $n \ge 1$, write $h_n = \left(\sum_{i=1}^n \lambda_i\right)^{-1/2}$. Show that

$$\log\mathbb{E}\left[\exp(tZ_n)\right] = \frac{e^{h_n t} - 1 - h_n t}{h_n^2}. \tag{654}$$

Now show the following implications:

$$\sum_{n=1}^{\infty} \lambda_n = \infty \qquad \Longleftrightarrow \qquad \log\mathbb{E}\left[\exp(tZ_n)\right] \underset{n\to\infty}{\to} \frac{t^2}{2} \quad \forall t \in \mathbb{R} \tag{655}$$

(You may use L'Hospital's rule.)

(iii) Use Theorem 2.5 to deduce (652) from (ii).

5. The SLLN and almost sure convergence

Let $(X_n)_{n\geq 1}$ be i.i.d. RVs with finite mean $\mathbb{E}[X_1] = \mu$ and let $S_n = X_1 + \cdots + X_n$ for all $n \geq 1$. The weak law of large numbers states that the sample mean S_n/n converges to μ in probability, that is,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) = 0 \quad \forall \varepsilon > 0. \tag{656}$$

On the other hand, the Strong Law of Large Numbers (SLLN) tells us that a similar statement holds where the limit is inside the probability bracket. Namely,

$$\mathbb{P}\left(\lim_{n\to\infty}\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) = 0 \quad \forall \varepsilon > 0. \tag{657}$$

If we view the limit on the left hand side as a RV, then (657) in fact states that this limit RV is 0 with probability 1:

$$\mathbb{P}\left(\lim_{n\to\infty}\left|\frac{S_n}{n} - \mu\right| = 0\right) = 1. \tag{658}$$

This is equivalent to the following familiar form of SLLN in Theorem 1.2:

$$\mathbb{P}\left(\lim_{n\to\infty}\frac{S_n}{n} = \mu\right) = 1. \tag{659}$$

Definition 5.1. Let $(X_n)_{n\geq 1}$ be a sequence of RVs and let a be a real number. We say that X_n converges to *a almost surely* (or *with probability* 1) if

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = a\right) = 1. \tag{660}$$

Example 5.2. In this example, we will see that convergence in probability does not necessarily imply convergence with probability 1. Define a sequence of RVs $(T_n)_{n\geq 1}$ as follows. Let $T_1 = 1$, and $T_2 \sim$ Uniform$(\{2,3\})$, $T_3 \sim$ Uniform$(\{4,5,6\})$, and so on. In general, $X_k \sim$ Uniform$\{(k-1)k/2, \cdots, k(k+1)/2\}$ for all $k \geq 2$. Let $X_n = \mathbf{1}$(some T_k takes value n). Think of

$$T_n = n\text{th arrival time of customers} \tag{661}$$

$$X_n = \mathbf{1}(\text{some customer arrives at time } n). \tag{662}$$

Then note that

$$\mathbb{P}(X_1 = 1) = 1, \tag{663}$$

$$\mathbb{P}(X_2 = 1) = \mathbb{P}(X_3 = 1) = 1/2, \tag{664}$$

$$\mathbb{P}(X_4 = 1) = \mathbb{P}(X_5 = 1) = \mathbb{P}(X_6 = 1) = 1/3, \tag{665}$$

and so on. Hence it is clear that $\lim_{n\to\infty}\mathbb{P}(X_n = 1) = 0$. Since X_n is an indicator variable, this yields that $\lim_{n\to\infty}\mathbb{P}(|X_n - 0| > \varepsilon) = 0$ for all $\varepsilon > 0$, that is, X_n converges to 0 in probability. On the other hand, $X_n \to 0$ a.s. means $\mathbb{P}(\lim_{n\to\infty} X_n = 0)$, which implies that $X_n = 0$ for all but finitely many n's. However, $X_n = 1$ for infinitely many n's since customer always arrive after any large time N. Hence X_n cannot converge to 0 almost surely. ▲

Exercise 5.3. Let $(X_n)_{n\geq 1}$ be a sequence of RVs and let a be a real number. Suppose X_n converges to a with probability 1.

(i) Show that

$$\mathbb{P}\left(\lim_{n\to\infty}|X_n - a| \leq \varepsilon\right) = 1 \quad \forall \varepsilon > 0. \tag{666}$$

(ii) Fix $\varepsilon > 0$. Let A_k be the event that $|X_n - a| \le \varepsilon$ for all $n \ge k$. Show that $A_1 \subseteq A_2 \subseteq \cdots$ and

$$\mathbb{P}\left(\lim_{n\to\infty} |X_n - a| \le \varepsilon\right) \le \mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right). \tag{667}$$

(iii) Show that for each $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(|X_n - a| \le \varepsilon\right) \ge \lim_{n\to\infty} \mathbb{P}\left(A_n\right) = \mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) \ge \mathbb{P}\left(\lim_{n\to\infty} |X_n - a| \le \varepsilon\right) = 1. \tag{668}$$

Conclude that $X_n \to a$ in probability.

A typical tool for proving convergence with probability 1 is the following.

Exercise 5.4 (Borel-Cantelli lemma)**.** Let $(A_n)_{n\ge1}$ be a sequence of events such that

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty. \tag{669}$$

We will show that

$$\mathbb{P}\left(A_n \text{ occurs only for finitely many } n\text{'s}\right) = 1. \tag{670}$$

(i) Let $N = \sum_{n=1}^{\infty} \mathbf{1}(A_n)$, which is the number of n's such that A_n occurs. Use Fubini's theorem to show that

$$\mathbb{E}[N] = \mathbb{E}\left[\sum_{n=1}^{\infty} \mathbf{1}(A_n)\right] = \sum_{n=1}^{\infty} \mathbb{E}[\mathbf{1}(A_n)] = \sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty. \tag{671}$$

(ii) Deduce that the RV N must not take ∞ with positive probability. Hence $\mathbb{P}(N < \infty) = 1$, as desired.

Exercise 5.5. Let $(X_n)_{n\ge1}$ be a sequence of RVs and fix $x \in \mathbb{R}$. We will show that $X_n \to x$ a.s. if the tail probabilities are 'summable'. (This is the typical application of the Borel-Cantelli lemma.)

(i) Fix $\varepsilon > 0$. Suppose $\sum_{n=1}^{\infty} \mathbb{P}(|X_n - x| > \varepsilon) < \infty$. Use Borel-Cantelli lemma to deduce that $|X_n - x| > \varepsilon$ for only finitely many n's.

(ii) Conclude that, if $\sum_{n=1}^{\infty} \mathbb{P}(|X_n - x| > \varepsilon) < \infty$ for all $\varepsilon > 0$, then $X_n \to x$ a.s.

Example 5.6. Let $(X_n)_{n\ge0}$ be a sequence of i.i.d. Exp(λ) RVs. Define $Y_n = \min(X_1, X_2, \cdots, X_n)$. Recall that in Exercise 3.6, we have shown that

$$\mathbb{P}(|Y_n - 0| > \varepsilon) = e^{-\lambda\varepsilon n}. \tag{672}$$

for all $\varepsilon > 0$ and that $Y_n \to 0$ in probability as $n \to \infty$. In fact, $Y_n \to 0$ with probability 1. To see this, we note that, for all $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}(|Y_n - 0| > \varepsilon) = \sum_{n=1}^{\infty} e^{-\lambda\varepsilon n} = \frac{e^{-\lambda\varepsilon}}{1 - e^{\lambda\varepsilon}} < \infty. \tag{673}$$

By Borel-Cantelli lemma (or Exercise 5.5), we conclude that $Y_n \to 0$ a.s. ▲

Example 5.7. Let $(X_n)_{n\ge0}$ be a sequence of i.i.d. Uniform($[0,1]$) RVs.

(i) We show that $X_n^{1/n}$ converges to 1 almost surely, as $n \to \infty$. Fix any $\epsilon > 0$. Since $X_n \ge 1$,

$$\begin{aligned} \mathbb{P}(|(X_n)^{1/n} - 1| > \epsilon) &= \mathbb{P}\big((X_n)^{1/n} > (1+\epsilon) \text{ or } (X_n)^{1/n} < (1-\epsilon)\big) = \mathbb{P}((X_n)^{1/n} < (1-\epsilon)) \\ &= \mathbb{P}\big(X_n < (1-\epsilon)^n\big) = (1-\epsilon)^n, \end{aligned} \tag{674}$$

which goes to 0 as $n \to \infty$. Therefore, $X_n^{1/n}$ converges to 1 in probability. However, since $\sum_{n=1}^{\infty}(1-\epsilon)^n < \infty$, we have that $X_n^{1/n}$ also converges to 1 almost surely.

(ii) Define $U_n = \max\{X_1, X_2^2, X_3^3, \dots, X_{n-1}^{n-1}, X_n^n\}$. We show that the sequence U_n converges in probability to 1. For an $\epsilon < 1$ fixed

$$\mathbb{P}(|U_n - 1| \geq \epsilon) = \mathbb{P}(U_n \leq 1-\epsilon) = \mathbb{P}(X_1 \leq 1-\epsilon, X_2^2 \leq 1-\epsilon, \dots, X_n^n \leq 1-\epsilon) \tag{675}$$

$$= \mathbb{P}(X_1 \leq 1-\epsilon)\mathbb{P}(X_2^2 \leq 1-\epsilon)\cdots\mathbb{P}(X_n^n \leq 1-\epsilon) \tag{676}$$

$$= \mathbb{P}(X_1 \leq 1-\epsilon)\mathbb{P}(X_2 \leq (1-\epsilon)^{1/2})\cdots\mathbb{P}(X_n \leq (1-\epsilon)^{1/n}) \tag{677}$$

$$= (1-\epsilon)\cdot(1-\epsilon)^{1/2}\cdots(1-\epsilon)^{1/n} = (1-\epsilon)^{1+1/2+\cdots+1/n} \to 0, \tag{678}$$

since $1 + 1/2 + \cdots + 1/n \to \infty$, as $n \to \infty$.

