INTRODUCTION TO PROBABILITY THEORY AND STOCHASTIC PROCESSES FOR UNDERGRADUATE STUDENTS

ANDREY SARANTSEV

August 25, 2019

Department of Mathematics and Statistics

University of Nevada in Reno

Email: asarantsev@unr.edu

Introduction

These lecture notes include both discrete- and continuous-time processes. We consider applications to insurance, finance, biology, and social sciences. These lecture notes are intended for junior- and senior-level undergraduate courses. They contain enough material for two semesters or three academic quarters.

The first 10 sections are devoted to Probability Theory (first semester), and the next 10 sections are devoted to Stochastic Processes (second semester). Appendix contains methods of simulations. These lecture notes are suitable for STEM majors, but are likely too hard for business and humanities majors.

A prerequisite for the first semester is calculus: a standard two-semester or three-quarter sequence, including multivariable calculus. For the second semester, we also need differential equations and matrix algebra. We tried to keep notes as condensed and concise as possible, very close to the actual notes written on the board.

Acknowledgements

- **University of Nevada in Reno:** I am thankful to Anna Panorska for guiding me through teaching Probability I in semester setting at UNR and providing me with her lecture notes, from which I borrowed some content. I am thankful to her, as well as to Birant Ramazan and Charlie Nazemian, for observing my classes and giving me useful advice.
- **University of California in Santa Barbara:** I am thankful to Tomoyuki Ichiba and Raya Feldman, who mentored me when I started teaching at the Department of Statistics and Applied Probability.
- **University of Washington in Seattle:** I was a PhD student at the Department of Mathematics, I am grateful to Judith Arms and Andrew Loveless, my teaching mentors, and to Boris Solomyak, for whom I was a TA for the first quarter.

I would like to thank Janko Gravner from Department of Mathematics, University of California in Davis, who put his lecture notes on Probability and Markov Chains on his web page. I use his materials a lot. Last but not least, I am thankful to Soumik Pal, my Ph.D. adviser, Ioannis Karatzas, his Ph.D. adviser, Jean-Pierre Fouque, my postdoctoral mentor, and all my graduate and undergraduate professors in Probability Theory and Stochastic Processes, for helping understand and appreciate this exciting area of mathematics.

1. Combinatorics

1.1. **Permutations.** A *permutation* of 123 is, say, 321 or 231: numbers cannot repeat. There are $3 \cdot 2 \cdot 1 = 3! = 6$ permutations: there are 3 choices for the first slot, 2 choices for the second slot (because one of the numbers is already in the first slot and cannot be repeated), and only one choice for the last, third slot.

In general, for $1, 2, 3, \ldots, n-1, n$ there are

$$\boxed{n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot (n-1) \cdot n}$$

permutations. This number is called n *factorial*. Examples:

$$1! = 1, \quad 2! = 2, \quad 3! = 6, \quad 4! = 24, \quad 5! = 120, \quad 6! = 720.$$

There is a convention that $0! = 1$. Indeed, we have the property $(n-1)!n = n!$ for $n = 2, 3, 4, \ldots$. It follows from the definition of the factorial, and this is the main property. We would like it to be true also for $n = 1$: $0! \cdot 1 = 1!$, so $0! = 1$. Factorial grows very quickly. Indeed, 100! is extremely large; no modern computer can go through permutations of $1, 2, \ldots, 100$. When a computer programming problem encounters search among permutations of n numbers, then this problem is deemed unsolvable. *Stirling's formula*:

$$\boxed{n! \backsim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \quad \text{as} \quad n \to \infty}$$

where $f(n) \backsim g(n)$ means $\lim_{n\to\infty} f(n)/g(n) = 1$.

If we have three slots for numbers $1, 2, 3, 4, 5, 6, 7$, and repetitions are not allowed, this is called an *arrangement*. Say, 364, 137, 634. There are $7 \cdot 6 \cdot 5 = 210$ such arrangements: 7 choices for the first slot, 6 for the second and 5 for the third. We can write this as

$$A_7^3 = 7 \cdot 6 \cdot 5 = \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{4 \cdot 3 \cdot 2 \cdot 1} = \frac{7!}{(7-3)!}.$$

In general, if there are k slots for $1, 2, \ldots, n$, then the number of arrangements is

$$A_n^k = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1) = \frac{n!}{(n-k)!}.$$

A permutation can be viewed as a particular case of an arrangement, when $k = n$: the number of slots is the same as the total number of elements.

1.2. **Subsets.** How many subsets of three elements are there in the set $\{1, 2, \ldots, 7\}$? The difference between an arrangement and a subset is that for a subset, order does not matter. (But in both of them, there are no repetitions.) For example, $\{3, 4, 6\}$ and $\{6, 3, 4\}$ is the same subset, but 346 and 634 are different arrangements. From any subset, we can create $3! = 6$ arrangements. The quantity of subsets is equal to the quantity of arrangements divided by 6:

$$\frac{A_7^3}{3!} = \frac{7!}{4!3!} = \frac{210}{6} = 35.$$

In general, the quantity of subsets of k elements in $\{1, \ldots, n\}$ is equal to

$$\boxed{\binom{n}{k} = \frac{A_n^k}{k!} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{k!} = \frac{n!}{(n-k)!k!}}$$

It is pronounced as "n choose k".

(i) $\binom{1}{0} = 1$, because there is only one subset of zero elements in $\{1\}$, and this is an empty set $\varnothing$.

(ii) $\binom{1}{1} = 1$, because there is only one subset of one element in $\{1\}$: the set $\{1\}$ itself.

(iii) $\binom{n}{0} = 1$, for the same reason as in (i);

(iv) $\binom{n}{n} = 1$, for the same reason as in (ii);

(v) $\binom{2}{1} = 2$, because there are two subsets of one element of $\{1, 2\}$: these are $\{1\}$ and $\{2\}$;

(vi) $\binom{n}{1} = n$, because there are n subsets of one element of $\{1, 2, \ldots, n\}$: $\{1\}$, $\{2\}, \ldots, \{n\}$;

(vii) $\binom{n}{n-1} = n$, because to choose a subset of $n-1$ elements out of $\{1, 2, \ldots, n\}$, we need to throw away one element, and it can be chosen in n ways;

(viii) $\binom{4}{2} = 4!/(2!2!) = 24/4 = 6$, and these subsets of $\{1, 2, 3, 4\}$ are

$$\{1, 2\}, \ \{1, 3\}, \ \{1, 4\}, \ \{2, 3\}, \ \{2, 4\}, \ \{3, 4\}.$$

1.3. **Symmetry.** We can say without calculations that $\binom{8}{2} = \binom{8}{6}$. Indeed, for every subset of $\{1, 2, \ldots, 8\}$ of two elements there is a subset of six elements: its complement. For example, $\{3, 5\}$ corresponds to $\{1, 2, 4, 6, 7, 8\}$. This is a one-to-one correspondence. So there are equally many subsets of two elements and subsets of six elements. Similarly, $\binom{8}{3} = \binom{8}{5}$. More generally,

$$\boxed{\binom{n}{k} = \binom{n}{n-k}}$$

1.4. **Power set.** How many subsets does the set $\{1, 2, \ldots, n\}$ contain? Answer: 2^n. Indeed, to construct an arbitrary subset E, you should answer n questions:

- Is $1 \in E$? Yes/No
- Is $2 \in E$? Yes/No
- ...
- Is $n \in E$? Yes/No

For each question, there are two possible answers. The total number of choices is $\underbrace{2 \cdot 2 \cdot \ldots \cdot 2}_{\text{n times}} = 2^n$. The set of all subsets of $\{1, \ldots, n\}$ is called a *power set*, and it contains 2^n elements. But we can also write the quantity of all subsets as the sum of binomial coefficients:

$$\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \ldots + \binom{n}{n}.$$

So we get the following identity:

$$\boxed{\binom{n}{0} + \binom{n}{1} + \ldots + \binom{n}{n} = 2^n}$$

Example 1.1. Let $n = 2$. Yes-Yes: $E = \{1, 2\}$, Yes-No: $E = \{2\}$, No-Yes: $E = \{1\}$, No-No: $E = \varnothing$. Total number of subsets: $2^2 = 4$. Two of them have one element: $\binom{2}{1} = 2$, one has two elements, $\binom{2}{2} = 1$, and one has zero elements, $\binom{2}{0} = 1$. Total: $1 + 2 + 1 = 4$.