(iii) Define $V_n = \max\{X_1, X_2^{2^2}, X_3^{3^2}, \dots, X_{n-1}^{(n-1)^2}, X_n^{n^2}\}$. Does the sequence V_n converges in probability to 1? Similarly as in the previous part, for a fixed $\epsilon < 1$ we have

$$\mathbb{P}(|U_n - 1| \geq \epsilon) = (1-\epsilon)^{1+1/2^2+\cdots+1/n^2}, \tag{679}$$

which doesn't converge to zero (but to a positive number), since $1 + 1/2^2 + \cdots + 1/n^2$ is a convergent series. Therefore, V_n doesn't converge to zero in probability. Hence it also doesn't converge to 0 a.s.

▲

Now we prove the strong law of large numbers. The proof of full statement (Theorem 1.2) with finite second moment assumption has extra technicality, so here we prove the result under a stronger assumption of finite fourth moment.

Theorem 5.8 (SLLN with fourth moment)**.** *Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. RVs such that $\mathbb{E}[X_n^4] < \infty$. Let $S_n = X_1 + \cdots + X_n$ for all $n \geq 1$. Then S_n/n converges to $\mathbb{E}[X_1]$ with probability 1.*

PROOF. Our aim is to show that

$$\sum_{n=1}^{\infty} \mathbb{E}[(S_n/n)^4] < \infty. \tag{680}$$

Then by Borel-Cantelli lemma, $(S_n/n)^4$ converges to 0 with probability 1. Hence S_n/n converges to 0 with probability 1, as desired.

For a preparation, we first verify that we have finite first and second moments for X_1. It is easy to verify the inequality $|x| \leq 1 + x^4$ for all $x \in \mathbb{R}$, so we have

$$\mathbb{E}[|X_1|] \leq 1 + \mathbb{E}[X_1^4] < \infty. \tag{681}$$

Hence $\mathbb{E}[X_1]$ exists. By shifting, we may assume that $\mathbb{E}[X_1] = 0$. Similarly, it holds that $x^2 \leq c + x^4$ for all $x \in \mathbb{R}$ if $c > 0$ is large enough. Hence $\mathbb{E}[X_1^2] < \infty$.

Note that

$$\mathbb{E}\left[S_n^4\right] = \mathbb{E}\left[\left(\sum_{k=1}^{n} X_k\right)^4\right] = \mathbb{E}\left[\sum_{1\leq i,j,k,\ell\leq n} X_i X_j X_k X_\ell\right] = \sum_{1\leq i,j,k,\ell\leq n} \mathbb{E}\left[X_i X_j X_k X_\ell\right]. \tag{682}$$

Note that by independence and the assumption that $\mathbb{E}[X_1] = 0$, $\mathbb{E}[X_i X_j X_k X_\ell] = 0$ if at least one of the four indices does not repeat. For instance,

$$\mathbb{E}[X_1 X_2^3] = \mathbb{E}[X_1]\mathbb{E}[X_2^3] = 0, \tag{683}$$

$$\mathbb{E}[X_1 X_2^2 X_3] = \mathbb{E}[X_1]\mathbb{E}[X_2^2]\mathbb{E}[X_3] = 0. \tag{684}$$

Hence by collecting terms based on number of overlaps, we have

$$\sum_{1\leq i,j,k,\ell\leq n} \mathbb{E}\left[X_i X_j X_k X_\ell\right] = \sum_{i=1}^{n} \mathbb{E}[X_i^4] + \binom{4}{2} \sum_{1\leq i<j\leq n} \mathbb{E}[X_i^2]\mathbb{E}[X_j^2] \tag{685}$$

$$= n\mathbb{E}[X_1^4] + 3n(n-1)\mathbb{E}[X_1^2]^2. \tag{686}$$

Thus for all $n \geq 1$,

$$\mathbb{E}\left[(S_n/n)^4\right] = \frac{n\mathbb{E}[X_1^4] + 3n(n-1)\mathbb{E}[X_1^2]^2}{n^4} \leq \frac{n^2\mathbb{E}[X_1^4] + 3n^2\mathbb{E}[X_1^2]^2}{n^4} = \frac{\mathbb{E}[X_1^4] + 3\mathbb{E}[X_1^2]}{n^2}. \tag{687}$$

Summing over all n, this gives

$$\sum_{n=1}^{\infty} \mathbb{E}[(S_n/n)^4] \leq (\mathbb{E}[X_1^4] + 3\mathbb{E}[X_1^2]) \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty. \tag{688}$$

Hence by Borell-Cantelli lemma, we conclude that $(S_n/n)^4$ converges to 0 with probability 1. The same conclusion holds for S_n/n. This shows the assertion. □

CHAPTER 7

Elementary Stochastic Processes

How can we use sequence of RVs to model real life situations? Say we would like to model the USD price of bitcoin. We could observe the actual price at every hour and record it by a sequence of real numbers $x_1, x_2, \cdots$. However, it is more interesting to build a 'model' that could predict the price of bitcoin at time t, or at least give some meaningful insight on how the actual bitcoin price behaves over time. Since there are so many factors affecting its price at every time, it might be reasonable that its price at time t should be given by a certain RV, say X_t. Then our sequence of predictions would be a sequence of RVs, $(X_t)_{t\geq 0}$. This is an example of what is called a *stochastic process*. Here 'process' means that we are not interested in just a single RV, that their sequence as a whole: 'stochastic' means that the way the RVs evolve in time might be random.

In this note, we will be studying three elementary stochastic processes: 1) Bernoulli process, 2) Poisson process, and 3) discrete-time Markov chain.

1. The Bernoulli process

1.1. Definition of Bernoulli process. Let $(X_t)_{t\geq 1}$ be a sequence of i.i.d. Bernoulli(p) variables. This is the *Bernoulli process* with parameter p, and that's it. Considering how simple it is conceptually, we can actually ask a lot of interesting questions about it.

First we envision this as a model of customers arriving at a register. Suppose a clerk rings a bell whenever she is done with her current customer or ready to take the next customer. Upon each bell ring, a customer arrives with probability p or no customer gets there with probability $1-p$, independently at each time. Then we can think of the meaning of X_t as

$$X_t = \mathbf{1}(\text{a customer arrives at the register after } t \text{ bell rings}). \tag{689}$$

To simplify terminology, let 'time' be measured by a nonnegative integer $t \in \mathbb{Z}_{\geq 0}$: time t means the time right after tth bell ring. Here are some of the *observables* for this process that we are interested in:

$$S_n = X_1 + \cdots + X_n = \#(\text{customers arriving at the register up to time } n) \tag{690}$$

$$T_i = \text{time that the } i\text{th customer arrives.} \tag{691}$$

$$\tau_i = T_i - T_{i-1} = \text{the inter-arrival time between the } i-1\text{st and } i\text{th customer.} \tag{692}$$

We also define $\tau_1 = T_1$. See Figure 1 for an illustration.

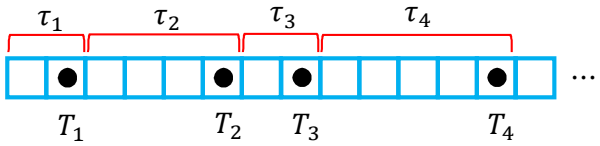

FIGURE 1. Illustration of Bernoulli process. First four customers arrive at times $T_1 = 2$, $T_2 = 6$, $T_3 = 8$, and $T_4 = 13$. The inter-arrival times are $\tau_1 = 2$, $\tau_2 = 4$, $\tau_3 = 2$, and $\tau_4 = 5$. There are $S_7 = 2$ customers up to time $t = 7$.

Exercise 1.1 (Independence)**.** Let $(X_t)_{t\geq 1}$ be a Bernoulli process with parameter p. Show the following.

(i) Let U and V be the number of customers at times $t \in \{1,2,\cdots,5\}$ and $t \in \{6,7,\cdots,10\}$, respectively. Show that U and V are independent.

(ii) Let U and V be the first odd and even time that a customer arrives, respectively. Show that U and V are independent.

(iii) Let S_5 be the number of customers up to time $t = 5$ and let $\tau_3 = T_3 - T_2$ be the inter-arrival time between the second and third customers. Are S_5 and τ_3 independent?

Exercise 1.2. Let $(X_t)_{t\geq 1}$ be a Bernoulli process with parameter p.

(i) Show that $S_n \sim \text{Binomial}(n,p)$.

(ii) Show that $T_1 \sim \text{Geom}(p)$.

(iii) Use conditioning on T_1 to show that $\tau_2 \sim \text{Geom}(p)$ and it is independent of τ_1.

(iv) Use conditioning on T_{k-1} to show that $\tau_k \sim \text{Geom}(p)$ and it is independent of $\tau_1, \tau_2, \cdots, \tau_{k-1}$. Conclude that τ_i's are i.i.d. with $\text{Geom}(p)$ distribution.