1.5. **Reduction property.** We can claim that

$$\binom{5}{2} = \binom{4}{2} + \binom{4}{1}.$$

Indeed, the total number of subsets $E \subseteq \{1, 2, 3, 4, 5\}$ which contain two elements is $\binom{5}{2}$. But there are two possibilities:

Case 1. $5 \in E$. Then $E \setminus \{5\}$ is a one-element subset of $\{1, 2, 3, 4\}$; there are $\binom{4}{1}$ such subsets.

Case 2. $5 \notin E$. Then E is a two-element subset of $\{1, 2, 3, 4\}$. There are $\binom{4}{2}$ such subsets.

So $\binom{4}{1} + \binom{4}{2} = \binom{5}{2}$. In general,

$$\boxed{\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}}$$

1.6. **Pascal's triangle.**

$n = 0$: 1

$n = 1$: 1 1

$n = 2$: 1 2 1

$n = 3$: 1 3 3 1

$n = 4$: 1 4 6 4 1

$n = 0$: $\binom{0}{0}$

$n = 1$: $\binom{1}{0}$ $\binom{1}{1}$

$n = 2$: $\binom{2}{0}$ $\binom{2}{1}$ $\binom{2}{2}$

$n = 3$: $\binom{3}{0}$ $\binom{3}{1}$ $\binom{3}{2}$ $\binom{3}{3}$

$n = 4$: $\binom{4}{0}$ $\binom{4}{1}$ $\binom{4}{2}$ $\binom{4}{3}$ $\binom{4}{4}$

Each element is the sum of two elements immediately above it: this is the reduction formula. We start from the edges, fill them with ones: $\binom{n}{0} = \binom{n}{n} = 1$, see the previous lecture. Then we fill the inside from top to bottom using this rule, which is the reduction formula.

1.7. **Newton's binomial formula.** We can expand $(x+y)^2 = x^2+2xy+y^2$, and $(x+y)^3 = x^3+3x^2y+3xy^2+y^3$. The coefficients are taken from corresponding lines in Pascal's triangle. Why is this? Let us show this for $n = 3$.

$$(x+y)^3 = (x+y)(x+y)(x+y) = xxx + xxy + xyx + yxx + xyy + yxy + yyx + yyy.$$

Each term has slots occupied by y: $xxy \leftrightarrow \{3\}$, $yxy \leftrightarrow \{1,3\}$. If there is one slot occupied by y, this corresponds to x^2y, and there are $\binom{3}{1}$ such combinations. So we have: $\binom{3}{1}x^2y$. Other terms give us:

$$\binom{3}{0}x^3 + \binom{3}{1}x^2y + \binom{3}{2}xy^2 + \binom{3}{3}y^3.$$

The general formula looks like this:

$$\boxed{(x+y)^n = \binom{n}{0}x^n + \binom{n}{1}x^{n-1}y + \ldots + \binom{n}{n}y^n}$$

Let $x = y = 1$. Then we get:

$$2^n = \binom{n}{0} + \binom{n}{1} + \ldots + \binom{n}{n}.$$

This formula was already proven above. Let $x = 1, y = -1$. Then

$$0 = \binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \binom{n}{3} + \ldots,$$

$$\binom{n}{0} + \binom{n}{2} + \ldots = \binom{n}{1} + \binom{n}{3} + \ldots$$

The quantity of subsets with even number of elements is equal to the quantity of subsets with odd number of elements.

1.8. **Combinatorics problems.** Here, we study a few combinatorics (counting) problems, which can be reduced to counting permutations and combinations.

Example 1.2. Five women and four men take an exam. We rank them from top to bottom, according to their performance. There are no ties.

(a) How many possible rankings?

(b) What if we rank men and women separately?

(c) As in (b), but Julie has the third place in women's rankings.

(a) A ranking is just another name for permutation of nine people. The answer is $\boxed{9!}$

(b) There are 5! permutations for women and 4! permutations for men. The total number is $\boxed{5!4!}$ We should multiply them, rather than add, because men's and women's rankings are independent: we are interested in pairs: the first item is a ranking for women, the second item is a ranking for men. If we needed to choose: either rank women or rank men, then the solution would be $5! + 4!$

(c) We exclude Julie from consideration, because her place is already reserved. There are four women remaining, so the number of permutations is 4! For men, it is also 4! The answer is $\boxed{4!^2}$

Example 1.3. A licence plate consists of seven symbols: digits or letters. How many licence plates are there if the following is true:

(a) there must be three letters and four digits, and symbols may repeat?

(b) there must be three letters and four digits, and symbols may not repeat?

(c) no restrictions on the quantity of letters and numbers, and symbols may repeat?

(d) no restrictions on the quantity of letters and numbers, and symbols may not repeat?

(a) Choose three slots among seven for letters; this may be done in $\binom{7}{3}$ ways. Fill each of these three slots with letters; there are 26^3 ways to do this, since letters can repeat. Fill each of the remaining four slots with digits; there are 10^4 ways to do this, since numbers can repeat. Answer: $\boxed{\binom{7}{3} \cdot 26^3 \cdot 10^4}$

(b) This case is different from (i) because there are $26 \cdot 25 \cdot 24 = 26!/23!$ ways to fill three chosen slots for letters, and $10 \cdot 9 \cdot 8 \cdot 7 = 10!/6!$ ways to fill four chosen slots for numbers. Answer: $\boxed{\binom{7}{3} 26 \cdot 25 \cdot 24 \cdot 10 \cdot 9 \cdot 8 \cdot 7}$

(c) This case is easier than the previous ones, since there are 36 symbols, and each of the seven slots can be filled with any of these symbols. Answer: $\boxed{36^7}$

(d) Similarly to (c), $\boxed{36 \cdot 35 \cdot 34 \cdot 33 \cdot 32 \cdot 31 \cdot 30 = 36!/29!}$

Example 1.4. We have five women and four men. We need to choose a committee of three women and two men. How many ways are there to do this if:

(a) there are no additional restrictions?

(b) Mike and Thomas refuse to serve together?

(c) Britney and Lindsey refuse to serve together?

(d) Andrew and Anna refuse to serve together?

(a) There are $\binom{4}{2} = 6$ ways to choose two men out of four, and there are $\binom{5}{3} = 10$ ways to choose two women out of five. So the answer is $\boxed{60}$

(b) How many committees are there for which this restriction is violated, so Mike and Thomas do serve together? If they are already chosen, then we do not need to choose any other man, and there are $\binom{5}{3} = 10$ ways to choose three women out of five. So the quantity of committees where Mike and Thomas do serve together is 10. The answer is $60 - 10 = \boxed{50}$

(c) Similarly to (b), the number of committees where Britney and Lindsay serve together is $\binom{3}{1}\binom{4}{2} = 18$, because you can choose one more woman out of the remaining three in $\binom{3}{1} = 3$ ways, and the number of choices for men is $\binom{4}{2}$. So the answer is $60 - 18 = \boxed{42}$

(d) Similarly to (c), the number of committees where Andrew and Anna serve together is $\binom{3}{1}\binom{4}{2} = 18$, because you can choose one more man out of the remaining three in $\binom{3}{1} = 3$ ways, and two more women out of the remaining four in $\binom{4}{2}$ ways. So the answer is $60 - 18 = \boxed{42}$

2. Axioms of Probability

2.1. **Set-theoretic notation.** A *set* is a collection of its *elements*: $A = \{1, 3, 4, 5\}$, $B = \{red, blue\}$. We say: $1 \in A$ (1 belongs to A), but $2 \notin A$ (2 does not belong to A). For two sets A and B, we can define their *union*: $A \cup B = \{x \mid x \in A \text{ or } x \in B\}$, *intersection*: $A \cap B := \{x \mid x \in A \text{ and } x \in B\}$, and *difference*: $A \setminus B = \{x \mid x \in A \text{ and } x \notin B\}$. The *empty set*, which does not contain any elements, is denoted by $\varnothing$.

Example 2.1. Let $A = \{0, 1, 2, 3\}$, $B = \{0 \le x \le 7 \mid \ x \text{ is odd}\} = \{1, 3, 5, 7\}$. Then

$$A \cup B = \{0, 1, 2, 3, 5, 7\},\ A \cap B = \{1, 3\},\ A \setminus B = \{0, 2\},\ B \setminus A = \{5, 7\}.$$

2.2. **Axioms of probability.** The following set of axioms was formulated by a Russian mathematician Andrey Kolmogorov. We have the set S of *elementary outcomes* $s \in S$. This set is called the *sample space*, and subsets $A \subseteq S$ are called *events*. Each event A has a number $\mathbf{P}(A)$, which is called the *probability* of A and satisfies the following axioms:

(a) $0 \le \mathbf{P}(A) \le 1$: every probability is between 0 and 1;

(b) $\mathbf{P}(\varnothing) = 0$: the probability of an empty set is 0;

(c) $\mathbf{P}(S) = 1$: the probability of the total event is 1;

(d) If two events A and B are *disjoint*: $A \cap B = \varnothing$, then $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$. The same is true for three or more disjoint events, or even for infinitely many.

By the axiom (d), if we taken the *complement* of the event A: $A^c = S \setminus A$, then $\mathbf{P}(A) + \mathbf{P}(A^c) = \mathbf{P}(A \cap A^c) = \mathbf{P}(S) = 1$, and therefore $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$.

Example 2.2. Toss a coin twice. Then $S = \{\text{HH, TT, HT, TH}\}$, and the probabilities of each outcome is $1/4$. For example, the event $A = \{\text{the same result}\} = \{\text{HH, TT}\}$ has probability $\mathbf{P}(A) = 1/2$.

Example 2.3. (Chevalier de Mere's Problem.) What is the probability that at least one six in four rolls of a fair die? Denote this event by A. The probability space is

$$S = \{(i_1, i_2, i_3, i_4) \mid i_1, i_2, i_3, i_4 \in \{1, 2, 3, 4, 5, 6\}\}.$$

The probability of each elementary outcome is $1/6^4$. The complement of the event A is $A^c = S \setminus A$, which means that there are no six in any of the four rolls. We have:

$$A^c = \{(i_1, i_2, i_3, i_4) \mid i_1, i_2, i_3, i_4 \in \{1, 2, 3, 4, 5\}\}.$$

The event A^c contains 5^4 outcomes. Each has probability $1/6^4$, so $\mathbf{P}(A^c) = 5^4/6^4$. But

$$\mathbf{P}(A) = 1 - \mathbf{P}(A^c) = 1 - \left(\frac{5}{6}\right)^4 \approx 0.5177.$$

This event has probability greater than 50%, so it is good to bet on it.

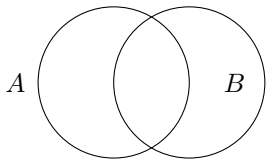
Example 2.4. Suppose we have $n = 365$ days, each day is equally likely to be a birthday. There are k people. Then

$$\begin{aligned}&\mathbf{P}(\text{there are two people with same birthdays})\\ &= 1 - \mathbf{P}(\text{there are no people with same birthdays})\\ &= 1 - \frac{n(n-1)\dots(n-k+1)}{n^k},\end{aligned}$$

because the number of birthday arrangements (assuming they are all different) is $A_n^k = n(n-1)\dots(n-k+1)$. This probability for $k = 23$ equals 0.5073, greater than 50%!

2.3. **Inclusion-exclusion formula for two events.** Consider two events, A and B, which can intersect. Then $\mathbf{P}(A \cup B) \neq \mathbf{P}(A) + \mathbf{P}(B)$. Indeed, toss two coins, let A be the event that the first toss is H, let B be the event that the second toss is H. Then $\mathbf{P}(A \cup B) = 3/4$, but $\mathbf{P}(A) = \mathbf{P}(B) = 1/2$. Instead, we count intersection $A \cap B$ twice in $\mathbf{P}(A) + \mathbf{P}(B)$, so we need to subtract it. Then we get real probability $\mathbf{P}(A \cup B)$.

Theorem 2.1. *We have:* $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$.



Proof. Indeed, let $C_1 = A \setminus B$, $C_2 = A \cap B$, $C_3 = B \setminus A$. Then

$$\begin{aligned}\mathbf{P}(A \cup B) = \mathbf{P}(C_1 \cup C_2 \cup C_3) &= \mathbf{P}(C_1) + \mathbf{P}(C_2) + \mathbf{P}(C_3),\\ \mathbf{P}(A) = \mathbf{P}(C_1 \cup C_2) &= \mathbf{P}(C_1) + \mathbf{P}(C_2),\\ \mathbf{P}(B) = \mathbf{P}(C_2 \cup C_3) &= \mathbf{P}(C_2) + \mathbf{P}(C_3),\\ \mathbf{P}(A \cap B) &= \mathbf{P}(C_2),\end{aligned}$$

and the formula follows from here. □

Example 2.5. $\mathbf{P}(A) = 30\%$, $\mathbf{P}(B) = 50\%$, $\mathbf{P}(A \cup B) = 70\%$. Find: $\mathbf{P}(A \cap B)$, $\mathbf{P}(A^c \cup B^c)$, and $\mathbf{P}(A^c \cap B)$. Solution: $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B) = 0.3 + 0.5 - 0.1 = 0.7$; $\mathbf{P}(A^c \cup B^c) = 1 - \mathbf{P}(A \cap B) = 1 - 0.1 = 0.9$; $\mathbf{P}(A^c \cap B) = \mathbf{P}(B) - \mathbf{P}(A \cap B) = 0.5 - 0.1 = 0.4$.