Let $(X_t)_{t\geq 1}$ be a Bernoulli process with parameter p. If we discard the first 5 observations and start the process at time $t = 6$, then the new process $(X_t)_{t\geq 6}$ is still a Bernoulli process with parameter p. Moreover, the new process is independent on the past RVs $X_1, X_2, \cdots, X_5$. The following exercise generalizes this observation.

Exercise 1.3. Let $(X_t)_{t\geq 1}$ be a Bernoulli process with parameter p. Show the following.

(i) (Renewal property of Bernoulli RV) For any integer $k \geq 1$, $(X_t)_{t\geq k}$ is a Bernoulli process with parameter p and it is independent from $X_1, X_2, \cdots, X_{k-1}$.

(ii) (Memoryless property of Geometric RV) For any integer $k \geq 1$, let $\bar{T}$ be the first time that a customer arrives after time $t = k$. Show that $\bar{T} - k \sim \text{Geom}(p)$ and it is independent from $X_1, X_2, \cdots, X_k$. (hint: use part (i))

Example 1.4 (Renewal property at a random time)**.** Let $(X_t)_{t\geq 1}$ be a Bernoulli process with parameter p. Suppose N is the first time that we see two consecutive customers, that is,

$$N = \min\{k \geq 2 \,|\, X_{k-1} = X_k = 1\}. \tag{693}$$

Then what is the probability $\mathbb{P}(X_{N+1} = X_{N+2} = 0)$ that no customers arrive at times $t = N+1$ and $t = N+2$? Intuitively, what's happening after time $t = N$ should be independent from what happened up to time $t = N$, so we should have $\mathbb{P}(X_{N+1} = X_{N+2} = 0) = (1-p^2)$. However, this is not entirely obvious since N is a random time.

Observe that the probability $\mathbb{P}(X_{N+1} = X_{N+2} = 0)$ depends on more than two source of randomness: N, X_{N+1}, and X_{N+2}. Our principle to handle this kind of situation was to use conditioning:

$$\mathbb{P}(X_{N+1} = X_{N+2} = 0) = \sum_{n=1}^{\infty} \mathbb{P}(X_{n+1} = X_{n+2} = 0 \,|\, N = n)\mathbb{P}(N = n) \tag{694}$$

$$= \sum_{n=1}^{\infty} \mathbb{P}(X_{n+1} = X_{n+2} = 0)\mathbb{P}(N = n) \tag{695}$$

$$= \sum_{n=1}^{\infty} (1-p)^2 \mathbb{P}(N = n) = (1-p)^2 \sum_{n=1}^{\infty} \mathbb{P}(N = n) = (1-p)^2. \tag{696}$$

Note that for the second equality we have used the renewal property of the Bernoulli process, namely, $(X_t)_{t\geq n+1}$ is a Bernoulli process with parameter p that is independent of $X_1, \cdots, X_n$, and the fact that the event $\{N = n\}$ is completely determined by the RVs $X_1, \cdots, X_n$.

Example 1.5 (Alternative definition of Bernoulli process)**.** Recall that if $(X_t)_{t\geq 1}$ is a sequence of i.i.d. Bernoulli(p) RVs, then the sequence of inter-arrival times $(\tau_k)_{k\geq 1}$ is i.i.d. with $\text{Geom}(p)$ distribution. In this example, we will show that the converse is true. In other words, we give an alternative definition of the Bernoulli processes in terms of the inter-arrival times τ_k, instead of the indicators X_t.

Let $(\tau_k)_{k\geq 0}$ be a sequence of i.i.d. Geom(p) variables. Our interpretation is that

$$\tau_k = (\text{location of the } k\text{th ball}) - (\text{location of the } k-1\text{st ball}). \tag{697}$$

So if we denote by T_k the location of the kth ball, then

$$T_k = \tau_1 + \tau_2 + \cdots + \tau_k. \tag{698}$$

Now define a sequence $(X_t)_{t\geq 0}$ of indicator RVs by

$$X_t = \mathbf{1}(\ T_k = 1 \text{ for some } k \geq 1). \tag{699}$$

Our claim is that X_t's are i.i.d. Bernoulli(p) RVs, so that $(X_t)_{t\geq 1}$ is a BP(p).

We show the claim by a strong induction. That is, suppose $X_1, \cdots, X_t$ are i.i.d. Bernoulli(p) RVs. Then we show $X_{t+1} \sim$ Bernoulli(p) and it is independent of all previous X_i's. To this end, fix a sequence of 0's and 1's, $(x_1, x_2, \cdots, x_t) \in \{0,1\}^t$. This is a particular sample path we observe from the first t boxes. Let k be the number of 1's and let s be the largest such that $x_s = 1$ among this sequence. Under this event, we will have a ball at box $t+1$ if and only if $T_{k+1} = T_k + \tau_{k+1} = s + \tau_{k+1} = t+1$. Hence we have

$$\mathbb{P}(X_{t+1} = 1 \,|\, X_1 = x_1, X_2 = x_2, \cdots, X_t = x_t) = \mathbb{P}(\tau_{k+1} = t+1-s \,|\, \tau_{k+1} \geq t+1-s) \tag{700}$$

$$= \frac{\mathbb{P}(\tau_{k+1} = t+1-s)}{\mathbb{P}(\tau_{k+1} \geq t+1-s)} = \frac{(1-p)^{t+1-s}p}{(1-p)^{t+1-s}} = p. \tag{701}$$

Since X_{t+1} is a 0-1 RV, this also shows that

$$\mathbb{P}(X_{t+1} = 0 \,|\, X_1 = x_1, X_2 = x_2, \cdots, X_t = x_t) = 1-p. \tag{702}$$

Since $(x_1, x_1, \cdots, x_t) \in \{0,1\}^t$ was arbitrary, this shows that X_{t+1} is independent of $X_1, \cdots, X_t$. Furthermore, iterated expectation shows that $\mathbb{P}(X_{t+1} = 1) = p$:

$$\mathbb{P}(X_{t+1} = 1) = \mathbb{E}_{X_t}\mathbb{E}_{X_{t-1}} \cdots \mathbb{E}_{X_1}[\mathbb{P}(X_{t+1} = 1 \,|\, X_1, X_2, \cdots, X_t)] = \mathbb{E}_{X_t}\mathbb{E}_{X_{t-1}} \cdots \mathbb{E}_{X_1}[p] = p. \tag{703}$$

This completes the induction. ▲

1.2. Splitting, merging, and limit theorems for BPs.

Example 1.6 (Splitting and merging of Bernoulli processes)**.** Let $(X_t)_{t\geq 1}$ be a Bernoulli process with parameter p. Let us flip an independent probability $q \in [0,1]$ coin at every t, and define

$$Y_t = X_t \mathbf{1}(\text{coin at time } t \text{ lands heads}) \tag{704}$$

$$Z_t = X_t \mathbf{1}(\text{coin at time } t \text{ lands tails}). \tag{705}$$

Moreover, we have

$$X_t = Y_t + Z_t. \tag{706}$$

Note that $(Y_t)_{t\geq 1}$ and $(Z_t)_{t\geq 1}$ are also Bernoulli processes with parameters pq and $p(1-q)$, respectively. In other words, we splitted the Bernoulli process $(X_t)_{t\geq 1}$ with parameter p into two Bernoulli processes with parameters pq and $p(1-q)$. However, note that the processes Y_t and Z_t are not independent.

Conversely, let $(Y_t)_{t\geq 1}$ and $(Z_t)_{t\geq 1}$ be *independent* Bernoulli processes with parameters p and q, respectively. Is it possible to merge them into a single Bernoulli process? Indeed, we define

$$X_t = \mathbf{1}(Y_t = 1 \text{ or } Z_t = 1). \tag{707}$$

Then $\mathbb{P}(X_t = 1) = 1 - \mathbb{P}(Y_t = 0)\mathbb{P}(Z_t = 0) = 1 - (1-p)(1-q) = p + q - pq$. By independence, X_t is a Bernoulli process with parameter $p + q - pq$.

Exercise 1.7. A transmitter sends a message every 10 minutes, and a receiver successfully obtains each message independently with probability p. Furthermore, each message is of size 1 or 2MB, independently in t with equal probability. Parameterize the time so that "time t" means "after $10t$ minutes". Define random variables

$$X_t = \mathbf{1}(\text{reciever obtains a message succesfully at time } t) \tag{708}$$

$$Y_t = \mathbf{1}(\text{reciever obtains a message of size 2MB succesfully at time } t). \tag{709}$$

(i) Verify that $(X_t)_{t\geq 1}$ is a BP(p).
(ii) Show that $(Y_t)_{t\geq 1}$ is a BP($p/2$).
(iii) What is the expected number of messages of size 2MB received successfully by time 10?
(iv) What is the expected total size of messages received successfully by time 10?

Let $\tau_i \sim \text{Geom}(p)$ for $i \geq 0$ and let $N \sim \text{Geom}(q)$. Suppose all RVs are independent. Let $Y = \sum_{k=1}^{N} \tau_k$. In Exercise 4.24, we have shown that $Y \sim \text{Geom}(pq)$ using MGFs. In the following exercise, we show this by using splitting of Bernoulli processes.

Exercise 1.8 (Sum of geometric number of geometric RVs)**.** Let $(X_t)_{t\geq 0}$ be Bernoulli process of parameter p. Give each ball color Blue and Red independently with probability q and $1-q$, respectively. Let $X_t^B = \mathbf{1}(\text{there is a blue ball in box } t)$.