Example 2.6. (SOA) 22% patients visit both physical therapist T and a chiropractor C, 12% visit neither. Next, $\mathbf{P}(C) = 0.14 + \mathbf{P}(T)$. Find $\mathbf{P}(T)$. Indeed, $\mathbf{P}(C \cap T) = 22\%$ and $\mathbf{P}(C^c \cap T^c) = 12\%$, so $\mathbf{P}(C \cup T) = 88\%$. Next, $\mathbf{P}(C) + \mathbf{P}(T) = \mathbf{P}(C \cup T) + \mathbf{P}(C \cap T) = 1.1$. But $\mathbf{P}(C) = 0.14 + \mathbf{P}(T)$. Thus $2\mathbf{P}(T) + 0.14 = 1.1$ and $\mathbf{P}(T) = 0.48$.

Example 2.7. Choose a random number from 1 to 1000. What is the probability that it is divisible either by 2 or by 3? Let $A =$ divisible by 2, and $B =$ divisible by 3. There are 500 numbers in A and 333 numbers in B, because $1000/3 = 333 + 1/3$ has integer part 333. More exactly,

$$A = \{2, 4, 6, \dots, 1000\}, \quad B = \{3, 6, 9, \dots, 996, 999\}.$$

$$\mathbf{P}(A) = \frac{500}{1000} = \frac{1}{2}, \quad \mathbf{P}(B) = \frac{333}{1000}.$$

In addition, $A \cap B = \{\text{divisible by } 6\}$ contains 166 numbers: $1000/6$ has integer part 166. Therefore,

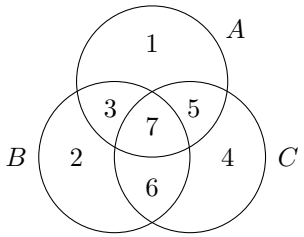
$$\mathbf{P}(A \cap B) = 166/1000, \quad \mathbf{P}(A \cup B) = \frac{500}{1000} + \frac{333}{1000} - \frac{166}{1000} = \boxed{\frac{667}{1000}}$$

2.4. **Inclusion-exclusion formula for three events.** We can state an analogous formula for three, four, or more events. For simplicity, let us state it for 3 events.

Theorem 2.2. *For any events A, B, C, we have:*

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

Informal explanation: To correct for double counting in double intersections, we remove them. But this removes $A \cap B \cap C$ three times, and it was initially counted three times too (as part of A, B, C). Thus we need to add it once again.



Proof. Let p_1 be the probability of the event 1 on the Venn diagram, which is $A \setminus (B \cup C)$. Let p_2 be the probability of the event 2, etc. Then

$$\begin{aligned}
\mathbf{P}(A) &= p_1 + p_3 + p_5 + p_7,\\
\mathbf{P}(B) &= p_2 + p_3 + p_6 + p_7,\\
\mathbf{P}(C) &= p_4 + p_6 + p_7 + p_5,\\
\mathbf{P}(A \cap B) &= p_3 + p_7,\\
\mathbf{P}(A \cap C) &= p_5 + p_7,\\
\mathbf{P}(B \cap C) &= p_6 + p_7,\\
\mathbf{P}(A \cap B \cap C) &= p_7.
\end{aligned}$$

Finally, $\mathbf{P}(A \cup B \cup C) = p_1 + p_2 + p_3 + p_4 + p_5 + p_6 + p_7$. Plugging this into the formula in the box above, we can check it is indeed true. □

Example 2.8. Out of numbers $1, \dots, 600$, how many are divisible by either 2, 3, or 5? We have: $A = \{\text{divisible by } 2\}$, $B = \{\text{divisible by } 3\}$, $C = \{\text{divisible by } 5\}$. Then $A \cap B = \{\text{divisible by } 6\}$, $A \cap C = \{\text{divisible by } 10\}$, $B \cap C = \{\text{divisible by } 15\}$, $A \cap B \cap C = \{\text{divisible by } 30\}$. Thus

$$\mathbf{P}(A) = \frac{1}{2},\ \mathbf{P}(B) = \frac{1}{3},\ \mathbf{P}(C) = \frac{1}{5},\ \mathbf{P}(A \cap B) = \frac{1}{6},\ \mathbf{P}(A \cap C) = \frac{1}{10},\ \mathbf{P}(B \cap C) = \frac{1}{15},\ \mathbf{P}(A \cap B \cap C) = \frac{1}{30}.$$

By the inclusion-exclusion formula,

$$\mathbf{P}(A \cup B \cup C) = \frac{1}{2} + \frac{1}{3} + \frac{1}{5} - \frac{1}{6} - \frac{1}{10} - \frac{1}{15} + \frac{1}{30} = \frac{11}{15}.$$

2.5. **Conditional probability.** A bag with 11 cubes:

- 3 red and fuzzy
- 2 red and smooth
- 4 blue and fuzzy
- 2 blue and smooth

Put your hand in the bag and randomly pick a cube. Let

$$R = \{\text{the cube is red}\},\ F = \{\text{the cube is fuzzy}\}.$$

Then the probability that it is red is $\mathbf{P}(R) = 5/11$. But if you feel that it is fuzzy, then the probability that it is red is $3/7$. This is called *conditional probability* of R given F:

$$\mathbf{P}(R \mid F) = \frac{\mathbf{P}(R \cap F)}{\mathbf{P}(F)} = \frac{3}{7}.$$

Definition 2.1. Conditional probability of A given B is defined as

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}.$$

Corollary 2.3. $\mathbf{P}(A \cap B) = \mathbf{P}(B)\mathbf{P}(A \mid B) = \mathbf{P}(A)\mathbf{P}(B \mid A)$.

Example 2.9. For familiar with two children, assume that all probabilities of two kids are equally likely: $S = \{GG, GB, BG, BB\}$. Let $A =$ at least one boy, $B =$ at least one is girl. Then

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} = \frac{2/4}{3/4} = \frac{3}{4}.$$

Example 2.10. $\mathbf{P}(A \backslash B) = 0.2, \mathbf{P}(B \backslash A) = 0.1, \mathbf{P}(A^c \cap B^c) = 0.6$. Find $\mathbf{P}(A \mid B)$. Solution: $\mathbf{P}(A \cup B) = 1 - 0.6 = 0.4$, and $\mathbf{P}(A \cap B) = 0.4 - 0.2 - 0.1 = 0.1$. Thus $\mathbf{P}(B) = 0.1 + 0.1 = 0.2$ and $\mathbf{P}(A \mid B) = 0.1/0.2 = 50\%$.

2.6. **Total probability formula.** Take have events $F_1, \ldots, F_n$ such that

$$F_1 \cup F_2 \cup \ldots \cup F_n = \Omega, \quad F_i \cap F_j = \varnothing, \ i \neq j.$$

This means one and only one of the events $F_1, \ldots, F_n$ happens. This collection of events is called a *partition.* An example is Heads or Tails coin toss; or numbers from 1 to 6 on a die.