(i) Show that $(X_t^B)_{t\geq 1}$ is a Bernoulli process of parameter pq.
(ii) Let T_1^B be the location of first blue ball. Show that $T_1^B \sim \text{Geom}(pq)$.
(iii) Let N denote the total number of balls (blue or red) in the first T_1^B boxes. Show that $N \sim \text{Geom}(q)$.
(iv) Let T_k be the location of kth ball, and let $\tau_k = T_k - T_{k-1}$. Show that τ_k's are i.i.d. Geom(p) RVs and they are independent of N. Lastly, show the identity

$$T_1^B = \sum_{k=1}^{N} \tau_k. \tag{710}$$

Hence the sum of geometric (N) number of geometric RVs (τ_k's) is distributed as another geometric RV (T_1^B).

Example 1.9 (Applying limit theorems to BP)**.** Let $(X_t)_{t\geq 1}$ be a Bernoulli process with parameter p. Let T_k be the the smallest integer m such that $X_1 + \cdots + X_m = k$, that is, the location of kth ball. Let $\tau_i = T_i - T_{i-1}$ for $i \geq 2$ and $\tau_0 = T_1$ be the inter-arrival times. Then

$$T_k = T_1 + (T_2 - T_1) + (T_3 - T_2) + \cdots + (T_k - T_{k-1}) \tag{711}$$

$$= \tau_1 + \tau_2 + \cdots + \tau_k. \tag{712}$$

Note that the τ_i's are i.i.d. Geom(p) variables. Hence we can apply all limit theorems to T_k to bound/approximate probabilities associated to it.

To begin, recall that $\mathbb{E}(\tau_i) = 1/p$ and $\text{Var}(\tau_i) = (1-p)/p^2 < \infty$. Hence

$$\mathbb{E}(T_k) = k/p, \qquad \text{Var}(T_k) = \frac{(1-p)k}{p^2}. \tag{713}$$

If we apply SLLN to T_k, we conclude that

$$\mathbb{P}\left(\lim_{k\to\infty} \frac{T_k}{k} = \frac{1}{p}\right) = 1. \tag{714}$$

So the line $y = x/p$ is the 'best fitting line' that explains the data points (k, T_k) (in the sense of linear regression). So we know that $1/p$ is a very good guess for T_k/k, which becomes more accurate as $k \to \infty$.

On the other hand, CLT describes how the sample mean T_k/k fluctuates around its mean $1/p$ as $k \to \infty$. The theorem says that as $k \to \infty$,

$$\frac{T_k - k/p}{\sqrt{k}\sqrt{(1-p)/p^2}} \Rightarrow Z \sim N(0,1). \tag{715}$$

What is this statement good for?

Lets take a concrete example by saying $p = 1/2$ and $k = 100$. Then $\mathbb{E}(T_{100}) = 200$ and $\text{Var}(T_{100}) = 200$. Hence we expect the probability $\mathbb{P}(T_k \geq 250)$ to be very small. For this kind of tail probability estimation, we so far have three devices: Markov's and Chebyshef's inequality, and CLT itself.

First, Markov says

$$\mathbb{P}(T_{100} \geq 250) \leq \frac{\mathbb{E}(T_{100})}{250} = \frac{200}{250} = \frac{4}{5} = 0.8. \tag{716}$$

So this bound is not very useful here. Next, Chebyshef says

$$\mathbb{P}(|T_{100} - 200| \geq 50) \leq \frac{\mathrm{Var}(T_{100})}{50^2} = \frac{200}{2500} = 0.08. \tag{717}$$

Moreover, an implication of CLT is that the distribution of T_k becomes more symmetric about its mean, so the probability on the left hand side is about twice of what we want.

$$\mathbb{P}(T_{100} \geq 250) \approx \frac{1}{2}\mathbb{P}(|T_{100} - 200| \geq 50) \leq 0.04. \tag{718}$$

So Chebyshef gives a much better bound.

But the truth is, the probability $\mathbb{P}(T_{100} \geq 250)$ in fact is extremely small. To see this, we apply CLT to get

$$\mathbb{P}(T_{100} \geq 250) = \mathbb{P}\left(\frac{T_{100} - 200}{\sqrt{200}} \geq \frac{50}{10\sqrt{2}}\right) \approx \mathbb{P}(Z \geq 3.5355). \tag{719}$$

From the table for standard normal distribution, we know that $\mathbb{P}(Z \geq 1.96) = 0.025$ and $\mathbb{P}(Z \geq 2.58) = 0.005$. Hence The probability on the right hand side even smaller than these values.

2. The Poisson process

2.1. Definition of Poisson process. An *arrival* process is a sequence of strictly increasing RVs $0 < T_1 < T_2 < \cdots$. For each integer $k \geq 1$, its kth *inter-arrival time* is defined by $\tau_k = T_k - T_{k-1}\mathbf{1}(k \geq 2)$. For a given arrival process $(T_k)_{k\geq 1}$, the associated *counting process* $(N(t))_{t\geq 0}$ is defined by

$$N(t) = \sum_{k=1}^{\infty} \mathbf{1}(T_k \leq t) = \#(\text{arrivals up to time } t). \tag{720}$$

Note that these three processes (arrival times, inter-arrival times, and counting) determine each other:

$$(T_k)_{k\geq 1} \Longleftrightarrow (\tau_k)_{k\geq 1} \Longleftrightarrow (N(t))_{t\geq 0}. \tag{721}$$

Exercise 2.1. Let $(T_k)_{k\geq 1}$ be any arrival process and let $(N(t))_{t\geq 0}$ be its associated counting process. Show that these two processes determine each other by the following relation

$$\{T_n \leq t\} = \{N(t) \geq n\}. \tag{722}$$

In words, nth customer arrives by time t if and only if at least n customers arrive up to time t.

Now we define Poisson process.

Definition 2.2 (Poisson process)**.** An arrival process $(T_k)_{k\geq 1}$ is a *Poisson process of rate* λ if its inter-arrival times are i.i.d. Exp(λ) RVs.

Exercise 2.3. Let $(T_k)_{k\geq 1}$ be a Poisson process with rate λ. Show that $\mathbb{E}[T_k] = k/\lambda$ and $\mathrm{Var}(T_k) = k/\lambda^2$. Furthermore, show that $T_k \sim \mathrm{Erlang}(k, \lambda)$, that is,

$$f_{T_k}(z) = \frac{\lambda^k z^{k-1} e^{-\lambda z}}{(k-1)!}. \tag{723}$$

The following exercise explains what is 'Poisson' about the Poisson process.

Exercise 2.4. Let $(T_k)_{k\geq 1}$ be a Poisson process with rate λ and let $(N(t))_{t\geq 0}$ be the associated counting process. We will show that $N(t) \sim \mathrm{Poisson}(\lambda t)$.

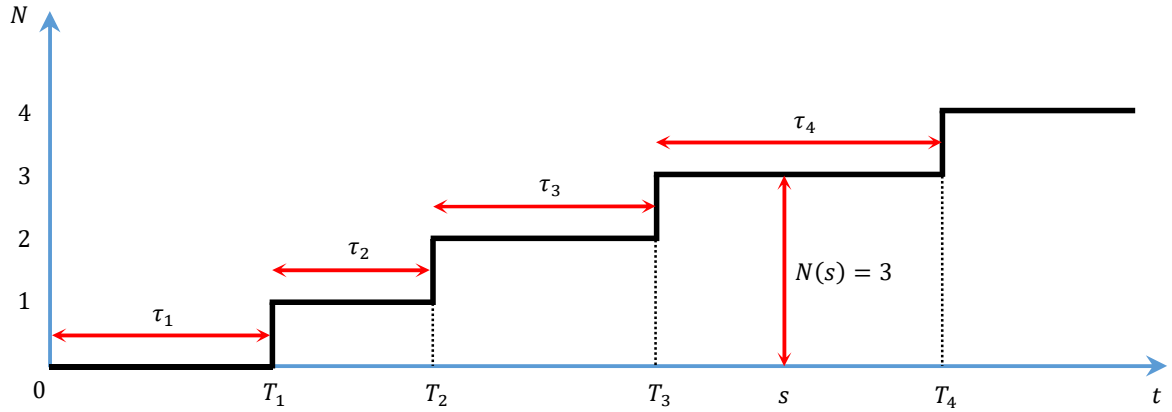

FIGURE 2. Illustration of a continuous-time arrival process $(T_k)_{k\geq 1}$ and its associated counting process $(N(t))_{t\geq 0}$. τ_k's denote inter-arrival times. $N(t) \equiv 3$ for $T_3 < t \leq T_4$.

(i) Using the relation $\{T_n \leq t\} = \{N(t) \geq n\}$ and Exercise 2.3, show that

$$\mathbb{P}(N(t) \geq n) = \mathbb{P}(T_n \leq t) = \int_0^t \frac{\lambda^n z^{n-1} e^{-\lambda z}}{(n-1)!}\, dz. \tag{724}$$

(ii) Let $G(t) = \sum_{m=n}^{\infty} (\lambda t)^m e^{-\lambda t}/m! = \mathbb{P}(\text{Poisson}(\lambda) \geq n)$. Show that

$$\frac{d}{dt} G(t) = \frac{\lambda^n t^{n-1} e^{-\lambda t}}{(n-1)!} = \frac{d}{dt} \mathbb{P}(T_n \leq t). \tag{725}$$

Conclude that $G(t) = \mathbb{P}(T_n \leq t)$.

(iii) From (i) and (ii), conclude that $N(t) \sim \text{Poisson}(\lambda t)$.

2.2. Memoryless property of PP. The choice of exponential inter-arrival times is special due to the following 'memoryless property' of exponential RVs.

Exercise 2.5 (Memoryless property of exponential RV)**.** A continuous positive RV X is say to have *memoryless property* if

$$\mathbb{P}(X \geq t_1 + t_2) = \mathbb{P}(X \geq t_1)\mathbb{P}(X \geq t_2) \quad \forall x_1, x_2 \geq 0. \tag{726}$$

(i) Show that (726) is equivalent to

$$\mathbb{P}(X \geq t_1 + t_2 \,|\, X \geq t_2) = \mathbb{P}(X \geq t_1) \quad \forall x_1, x_2 \geq 0. \tag{727}$$

(ii) Show that exponential RVs have memoryless property.

(iii) Suppose X is continuous, positive, and memoryless. Let $g(t) = \log \mathbb{P}(X \geq t)$. Show that g is continuous at 0 and

$$g(x + y) = g(x) + g(y) \quad \text{for all } x, y \geq 0. \tag{728}$$

Using the following exercise, conclude that X must be an exponential RV.

Exercise 2.6. Let $g : \mathbb{R}_{\geq 0} \to \mathbb{R}$ be a function with the property that $g(x + y) = g(x) + g(y)$ for all $x, y \geq 0$. Further assume that g is continuous at 0. In this exercise, we will show that $g(x) = cx$ for some constant c.

(i) Show that $g(0) = g(0 + 0) = g(0) + g(0)$. Deduce that $g(0) = 0$.

(ii) Show that for all integers $n \geq 1$, $g(n) = ng(1)$.

(iii) Show that for all integers $n, m \geq 1$,

$$ng(1) = g(n \cdot 1) = g(m(n/m)) = mg(n/m). \tag{729}$$

Deduce that for all nonnegative rational numbers r, we have $g(r) = rg(1)$.

(iv) Show that g is continuous.

(v) Let x be nonnegative real number. Let r_k be a sequence of rational numbers such that $r_k \to x$ as $k \to \infty$. By using (iii) and (iv), show that

$$g(x) = g\left(\lim_{k\to\infty} r_k\right) = \lim_{k\to\infty} g(r_k) = g(1) \lim_{k\to\infty} r_k = x \cdot g(1). \tag{730}$$

Given a Poisson process, we can restart it at any given time t. Then the first arrival time after t is simply the remaining inter-arrival time after time t. By memoryless property of exponential RVs, we see that this remaining time is also an exponential RV that is independent of what have happend so far. We will show this in the following proposition. The proof is essentially a Poisson version of Exercise 1.5.

Proposition 2.7 (Memoryless property of PP)**.** *Let $(T_k)_{k\geq 1}$ be a Poisson process of rate λ and let $(N(t))_{t\geq 0}$ be the associated counting process.*

(i) *For any $t \geq 0$, let $Z(t) = \inf\{s > 0 : N(t+s) > N(t)\}$ be the waiting time for the first arrival after time t. Then $Z(t) \sim \mathrm{Exp}(\lambda)$ and it is independent of the process up to time t.*

(ii) *For any $s \geq 0$, $(N(t+s) - N(t))_{s\geq 0}$ is the counting process of a Poisson process of rate λ, which is independent of the process $(N(u))_{t\leq u}$.*

PROOF. We first show (ii). Note that

$$T_{N(t)} \leq t < T_{N(t)+1}. \tag{731}$$

Hence we may write

$$Z(t) = T_{N(t)+1} - t = \tau_{N(t)+1} - (t - T_{N(t)}). \tag{732}$$

Namely, $Z(t)$ is the remaining portion of the $N(t)+1$st inter-arrival time $\tau_{N(t)+1}$ after we waste the first $t - T_{N(t)}$ of it. (See Figure 5).

Now consider restarting the arrival process $(T_k)_{k\geq 0}$ at time t. The first inter-arrival time is $T_{N(t)+1} - t = Z(t)$, which is $\mathrm{Exp}(\lambda)$ and independent from the past by (i). The second inter-arrival time is $T_{N(t)+2} - T_{N(t)+1}$, which is $\mathrm{Exp}(\lambda)$ and is independent of everything else by assumption. Likewise, the following inter-arrival times for this restarted arrival process are i.i.d. $\mathrm{Exp}(\lambda)$ variables. This shows (ii).

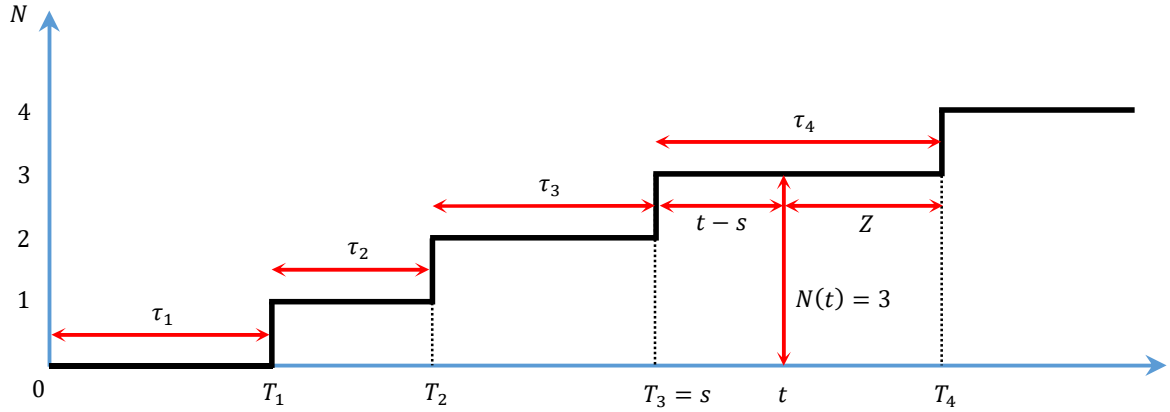

FIGURE 3. Assuming $N(t) = 3$ and $T_3 = s \leq t$, we have $Z = \tau_4 - (t - s)$. By memoryless property of exponential RV, Z follows $\mathrm{Exp}(\lambda)$ on this conditioning.

Next, we show (i). Let E by any event for the counting process $(N(s))_{0\leq s\leq t}$ up to time t. In order to show the remaining waiting time $Z(t)$ and the past process up to time t are independent and $Z(t) \sim \mathrm{Exp}(\lambda)$, we want to show that

$$\mathbb{P}(Z(t) \geq x \,|\, (N(s))_{0\leq s\leq t} \in E) = \mathbb{P}(Z(t) \geq x) = e^{-\lambda x}. \tag{733}$$

for any $x \geq 0$. To this end,

As can be seen from (2.2), $Z(t)$ depends on three random variables: $\tau_{N(t)+1}$, $N(t)$, and $T_{N(t)}$. To show , we argue by conditioning the last two RVs and use iterated expectation. Using 2.2, note that

$$\mathbb{P}\left(Z(t) \geq x \,\middle|\, (N(s))_{0\leq s\leq t} \in E,\, N(t) = n,\, T_{N(t)} = u\right) \tag{734}$$

$$= \mathbb{P}\left(\tau_{n+1} - (t-s) \geq x \,\middle|\, (N(s))_{0\leq s\leq t} \in E,\, N(t) = n,\, T_n = u\right) \tag{735}$$

$$= \mathbb{P}\left(\tau_{n+1} - (t-s) \geq x \,\middle|\, (N(s))_{0\leq s\leq t} \in E,\, T_{n+1} > t,\, T_n = u\right) \tag{736}$$

$$= \mathbb{P}\left(\tau_{n+1} - (t-s) \geq x \,\middle|\, (N(s))_{0\leq s\leq t} \in E,\, \tau_{n+1} > t-s,\, T_n = u\right). \tag{737}$$

Conditioned on $N(t) = n$, the event that $(N(s))_{0\leq s\leq t} \in E$ is determined by the arrival times $T_1, \cdots, T_n$ and the fact that $T_{n+1} \geq t$. Hence we can rewrite

$$\{(N(s))_{0\leq s\leq t} \in E,\, \tau_{n+1} > t-u,\, T_n = u\} = \left\{(\tau_1, \cdots, \tau_n) \in E',\, \tau_{n+1} > t-u\right\} \tag{738}$$

for some event E' to be satisfied by the first n inter-arrival times. Since inter-arrival times are independent, this gives

$$\mathbb{P}\left(Z(t) \geq x \,\middle|\, (N(s))_{0\leq s\leq t} \in E,\, N(t) = n,\, T_{N(t)} = u\right) \tag{739}$$

$$= \mathbb{P}\left(\tau_{n+1} - (t-u) \geq x \,\middle|\, \tau_{n+1} \geq t-u\right) \tag{740}$$

$$= \mathbb{P}\left(\tau_{n+1} \geq x\right) = e^{-\lambda x}, \tag{741}$$

where we have used the memoryless property of exponential variables. Hence by iterated expectation,

$$\mathbb{P}\left(Z(t) \geq x \,\middle|\, (N(s))_{0\leq s\leq t} \in E,\, N(t) = n\right) = \mathbb{E}_{T_{N(t)}}\left[\mathbb{P}\left(Z(t) \geq x \,\middle|\, (N(s))_{0\leq s\leq t} \in E,\, N(t) = n,\, T_{N(t)}\right)\right] \tag{742}$$

$$= \mathbb{E}_{T_{N(t)}}[e^{-\lambda x}] = e^{-\lambda x}. \tag{743}$$

By using iterated expectation once more,

$$\mathbb{P}\left(Z(t) \geq x \,\middle|\, (N(s))_{0\leq s\leq t} \in E,\right) = \mathbb{E}_{N(t)}\left[\mathbb{P}\left(Z(t) \geq x \,\middle|\, (N(s))_{0\leq s\leq t} \in E,\, N(t)\right)\right] \tag{744}$$

$$= \mathbb{E}_{N(t)}[e^{-\lambda x}] = e^{-\lambda x}. \tag{745}$$

By taking E to be the entire sample space, this also gives

$$\mathbb{P}\left(Z(t) \geq x\right) = e^{-\lambda x}. \tag{746}$$

This shows (733). □

Exercise 2.8 (Sum of independent Poisson RV's is Poisson)**.** Let $(T_k)_{k\geq 1}$ be a Poisson process with rate λ and let $(N(t))_{t\geq 0}$ be the associated counting process. Fix $t, s \geq 0$.