Theorem 2.4. *For an event A and a partition $F_1, \ldots, F_n$, we have:*

$$\mathbf{P}(A) = \sum_{k=1}^{n} \mathbf{P}(A \cap F_k) = \sum_{k=1}^{n} \mathbf{P}(A \mid F_k)\mathbf{P}(F_k).$$

Proof. We have from conditional probability definition: $\mathbf{P}(A \mid F_i)\mathbf{P}(F_i) = \mathbf{P}(A \cap F_i)$. But

$$\mathbf{P}(A \cap F_1) + \ldots + \mathbf{P}(A \cap F_n) = \mathbf{P}(A),$$

because the union of events $A \cap F_1, \ldots, A \cap F_n$ is A, and these do not intersect. □

Example 2.11. A policyholder is classified as *prime* or *subprime*, with probabilities 20% and 80%, respectively. Prime policyholders have accidents with probability 2%, and subprime with probability 10%. Find the probability that a random policyholder has an accident. Here, $F_1 = \{\text{prime}\}$, $F_2 = \{\text{subprime}\}$, $A = \{\text{accident}\}$. Then

$$\mathbf{P}(F_1) = 0.2, \ \mathbf{P}(F_2) = 0.8, \ \mathbf{P}(A \mid F_1) = 0.02, \ \mathbf{P}(A \mid F_2) = 0.1.$$

Thus by the total probability formula

$$\mathbf{P}(A) = \mathbf{P}(A \mid F_1)\mathbf{P}(F_1) + \mathbf{P}(A \mid F_2)\mathbf{P}(F_2) = 0.2 \cdot 0.02 + 0.8 \cdot 0.1 = 8.4\%.$$

Example 2.12. Flip a fair coin. If H, roll one die. If T, roll two dice. What is the probability p that there is at least one six? Let X_1, X_2 be the first and the second dice. Let $F_1 =$ heads, $F_2 =$ tails, $A =$ at least one six. But $\mathbf{P}(X_1 = 6) = 1/6$, and by inclusion-exclusion formula we have:

$$\mathbf{P}(X_1 = 6 \,\text{or}\, X_2 = 6) = \mathbf{P}(X_1 = 6) + \mathbf{P}(X_2 = 6) - \mathbf{P}(X_1 = 6, \ X_2 = 6) = \frac{1}{6} + \frac{1}{6} - \frac{1}{36} = \frac{11}{36}.$$

Thus by the total probability formula,

$$p = \frac{1}{2} \cdot \frac{1}{6} + \frac{1}{2} \cdot \frac{11}{36} = \frac{17}{72}.$$

2.7. **Bayes' formula.** In the setting above, we have:

$$\mathbf{P}(F_1 \mid A) = \frac{\mathbf{P}(F_1 \cap A)}{\mathbf{P}(A)} = \frac{\mathbf{P}(A \mid F_1)\mathbf{P}(F_1)}{\mathbf{P}(A \mid F_1)\mathbf{P}(F_1) + \ldots + \mathbf{P}(A \mid F_n)\mathbf{P}(F_n)}.$$

Example 2.13. 1% of people have a disease. If a person is sick, the test is positive with probability 90%, and if a person is healthy, the test is positive only with probability 20%. A random person is selected from the population and is tested positive. What is the probability that he is sick? Let $F_1 =$ sick, $F_2 =$ healthy, $A =$ tested positive.

$$\mathbf{P}(F_1) = 0.01, \ \mathbf{P}(F_2) = 0.99, \ \mathbf{P}(A \mid F_1) = 0.9, \ \mathbf{P}(A \mid F_2) = 0.2.$$

$$\mathbf{P}(F_1 \mid A) = \frac{0.01 \cdot 0.9}{0.01 \cdot 0.9 + 0.99 \cdot 0.2} \approx 4.3\%$$

We updated the probability of our *hypothesis* (that he is sick) from 1% to 4.3%, using new information that the test is positive. The probability $\mathbf{P}(F_1)$ is called a *prior probability*, and $\mathbf{P}(F_1 \mid A)$ is called a *posterior probability.* But this probability is still small, because the original probability was small. This is *prosecutor's fallacy:* Even though the probability 90% is large, the posterior probability is small, because the prior is small.

Example 2.14. We have a fair coin and a magic coin, which always comes out H. Choose one at random (each can be chosen with probability 1/2), flip it twice. It comes out H both times. What is the probability it is fair? Let F_1 = fair, F_2 = magic, A = both heads. Then

$$\mathbf{P}(F_1) = \mathbf{P}(F_2) = \frac{1}{2}, \quad \mathbf{P}(A \mid F_1) = \frac{1}{4}, \ \mathbf{P}(A \mid F_2) = 1.$$

Therefore, by Bayes' formula we get:

$$\mathbf{P}(F_1 \mid A) = \frac{\frac{1}{2} \cdot \frac{1}{4}}{\frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2} \cdot 1} = \frac{1}{5} = 20\%.$$

2.8. **Independence for two events.** Events A and B are called *independent* if knowledge of whether A happened or not does not influence the probability of B: $\mathbf{P}(B \mid A) = \mathbf{P}(B)$. Since $\mathbf{P}(B \mid A) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(A)}$, we can rewrite this as

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B).$$

Example 2.15. Toss the coin twice. Let A = first toss H, B = second toss H, C = both tosses the same. Then A and B are independent. Indeed, the probability space (the space of all outcomes) is $S = \{HH, HT, TH, TT\}$, and $A = \{\text{HH, HT}\}$, $B = \{\text{TH, HH}\}$, $C = \{\text{TT, HH}\}$. Next,

$$\mathbf{P}(A \cap B) = \mathbf{P}\{HH\} = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \mathbf{P}(A) \cdot \mathbf{P}(B).$$

It is obvious that these events are independent, because they result in different tosses of the coin. In some other cases, it is not obvious. For example, A and C are also independent. Indeed,

$$\mathbf{P}(A \cap C) = \mathbf{P}\{HH\} = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \mathbf{P}(A) \cdot \mathbf{P}(C).$$

Similarly, B and C are independent.

Example 2.16. A collateralized debt obligation (CDO) is backed by 5 subprime mortgages from California, each defaults with probability 40%. A *senior tranch* defaults only if *all* 5 mortgages default. Find the probability of this in two cases: (a) they are independent; (b) they default simultaneously because of housing bubble. Solution: (a) $0.4^5 \approx 1\%$; (b) 40%. The case (b) is roughly what happened in 2007-8.

Example 2.17. Mary and Carlos are debugging a code. Their result are independent. Mary can spot a mistake with probability 60%, and Carlos with probability 80%. What is the probability that this mistake remains? Answer: $(1 - 60\%) \cdot (1 - 80\%) = 0.08$.

Example 2.18. A gene exhibits two *alleles:* Dominant W and recessive w, with

$$\mathbf{P}(WW) = 50\%, \ \mathbf{P}(Ww) = 20\%, \ \mathbf{P}(ww) = 30\%.$$

What is the chance that the offspring is ww? Needs to receive w from father and w from mother, which are independent. The probability of getting w from the father is

$$\frac{1}{2} \cdot 0.2 + 0.3 = 0.4.$$

The same probability is of getting w from mother. Thus the final probability is $0.4^2 = 0.16$.

2.9. **Independence for three or more events.** Events A, B and C are called *independent* if

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B), \quad \mathbf{P}(A \cap C) = \mathbf{P}(A)\mathbf{P}(C), \quad \mathbf{P}(B \cap C) = \mathbf{P}(B)\mathbf{P}(C),$$

$$\mathbf{P}(A \cap B \cap C) = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C).$$

This last condition is important, because it does not automatically follow from the first three conditions. For example, if A, B and C are the events from the example, then A and B are independent, B and C are independent, A and C are independent, so these events are *pairwise independent*. But $A \cap B \cap C = \text{HH}$, so

$$\mathbf{P}(A \cap B \cap C) = \frac{1}{4} \neq \frac{1}{8} = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C).$$

Example 2.19. Three clients, each has accident independently with probability 20%. Let X be the number of total accidents. Then $\mathbf{P}(X = 0) = (1 - 0.2)^3 = 0.512$, and $\mathbf{P}(X = 3) = 0.2^3 = 0.008$.