(i) Use memoryless property to show that $N(t)$ and $N(t+s) - N(t)$ are independent Poisson RVs of rates λt and λs.

(ii) Note that the total number of arrivals during $[0, t+s]$ can be divided into the number of arrivals during $[0, t]$ and $[t, t+s]$. Conclude that if $X \sim \text{Poisson}(\lambda t)$ and $Y \sim \text{Poisson}(\lambda s)$ and if they are independent, then $X + Y \in \text{Poisson}(\lambda(t+s))$.

2.3. Splitting and merging of Poisson process. Recall the splitting of Bernoulli processes: If balls are given by BP(p) and we color each ball with blue and red independently with probability q and $1-q$, respectively, then the process restricted on blue and red balls are BP(pq) and BP($p(1-q)$), respectively. Considering blue balls process is sometimes called 'thinning' of the original BP. The same construction naturally works for Poisson processes as well. If customers arrive at a bank according to PP(λ) and if each one is male or female independently with probability q and $1-q$, then the 'thinned out' process of only male customers is a PP($q\lambda$); the process of female customers is a PP($(1-q)\lambda$).

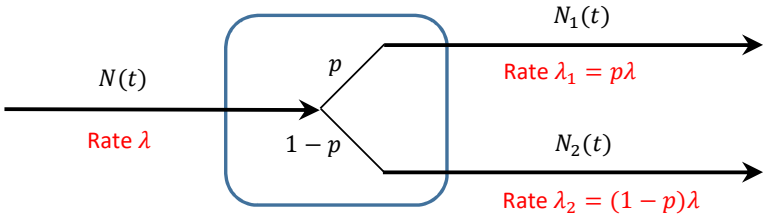

FIGURE 4. Splitting of Poisson process $N(t)$ of rate λ according to an independent Bernoulli process of parameter p.

The reverse operation of splitting a given PP into two complementary PPs is call the 'merging'. Namely, imagine customers arrive at a register through two doors A and B independently according to PPs of rates λ_A and λ_B, respectively. Then the combined arrival process of entire customers is again a PP of the added rate.

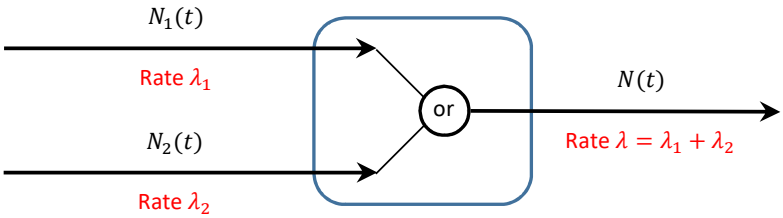

FIGURE 5. Merging two independent Poisson processes of rates λ_1 and λ_2 gives a new Poisson process of rate $\lambda_1 + \lambda_2$.

Exercise 2.9 (Excerpted from [**?**])**.** Transmitters A and B independently send messages to a single receiver according to Poisson processes with rates $\lambda_A = 3$ and $\lambda_B = 4$ (messages per min). Each message (regardless of the source) contains a random number of words with PMF

$$\mathbb{P}(1 \text{ word}) = 2/6, \quad \mathbb{P}(2 \text{ words}) = 3/6, \quad \mathbb{P}(3 \text{ words}) = 1/6, \tag{747}$$

which is independent of everything else.

(i) Find $\mathbb{P}$(total nine messages are recieved during $[0, t]$).
(ii) Let $M(t)$ be the total number of words received during $[0, t]$. Find $\mathbb{E}[M(t)]$.
(iii) Let T be the first time that the receiver receives exactly three messages consisting of three words from transmitter A. Find distribution of T.
(iv) Compute $\mathbb{P}$(exactly seven messages out of the first ten messages are from A).

Exercise 2.10 (Order statistics of i.i.d. Exp RVs)**.** One hundred light bulbs are simultaneously put on a life test. Suppose the lifetimes of the individual light bulbs are independent Exp(λ) RVs. Let T_k be the kth time that some light bulb fails. We will find the distribution of T_k using Poisson processes.

(i) Think of T_1 as the first arrival time among 100 independent PPs of rate λ. Show that $T_1 \sim \text{Exp}(100\lambda)$.
(ii) After time T_1, there are 99 remaining light bulbs. Using memoryless property, argue that $T_2 - T_1$ is the first arrival time of 99 independent PPs of rate λ. Show that $T_2 - T_1 \sim \text{Exp}(99\lambda)$ and that $T_2 - T_1$ is independent of T_1.
(iii) As in the coupon collector problem, we break up

$$T_k = \tau_1 + \tau_2 + \cdots + \tau_k, \tag{748}$$

where $\tau_i = T_i - T_{i-1}$ with $\tau_1 = T_1$. Note that τ_i is the waiting time between $i-1$st and ith failures. Using the ideas in (i) and (ii), show that τ_i's are independent and $\tau_i \sim \text{Exp}((100-i)\lambda)$. Deduce

that

$$\mathbb{E}[T_k] = \frac{1}{\lambda}\left(\frac{1}{100} + \frac{1}{99} + \cdots + \frac{1}{(100-k+1)}\right), \tag{749}$$

$$\mathrm{Var}[T_k] = \frac{1}{\lambda^2}\left(\frac{1}{100^2} + \frac{1}{99^2} + \cdots + \frac{1}{(100-k+1)^2}\right). \tag{750}$$

(iv) Let $X_1, X_2, \cdots, X_{100}$ be i.i.d. Exp(λ) variables. Let $X_{(1)} < X_{(2)} < \cdots < X_{(100)}$ be their order statistics, that is, $X_{(k)}$ is the ith smallest among the X_i's. Show that $X_{(k)}$ has the same distribution as T_k, the kth time some light bulb fails. (So we know what it is from the previous parts.)

In the next two exercises, we rigorously justify splitting and merging of Poisson processes.

Exercise 2.11 (Splitting of PP)**.** Let $(N(t))_{t\geq 0}$ be the counting process of a PP(λ), and let $(X_k)_{k\geq 0}$ be an independent BP(p). We define two counting processes $(N_1(t))_{t\geq 0}$ and $(N_2(t))_{t\geq 0}$ by

$$N_1(t) = \sum_{k=1}^{\infty} \mathbf{1}(T_k \leq t)\mathbf{1}(X_k = 1) = \#(\text{arrivals with coin landing on heads up to time } t), \tag{751}$$

$$N_2(t) = \sum_{k=1}^{\infty} \mathbf{1}(T_k \leq t)\mathbf{1}(X_k = 0) = \#(\text{arrivals with coin landing on heads up to time } t). \tag{752}$$

In this exercise, we show that $(N_1(t))_{t\geq 0} \sim \mathrm{PP}(p\lambda)$ and $(N_2(t))_{t\geq 0} \sim \mathrm{PP}((1-p)\lambda)$.

(i) Let τ_k and $\tau_k^{(1)}$ be the kth inter-arrival times of the counting processes $(N(t))_{t\geq 0}$ and $(N_1(t))_{t\geq 0}$. Let Y_k be the location of kth ball for the BP $(X_t)_{t\geq 0}$. Show that

$$\tau_1^{(1)} = \sum_{i=1}^{Y_1} \tau_i. \tag{753}$$

(ii) Show that

$$\tau_2^{(1)} = \sum_{k=Y_1+1}^{Y_2} \tau_i. \tag{754}$$

(iii) Show that in general,

$$\tau_k^{(1)} = \sum_{k=Y_{k-1}+1}^{Y_k} \tau_i. \tag{755}$$

(iv) Recall that $Y_k - Y_{k-1}$'s are i.i.d. Geom(p) RVs. Use Exercise 4.23 and (iii) to deduce that $\tau_k^{(1)}$'s are i.i.d. Exp($p\lambda$) RVs. Conclude that $(N_1(t)) \sim \mathrm{PP}(p\lambda)$. (The same argument shows $(N_2(t))_{t\geq 0} \sim \mathrm{PP}((1-p)\lambda)$.)

Exercise 2.12 (Merging of independent PPs)**.** Let $(N_1(t))_{t\geq 0}$ and $(N_2(t))_{t\geq 0}$ be the counting processes of two independent PPs of rates λ_1 and λ_2, respectively. Define a new counting process $(N(t))_{t\geq 0}$ by

$$N(t) = N_1(t) + N_2(t). \tag{756}$$

In this exercise, we show that $(N(t))_{t\geq 0} \sim \mathrm{PP}(p\lambda)$.

(i) Let $\tau_k^{(1)}$, $\tau_k^{(2)}$, and τ_k be the kth inter-arrival times of the counting processes $(N_1(t))_{t\geq 0}$, $(N_2(t))_{t\geq 0}$, and $(N(t))_{t\geq 0}$. Show that $\tau_1 = \min(\tau_1^{(1)}, \tau_1^{(2)})$. Conclude that $\tau_1 \sim \mathrm{Exp}(\lambda_1 + \lambda_2)$.

(ii) Let T_k be the kth arrival time for the joint process $(N(t))_{t\geq 0}$. Use memoryless property of PP to deduce that N_1 and N_2 restarted from time T_k are independent PPs of rates λ_1 and λ_2, which are also independent from the past (before time t).

(iii) From (ii), show that

$$\tau_{k+1} = \min(\tilde{\tau}_1, \tilde{\tau}_2), \tag{757}$$

where $\tilde{\tau}_1$ is the waiting time for the first arrival after time T_k for N_1, and similarly for $\tilde{\tau}_2$. Deduce that $\tau_{k+1} \sim \text{Exp}(\lambda_1+\lambda_2)$ and it is independent of $\tau_1,\cdots,\tau_k$. Conclude that $(N(t))_{t\geq 0} \sim \text{PP}(\lambda_1+\lambda_2)$.

3. Discrete-time Markov chains

3.1. Definition and examples of MCs. In this sebsection, we change our gear from arrival processes to *Markov processes*. Roughly speaking, Markov processes are used to model temporally changing systems where future state only depends on the current state. For instance, if the price of bitcoin tomorrow depends only on its price today, then bitcoin price can be modeled as a Markov process. (Of course, the entire history of price often affects decisions of buyers/sellers so it may not be a realistic assumption.)

Even through Markov processes can be defined in vast generality, we concentrate on the simplest setting where the state and time are both discrete. Let $\Omega = \{1,2,\cdots,m\}$ be a finite set, which we call the *state space*. Consider a sequence $(X_t)_{t\geq 0}$ of Ω-valued RVs, which we call a *chain*. We call the value of X_t the *state* of the chain at time t. In order to narrow down the way the chain $(X_t)_{t\geq 0}$ behaves, we introduce the following properties:

(i) (Markov property) The distribution of X_{t+1} given the history $X_0, X_1, \cdots, X_t$ depends only on X_t. That is,

$$\mathbb{P}(X_{t+1}=k \,|\, X_t=j_t, X_{t-1}=j_{t-1},\cdots,X_1=j_1) = \mathbb{P}(X_{t+1}=k \,|\, X_t=j_t). \tag{758}$$

(ii) (Time-homogeneity) The transition probabilities

$$p_{ij} = \mathbb{P}(X_{t+1}=j \,|\, X_t=i) \qquad i,j\in\Omega \tag{759}$$

do not depend on t.

When the chain $(X_t)_{t\geq 0}$ satisfies the above two properties, we say it is a (discrete-time and time-homogeneous) *Markov chain*. Note that the Markov property (i) is a kind of a one-step complication of the memoryless property: We now forget all the past but we do remember the present. On the other hand, time-homogeneity (ii) states that the behavior of the chain does not depend on time. In this case, we define the *transition matrix* P to be the $m\times m$ matrix of transition probabilities:

$$P = (p_{ij})_{1\leq i,j\leq m} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{bmatrix}. \tag{760}$$

Finally, since the state X_t of the chain is a RV, we represent its PMF via a row vector

$$\mathbf{r}_t = [\mathbb{P}(X_t=1), \mathbb{P}(X_t=2), \cdots, \mathbb{P}(X_t=m)]. \tag{761}$$

Example 3.1. Let $\Omega=\{1,2\}$ and let $(X_t)_{t\geq 0}$ be a Markov chain on Ω with the following transition matrix

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}. \tag{762}$$

We can also represent this Markov chain pictorially as in Figure 8, which is called the 'state space diagram' of the chain $(X_t)_{t\geq 0}$.

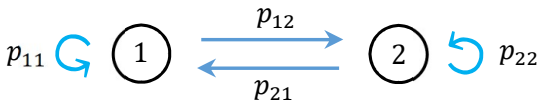

FIGURE 6. State space diagram of a 2-state Markov chain

For some concrete example, suppose

$$p_{11} = 0.2, \quad p_{12} = 0.8, \quad p_{21} = 0.6, \quad p_{22} = 0.4. \tag{763}$$

If the initial state of the chain X_0 is 1, then

$$\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = 1 \mid X_0 = 1)\mathbb{P}(X_0 = 1) + \mathbb{P}(X_1 = 1 \mid X_0 = 2)\mathbb{P}(X_0 = 2) \tag{764}$$

$$= \mathbb{P}(X_1 = 1 \mid X_0 = 1) = p_{11} = 0.2 \tag{765}$$

and similarly,

$$\mathbb{P}(X_1 = 2) = \mathbb{P}(X_1 = 2 \mid X_0 = 1)\mathbb{P}(X_0 = 1) + \mathbb{P}(X_1 = 2 \mid X_0 = 2)\mathbb{P}(X_0 = 2) \tag{766}$$

$$= \mathbb{P}(X_1 = 2 \mid X_0 = 1) = p_{12} = 0.8. \tag{767}$$

Also we can compute the distribution of X_2. For example,

$$\mathbb{P}(X_2 = 1) = \mathbb{P}(X_2 = 1 \mid X_1 = 1)\mathbb{P}(X_1 = 1) + \mathbb{P}(X_2 = 1 \mid X_1 = 2)\mathbb{P}(X_1 = 2) \tag{768}$$

$$= p_{11}\mathbb{P}(X_1 = 1) + p_{21}\mathbb{P}(X_1 = 2) \tag{769}$$

$$= 0.2 \cdot 0.2 + 0.6 \cdot 0.8 = 0.04 + 0.48 = 0.52. \tag{770}$$

In general, the distribution of X_{t+1} can be computed from that of X_t via a simple linear algebra. Note that for $i = 1, 2$,

$$\mathbb{P}(X_{t+1} = i) = \mathbb{P}(X_{t+1} = i \mid X_t = 1)\mathbb{P}(X_t = 1) + \mathbb{P}(X_{t+1} = i \mid X_t = 2)\mathbb{P}(X_t = 2) \tag{771}$$

$$= p_{1i}\mathbb{P}(X_t = 1) + p_{2i}\mathbb{P}(X_t = 2). \tag{772}$$

This can be written as

$$[\mathbb{P}(X_{t+1} = 2), \mathbb{P}(X_{t+1} = 2)] = [\mathbb{P}(X_{t+1} = 2), \mathbb{P}(X_{t+1} = 2)] \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}. \tag{773}$$

That is, if we represent the distribution of X_t as a row vector, then the distribution of X_{t+1} is given by multiplying the transition matrix P to the left.

We generalize this observation in the following exercise.

Exercise 3.2. Let $(X_t)_{t\geq 0}$ be a Markov chain on state space $\Omega = \{1, 2, \cdots, m\}$ with transition matrix $P = (p_{ij})_{1\leq i,j\leq m}$. Let $\mathbf{r}_t = [\mathbb{P}(X_t = 1), \cdots, \mathbb{P}(X_t = m)]$ denote the row vector of the distribution of X_t.

(i) Show that for each $i \in \Omega$,

$$\mathbb{P}(X_{t+1} = i) = \sum_{j=1}^{m} p_{ji}\mathbb{P}(X_t = j). \tag{774}$$

(ii) Show that for each $t \geq 0$,

$$\mathbf{r}_{t+1} = \mathbf{r}_t P. \tag{775}$$

(iii) Show by induction that for each $t \geq 0$,

$$\mathbf{r}_t = \mathbf{r}_0 P^t. \tag{776}$$

Exercise 3.3. Let $\Omega = \{1, 2\}$ and let $(X_t)_{t\geq 0}$ be a Markov chain on Ω with the following transition matrix

$$P = \begin{bmatrix} 0.2 & 0.8 \\ 0.6 & 0.4 \end{bmatrix}. \tag{777}$$

(i) Show that P admits the following diagonalization

$$P = \begin{bmatrix} 1 & -4/3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -2/5 \end{bmatrix} \begin{bmatrix} 1 & -4/3 \\ 1 & 1 \end{bmatrix}^{-1}. \tag{778}$$

(ii) Show that P^t admits the following diagonalization

$$P^t = \begin{bmatrix} 1 & -4/3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (-2/5)^t \end{bmatrix} \begin{bmatrix} 1 & -4/3 \\ 1 & 1 \end{bmatrix}^{-1}. \tag{779}$$

(iii) Let $\mathbf{r}_t$ denote the row vector of distribution of X_t. Use Exercise 3.2 to deduce that

$$\mathbf{r}_t = \mathbf{r}_0 \begin{bmatrix} 1 & -4/3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (-2/5)^t \end{bmatrix} \begin{bmatrix} 1 & -4/3 \\ 1 & 1 \end{bmatrix}^{-1}. \tag{780}$$

Also show that

$$\lim_{t\to\infty} \mathbf{r}_t = \mathbf{r}_0 \begin{bmatrix} 3/7 & 4/7 \\ 3/7 & 4/7 \end{bmatrix} = [3/7, 4/7]. \tag{781}$$

Conclude that regardless of the initial distribution $\mathbf{r}_0$, the distribution of the Markov chain $(X_t)_{t\geq 0}$ converges to $[3/7, 4/7]$. This limiting distribution $\pi = [3/7, 4/7]$ is called the *stationary distribution* of the chain $(X_t)_{t\geq 0}$.