3. Random Variables

3.1. **Definition.** Take a sample space S.

Definition 3.1. Any function $X : S \to \mathbb{R}$ is called a *random variable*. The *distribution* of a random variable is the list of its values, together with probabilities of each value.

Example 3.1. For $S = \{\text{HH, HT, TH, TT}\}$ (two coin tosses), let X be the number of Heads. Then

$$X(HH) = 2, \quad X(TH) = 1, \quad X(HT) = 1, \quad X(TT) = 0.$$

The distribution, otherwise called the *probability mass function*, is

$$P(0) = \frac{1}{4}, \quad P(1) = \frac{1}{2}, \quad P(2) = \frac{1}{4}.$$

Example 3.2. Two dice rolled, $S = \{11, 12, \dots, 66\}$. Let X be the difference. Then $X(12) = 1$, $X(31) = 2$.

Some random variables are *continuous*. They take infinitely many values, and have *density* $f(x)$: For all a, b,

$$\mathbf{P}(a \le X \le b) = \int_a^b f(x)\,\mathrm{d}x.$$

Thus the probability of getting between a and b is the area under the density curve. For such random variables, $\mathbf{P}(X = c) = 0$ for all c. Any particular value is taken with probability 0. To be a density, a function needs to be nonnegative and integrate from $-\infty$ to $+\infty$ up to 1. For example, $f(x) = x$ or $f(x) = x^2$ do not pass this test.

Example 3.3. For an exponential random variable with unit mean,

$$f(x) = \begin{cases} e^{-x}, \ x > 0; \\ 0, \ x \le 0. \end{cases}$$

Then $\mathbf{P}(X \le 1) = \int_0^1 f(x)\,\mathrm{d}x = \int_0^1 e^{-x}\,\mathrm{d}x = 1 - e^{-1}$.

Actually, random variables can be neither discrete nor continuous, but a mixture of both. Say, an accident happens with probability 10%. In this case, losses are exponential, with distribution as above. Thus the amount X of losses is distributed as (for $0 < a < b$)

$$\mathbf{P}(X = 0) = 0.9, \quad \mathbf{P}(a \le X \le b) = \frac{1}{10}\int_a^b e^{-x}\,\mathrm{d}x.$$

Example 3.4. We have distribution of X:

$$p(0) = 0.5,\ p(1) = 0.2,\ p(2) = p(3) = p(4) = 0.1$$

Find $\mathbf{P}(X \ge 3 \mid X \ge 1)$: this is equal to

$$\frac{\mathbf{P}(X \ge 3)}{\mathbf{P}(X \ge 1)} = \frac{0.1 + 0.1}{1 - \mathbf{P}(X = 0)} = \frac{0.2}{0.5} = 40\%.$$

Example 3.5. A random variable X has density $f(x) = 2x$, $0 \le x \le 1$. Outside of this interval, the density is 0 (often we skip this 0 outside of effective interval). Then

$$\mathbf{P}(X \ge 0.5) = \int_{0.5}^1 2x\,\mathrm{d}x = \frac{3}{4}.$$

3.2. **Cumulative distribution function.** For a random variable X, its *cumulative distribution function* (CDF) is $F_X(x) := \mathbf{P}(X \le x)$. It exists for every random variable, discrete, continuous, or otherwise (unlike PDF, which exists only for continuous distributions), and it uniquely determines the distribution: If $F_X(x) = F_Y(x)$ for all x, then random variables X and Y have the same distribution.

Theorem 3.1. *(a) This function is nondecreasing: For $x \le y$, we have $F_X(x) \le F_X(y)$.*
(b) $F_X(x) \to 1$ as $x \to +\infty$. In other words, $F_X(+\infty) = 1$.
(c) $F_X(x) \to 0$ as $x \to -\infty$. In other words, $F_X(-\infty) = 0$.

Proof. (a) If $X \le x$, then $X \le y$. In other words, $\{X \le x\} \subseteq \{X \le y\}$. Thus $\mathbf{P}(X \le x) \le \mathbf{P}(X \le y)$. In other words, $F_X(x) \le F_X(y)$. (b) Always $X \le +\infty$ true, with probability 1. (c) Never $X \le -\infty$, with probability 0. □

Example 3.6. Take X which takes values 0 and 1 with equal probabilities 0.5. Then:

(1) $x < 0$. Then the event $X \le x$ cannot happen, because X can take only values 0 or 1. Therefore, the probability of this event is zero: $F_X(x) = 0$.

(2) $x \in [0,1)$. Then the event $X \le x$ is equivalent to $X = 0$, which has probability $1/2$. Therefore, the probability of this event is 0.5: $F_X(x) = 0.5$.
(3) $x \ge 1$. Then the event $X \le x$ always happens, because X can take only values 0 or 1. Therefore, the probability of this event is one: $F_X(x) = 1$.

$$F_X(x) = \begin{cases} 0, \, x < 0; \\ 1/2, \, 0 \le x < 1; \\ 1, \ x \ge 1. \end{cases}$$

Example 3.7. Take X the (random) number of successful trials in two independent trials, each with probability $1/3$ of success (and $2/3$ of failure). Then

$$\mathbf{P}(X = 0) = \left(\frac{2}{3}\right)^2 = \frac{4}{9}, \, \mathbf{P}(X = 1) = 2 \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{4}{9}, \, \mathbf{P}(X = 2) = \left(\frac{1}{3}\right)^2 = \frac{1}{9}.$$

(1) $x < 0$. Then the event $X \le x$ cannot happen. Therefore, the probability of this event is zero: $F_X(x) = 0$.
(2) $x \in [0,1)$. Then the event $X \le x$ is equivalent to $X = 0$, which has probability $4/9$. Therefore, $F_X(x) = 4/9$.
(3) $x \in [1,2)$. Then the event $X \le x$ is equivalent to $X \in \{0,1\}$, which has probability $4/9 + 4/9 = 8/9$. Therefore, $F_X(x) = 4/9$.
(4) $x \ge 2$. Then $X \le x$ always happens, with probability 1; therefore, $F_X(x) = 1$.

$$F_X(x) = \begin{cases} 0, \, x < 0; \\ 4/9, \, 0 \le x < 1; \\ 8/9, \, 1 \le x < 2; \\ 1, x \ge 2. \end{cases}$$

We can see that for a discrete random variable X, the CDF is piecewise constant. Jump points correspond to the values which X takes, and the size of the jump corresponds to the probability that this value is taken.

Example 3.8. Let us solve the opposite problem: Take a cumulative distribution function of a random variable X:

$$F_X(x) = \begin{cases} 0, \, x < 1; \\ 0.4, \, x \in [1,2); \\ 1, \, x \ge 2. \end{cases}$$

Let us find the distribution of X: which values it takes and with which probabilities. It jumps at 1 and 2, and the size of jumps is 0.4 and 0.6 respectively. Thus $\mathbf{P}(X = 1) = 0.4$ and $\mathbf{P}(X = 2) = 0.6$.