3.2. Stationary distribution and examples. Let $(X_t)_{t\geq 0}$ be a Markov chain on a finite state space $\Omega = \{1, 2, \cdots, m\}$ with transition matrix $P = (p_{ij})_{1\leq i,j\leq m}$. If π is a distribution on Ω such that

$$\pi = \pi P, \tag{782}$$

then we say π is a *stationary distribution* of the Markov chain $(X_t)_{t\geq 0}$.

Example 3.4. In Exercise 3.3, we have seen that the distribution of the 2-state Markov chain $(X_t)_{t\geq 0}$ with transition matrix

$$P = \begin{bmatrix} 0.2 & 0.8 \\ 0.6 & 0.4 \end{bmatrix}. \tag{783}$$

converges to $\pi = [3/7, 4/7]$. Since this is the limiting distribution, it should be invariant under left multiplication by P. Indeed, one can easily verify

$$[3/7, 4/7] = [3/7, 4/7] \begin{bmatrix} 0.2 & 0.8 \\ 0.6 & 0.4 \end{bmatrix}. \tag{784}$$

Hence π is a stationary distribution for the Markov chain $(X_t)_{t\geq 0}$. Furthermore, in Exercise 3.3 we also have shown the uniqueness of stationary distribution. However, this is not always the case.

Example 3.5. Let $(X_t)_{t\geq 0}$ be a 2-state Markov chain with transition matrix

$$P = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{785}$$

Then any distribution $\pi = [p, 1-p]$ is a stationary distribution for the chain $(X_t)_{t\geq 0}$.

In Exercise 3.3, we used diagonalization of the transition matrix to compute the limiting distribution, which must be a stationary distribution. However, we can simply use the definition (782) to algebraically compute stationary distribution(s). Namely, by taking transpose,

$$\pi^T = P^T \pi^T. \tag{786}$$

Namely, the transpose of any stationary distribution is an eigenvector of P^T associated with eigenvalue 1. We record some properties of stationary distributions using some linear algebra stuff.

Exercise 3.6. Let $(X_t)_{t\geq 0}$ be a Markov chain on state space $\Omega = \{1, 2, \cdots, m\}$ with transition matrix $P = (p_{ij})_{1\leq i,j\leq m}$.

(i) Show that a distribution π on Ω is a stationary distribution for the chain $(X_t)_{t\geq 0}$ if and only if it is a left eigenvector of P associated with left eigenvalue 1.

(ii) Show that 1 is a right eigenvalue of P with right eigenvector $[1, 1, \cdots, 1]^T$.

(iii) Recall that a square matrix and its transpose have the same (right) eigenvalues and corresponding (right) eigenspaces have the same dimension. Show that the Markov chain $(X_t)_{t\geq 0}$ has a unique stationary distribution if any only if $[1,1,\cdots,1]^T$ spans the (right) eigenspace of P associated with (right) eigenvalue 1.

Now we look at some important examples.

Exercise 3.7 (Birth-Death chain)**.** Let $\Omega=\{0,1,2,\cdots,N\}$ be the state space. Let $(X_t)_{t\geq 0}$ be a Markov chain on Ω with transition probabilities

$$\begin{cases} \mathbb{P}(X_{t+1}=k+1\,|\,X_t=k)=p & \forall 0\leq k<N \\ \mathbb{P}(X_{t+1}=k-1\,|\,X_t=k)=1-p & \forall 1\leq k\leq N \\ \mathbb{P}(X_{t+1}=0\,|\,X_t=0)=1-p & \\ \mathbb{P}(X_{t+1}=N\,|\,X_t=N)=p. & \end{cases} \tag{787}$$

This is called a Birth-Death chain. Its state space diagram is as below.

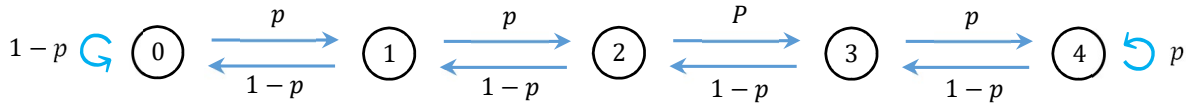

FIGURE 7. State space diagram of a 5-state Birth-Death chain

(i) Let $\pi=[\pi_0,\pi_1,\cdots,\pi_N]$ be a distribution on Ω. Show that π is a stationary distribution of the Birth-Death chain if and only if it satisfy the following 'balance equation'

$$p\pi_k=(1-p)\pi_{k+1} \qquad 0\leq k<N. \tag{788}$$

(ii) Let $\rho=p/(1-p)$. From (ii), deduce that $\pi_k=\rho^k\pi_0$ for all $0\leq k<N$.

(iii) Using the normalization condition $\pi_0+\pi_1+\cdots+\pi_N$, show that $\pi_0=1/(1+\rho+\rho^2+\cdots+\rho^N)$. Conclude that

$$\pi_k=\frac{\rho^k}{1+\rho+\rho^2+\cdots+\rho^N}=\rho^k\frac{1-\rho}{1-\rho^{N+1}} \qquad 0\leq k\leq N. \tag{789}$$

Conclude that the Birth-Death chain has a unique stationary distribution given by (789).

In the following example, we will encounter a new concept of 'absorption' of Markov chains.

Exercise 3.8 (Gambler's ruin)**.** Suppose a gambler has fortune of k dolors initially and starts gambling. At each time he wins or loses 1 dolor independently with probability p and $1-p$, respectively. The game ends when his fortune reaches either 0 or N dolors. What is the probability that he wins N dolors and goes home happy?

We use Markov chains to model his fortune after betting t times. Namely, let $\Omega=\{0,1,2,\cdots,N\}$ be the state space. Let $(X_t)_{t\geq 0}$ be a sequence of RVs where X_t is the gambler's fortune after betting t times. Note that the transition probabilities are similar to that of the Birth-Death chain, except the 'absorbing boundary' at 0 and N. Namely,

$$\begin{cases} \mathbb{P}(X_{t+1}=k+1\,|\,X_t=k)=p & \forall 1\leq k<N \\ \mathbb{P}(X_{t+1}=k\,|\,X_t=k-1)=1-p & \forall 1\leq k<N \\ \mathbb{P}(X_{t+1}=0\,|\,X_t=0)=1 & \\ \mathbb{P}(X_{t+1}=N\,|\,X_t=N)=1. & \end{cases} \tag{790}$$

Call the resulting Markov chain $(X_t)_{t\geq 0}$ the *gambler*'s chain. Its state space diagram is given below.

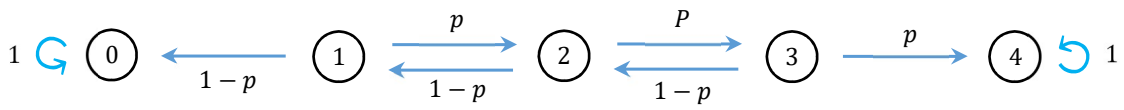

FIGURE 8. State space diagram of a 5-state gambler's chain

(i) Show that any distribution $\pi = [a, 0, 0, \cdots, 0, b]$ on Ω is stationary with respect to the gambler's chain. Also show that any stationary distribution of this chain should be of this form.

(ii) Clearly the gambler's chain eventually visits state 0 or N, and stays at that boundary state thereafter. This is called *absorbtion*. Let τ_i denote the time until absorbtion starting from state i:

$$\tau_i = \min\{t \geq 0 : X_t \in \{0, N\} \,|\, X_0 = i\}. \tag{791}$$

We are going to compute the 'winning probabilities': $q_i := \mathbb{P}(X_{\tau_i} = N)$.
By considering what happens in one step, show that they satisfy the following recursion

$$\begin{cases} q_i = p\,q_{i+1} + (1-p)\,q_{i-1} & \forall 1 \leq i < N \\ q_0 = 0, \quad q_N = 1 \end{cases}. \tag{792}$$

(iii) Denote $\rho = (1-p)/p$. Show that

$$q_{i+1} - q_i = \rho(q_i - q_{i-1}) \qquad \forall 1 \leq i < N. \tag{793}$$

Deduce that

$$q_{i+1} - q_i = \rho^i (q_1 - q_0) = \rho^i q_1 \qquad \forall 1 \leq i < N, \tag{794}$$

and that

$$q_i = q_1(1 + \rho + \cdots + \rho^{i-1}) \qquad \forall 1 \leq i \leq N. \tag{795}$$

(iv)* Conclude that

$$q_i = \begin{cases} \frac{1-\rho^i}{N - \rho(1-\rho^N)/(1-\rho)} & \text{if } p \neq 1/2 \\ \frac{2i}{N(N-1)} & \text{if } p = 1/2. \end{cases} \tag{796}$$

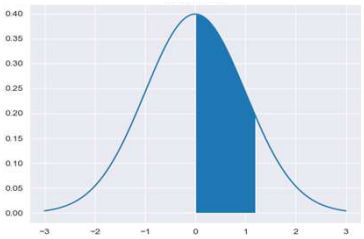

Table of standard normal probabilities $\mathbb{P}(0 \leq Z \leq z)$, where $Z \sim N(0,1)$.

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000

TABLE 1. Standard normal table