For a continuous random variable X with density p_X, then

$$F_X(x) = \mathbf{P}(X \le x) = \int_{-\infty}^{x} p_X(y) \, \mathrm{d}y.$$

Thus F_X is continuous. If p_X is continuous at the point x, then $F_X'(x) = p_X(x)$. That is, the derivative of this CDF is the PDF: probability density function, or simply density.

Example 3.9. Find CDF for X with the density

$$p(x) = \begin{cases} 2e^{-2x}, \, x \ge 0; \\ 0, \, x < 0. \end{cases}$$

Thus for $x < 0$ the event $\{X \le x\}$ is impossible, because $X \ge 0$ always. Therefore, $F_X(x) = \mathbf{P}(X \le x) = 0$. And for $x \ge 0$ we have:

$$F_X(x) = \int_0^x (2e^{-2y}) \, \mathrm{d}y = \left(-e^{-2y}\right)\Big|_{y=0}^{y=x} = 1 - e^{-2x}.$$

Let us summarize:

$$F_X(x) = \begin{cases} 0, \, x < 0; \\ 1 - e^{-2x}, \, x \ge 0. \end{cases}$$

Unlike in the exampels above, $F_X(x) < 1$ for all x, but $F_X(x) \to 1$ as $x \to +\infty$.

Example 3.10. Find the distribution of a random variable X with CDF

$$F_X(x) = \begin{cases} 0, \, x < 0; \\ x^2, \, x \in [0, 1); \\ 1, \, x \geq 1. \end{cases}$$

We have: $\mathbf{P}(X \leq 0) = F_X(0) = 0$, thus $X > 0$ with probability 1. Next, $\mathbf{P}(X \leq 1) = F_X(1) = 1$, thus $X \leq 1$ with probability 1. Thus with probability 1, $0 < X \leq 1$. And for $x \in (0, 1)$, the density is the derivative of F_X: $p_X(x) = (x^2)' = 2x$. Thus,

$$p_X(x) = \begin{cases} 2x, \, 0 \leq x \leq 1; \\ 0, \, \text{else}. \end{cases}$$

A common mistake is to assign, say, for random variable X lying between 0 and 2, $F_X(x) = 0$ for $x \geq 2$. This confuses CDF F_X with PDF p_X. But while $p_X(x) = 0$ in this case for $x \leq 0$ and $x \geq 2$, instead for CDF we have:

$$F_x(x) = \begin{cases} 0, \, x < 0; \\ 1, \, x \geq 2. \end{cases}$$

3.3. **Expectation.** Take a random variable $X : S \to \mathbb{R}$ on a sample space S.

Definition 3.2. The *expectation* of X is the sum over all elementary outcomes:

$$\mathbb{E}\,[X] = \sum_{s \in S} p(s) X(s).$$

This is also called *mean*, or *average*, or *expected value*.

Example 3.11. $S = \{\text{HH, HT, TH, TT}\}$ (two fair coins). Let X be the number of Heads. Then

$$X(HH) = 2, \quad X(TH) = 1, \quad X(HT) = 1, \quad X(TT) = 0.$$

Each of these 4 elementary outcomes has probabiltiy 1/4. The distribution, otherwise called the *probability mass function*, is

$$P(0) = \frac{1}{4}, \quad P(1) = \frac{1}{2}, \quad P(2) = \frac{1}{4}.$$

The expectation can be computed in two equivalent ways:

$$\mathbb{E}[X] = 0 \cdot P(0) + 1 \cdot P(1) + 2 \cdot P(2) = \frac{1}{4} \cdot X(TT) + \frac{1}{4} \cdot X(HT) + \frac{1}{4} \cdot X(TH) + \frac{1}{4} \cdot X(HH) = 1.$$

In general, we have the following results.

Lemma 3.2. *Assume X takes values $c_1, \ldots, c_n$, with probabilities $p_1, \ldots, p_n$. Then the expectation is*

$$\mathbb{E}[X] = c_1 p_1 + \ldots + c_n p_n.$$

If X is not discrete (that is, does not take finitely or countably many values), then it is harder to define expectation (we need *measure theory*). However, it is easy to compute it.

Lemma 3.3. *If X is continuous with density f, then*

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f(x) \, \mathrm{d}x.$$

Indeed, split real line into small subintervals $[y, y + \mathrm{d}y]$. Then $\mathbb{P}(y \leq X \leq y + \mathrm{d}y) \approx f(y) \, \mathrm{d}y$. Sum this over all such small intervals and get the integral.

Example 3.12. For a uniform random variable X on $[0, 1]$ with density $f(x) = 1$, $0 \leq x \leq 1$. Then

$$\mathbb{E}[X] = \int_0^1 x f(x) \, \mathrm{d}x = \int_0^1 x \, \mathrm{d}x = \frac{1}{2}.$$

Example 3.13. Sometimes the expectation can be infinite. Take density $f(x) = x^{-2}$ for $x \geq 1$. Then

$$\mathbb{E}[X] = \int_1^\infty x f(x) \, \mathrm{d}x = \int_1^\infty x^{-1} \, \mathrm{d}x = \ln x |_{x=1}^{x=\infty}.$$

Example 3.14. (St Petersburg Paradox.) A fair coin is flipped until first tail. Win 2 if it appears on the first toss (which happens with probability 1/2); Win 4 if on the second toss (which happens with probability 1/4); etc. Thus $\mathbb{E}[X] = 2 \cdot \frac{1}{2} + 4 \cdot \frac{1}{4} + \ldots = \infty$. Thus the value of this game is infinite. But would you pay, say 1000 to play it?

Example 3.15. Alaska Airlines flight Reno-Seattle on Embraer 75 has 64 economy class seats. Each ticket costs 100$. The airline overbooks: sells an additional tickets for full flight. If all passengers show up, the airline needs to refund 100$ and give a voucher 150$ to this bumped passenger. Each passenger will show up independently with probability 98%. Is this airline better off overbooking? Solution: Airline gains 100$ if at least one does not show up. But it loses 150$ if all show up, which happens with probability $0.98^{64} \approx 0.27$. Thus it is better to overbook: Expected gain is

$$100 \cdot (1 - 0.27) + (-150) \cdot 0.27 > 0.$$

Lemma 3.4. *For a function $g : \mathbb{R} \to \mathbb{R}$, we get:*

$$\mathbb{E}[g(X)] = \begin{cases} \int_{-\infty}^{+\infty} g(y) f(y)\,\mathrm{d}y, \\ \sum_c g(c)\mathbf{P}(X = c), \end{cases}$$

if X has density f or discrete distribution, respectively.

Example 3.16. For X with density $f(x) = e^{-x}$, $x > 0$, we get:

$$\mathbb{E}\big[e^{X/2}\big] = \int_0^\infty e^{x/2} e^{-x}\,\mathrm{d}x = \int_0^\infty e^{-x/2}\,\mathrm{d}x = 2.$$

3.4. **Variance.** This quantity measures the concentration of random variable around its mean.

Definition 3.3. For a random variable X, if $\mu = \mathbb{E}[X]$, then $\operatorname{Var} X = \mathbb{E}(X - \mu)^2$ is called *variance*.

This quantity is always nonnegative, and it can be zero only when $X = \mu$ is constant, not random.

Theorem 3.5. *One can compute variance as follows:* $\operatorname{Var} X = \mathbf{E}[X^2] - (\mathbf{E}X)^2$.

Proof. Note that $(X-\mu)^2 = X^2 - 2X\mu + \mu^2$. Take expectation and get: $\mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2 = \mathbb{E}[X^2] - 2\mu\cdot\mu + \mu^2 = \mathbb{E}[X^2] - \mu^2$. □

Example 3.17. For a uniform random variable X on $[0, 1]$, its density is $f(x) = 1$ for $0 \le x \le 1$. Thus

$$\mathbb{E}[X^2] = \int_0^1 x^2 f(x)\,\mathrm{d}x = \int_0^1 x^2\,\mathrm{d}x = \frac{1}{3}.$$

We can compute variance: $\operatorname{Var}(X) = 1/3 - (1/2)^2 = 1/12$.

Example 3.18. Let X be the number of Heads in two coin tosses. Its distribution is $p(0) = 1/4$, $p(1) = 1/2$, $p(2) = 1/4$. Thus

$$\mathbb{E}[X^2] = \frac{1}{4} \cdot 0^2 + \frac{1}{2} \cdot 1^2 + \frac{1}{4} \cdot 2^2 = \frac{3}{2},$$

and the variance is equal to $3/2 - 1^2 = 1/2$.

3.5. **Median.** For a continuous random variable X, its median is defined as the m such that

$$\mathbf{P}(X \ge m) = \int_m^\infty f(x)\,\mathrm{d}x = \frac{1}{2}.$$

Example 3.19. $f(x) = e^{-x}$, $x \ge 0$. Then $\mathbf{P}(X \ge m) = e^{-m} = 1/2$ and thus $m = \ln 2$. This is different from mean $\mathbb{E}[X] = 1$.

For discrete random variables, it is sometimes not possible to find m such that $\mathbf{P}(X \ge m) = \mathbf{P}(X \le m) = 1/2$. Instead, we should start from the smallest value and add probabilities until we reach or exceed $1/2$.

Example 3.20. For the distribution $p(-2) = 0.2$, $p(-1) = 0.1$, $p(0) = 0.4$, $p(2) = 0.2$, $p(3) = 0.1$, we get to 0 with probability $0.2 + 0.1 + 0.4 = 0.7$. Thus the median is $(-1 + 0)/2 = -0.5$.

Example 3.21. For the distribution $p(-2) = 0.2$, $p(-1) = 0.1$, $p(0) = 0.2$, $p(2) = 0.2$, $p(3) = 0.3$, we get to 0 with probability $0.2 + 0.1 + 0.4 = 0.5$. Thus the median is 0.

Definition 3.4. A mode x is the x such that $f(x)$ is maximal, or $\mathbf{P}(X = x)$ is maximal.

For example, for $f(x) = e^{-x}$, $x \ge 0$, and the mode is 0.

3.6. **Survivial function.** For a random variable $X \geq 0$, the function $S(c) = \mathbf{P}(X > c)$ is called the *survival* or *tail function*. Let X be the lifetime with density f. The *mortality rate* is defined as

$$m(x) = \frac{f(x)}{S(x)},$$

which has meaning

$$m(x)\,\mathrm{d}x = \mathbb{P}(x < X < x + \mathrm{d}x \mid X > x).$$

This is the conditional probability that I will survive for the additional $\mathrm{d}x$, given that I survived until x. These survival functions and mortality rates are available in the actuarial tables in Social Security Administration.

3.7. **Deductibles and copays.** An insurance policy has out-of-pocket maximum of 2 and deductible 1, and actuarial value 70%. This means the follows: The first 1 is paid by the client, then the company pays 70% until the out-of-pocket reaches 2. Starting from this, the company pays everything. Assume the loss distribution is

$$\mathbf{P}(X = n) = \frac{1}{(n+1)(n+2)},\ n = 0, 1, 2, \ldots$$

Let Y be the payment by the client. The distributions of X and Y are given in Table below.

X	Y	P
0	0	1/2
1	1	1/6
2	1.3	1/12
3	1.6	1/20
4	1.9	1/30
≥ 5	2	p

The probability p is equal to

$$p = 1 - \frac{1}{2} - \frac{1}{6} - \frac{1}{12} - \frac{1}{20} - \frac{1}{30} = \frac{1}{6}.$$

We can see that the expected out-of-pocket payment is

$$\mathbb{E}[Y] = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{6} + 1.3 \cdot \frac{1}{12} + 1.6 \cdot \frac{1}{20} + 1.9 \cdot \frac{1}{30} + 2 \cdot \frac{1}{6} \approx 0.75.$$

3.8. **Transformations.** Here we take functions of continuous random variables with known densities. We find densities of new random variables. Indeed, take X with $f_X(x) = 2e^{-2x}$, $x \geq 0$. Find density of $X^2 = Y$. There are two ways: change-of-variables and CDF.

Change of variables. For every $0 < a < b$,

$$\mathbf{P}(a \leq X \leq b) = \int_a^b f_X(x)\,\mathrm{d}x.$$

Try to do the same for Y instead of X:

$$\mathbf{P}(a \leq Y \leq b) = \mathbb{P}\left(a^{1/2} \leq X \leq b^{1/2}\right) = \int_{a^{1/2}}^{b^{1/2}} f_X(x)\,\mathrm{d}x.$$

Next, change variables in the integral:

$$y^{1/2} = x;\ a^{1/2} \leq x \leq b^{1/2} \Leftrightarrow a \leq y \leq b;\ \mathrm{d}x = \frac{1}{2} y^{-1/2}\,\mathrm{d}y.$$

Thus we get:

$$\int_{a^{1/2}}^{b^{1/2}} f_X(x)\,\mathrm{d}x = \int_a^b 2e^{-2y^{1/2}} \frac{1}{2} y^{-1/2}\,\mathrm{d}y.$$

Combining this, we get:

$$\mathbf{P}(a \leq Y \leq b) = \int_a^b e^{-2y^{1/2}} y^{-1/2}\,\mathrm{d}y.$$

Since this is true for all $a < b$, we get that the density of Y is

$$f_Y(y) = e^{-2y^{1/2}} y^{-1/2},\ y > 0.$$

CDF. Compute CDF for X: $\mathbf{P}(X \leq c) = 1 - e^{-2c}$, $c > 0$. Thus

$$\mathbf{P}(Y \leq c) = \mathbf{P}(X \leq c^{1/2}) = 1 - e^{-2c^{1/2}},\ c > 0.$$

The derivative of this function is the density of Y:

$$f_Y(c) = \left(1 - e^{-2c^{1/2}}\right)' = -(-2c^{1/2})' e^{-2c^{1/2}} = c^{-1/2} e^{-2c^{1/2}},\ c > 0.$$

The first method seems to be better when you cannot compute the CDF explicitly by integrating density.

4. Joint Distributions

4.1. **Joint distributions.** Take two random variables X and Y with joint distribution: If they are discrete, they have *joint probability mass function*

$$\mathbb{P}(X = x,\, Y = y) = p(x, y).$$

If they are continuous, they have *joint density*, or *joint probability density function* $f(x, y)$ such that for all a, b, c, d,

$$\mathbf{P}(a \le X \le b,\, c \le Y \le d) = \int_a^b \int_c^d f(x, y)\, \mathrm{d}y\, \mathrm{d}x.$$

To qualify as joint density, a function needs to be nonnegative: $f \ge 0$, and integrate up to 1 over the coordinate plane $\mathbb{R}^2$:

$$\iint_{\mathbb{R}^2} f(x, y)\, \mathrm{d}x\, \mathrm{d}y = 1.$$

Example 4.1. Toss a coin twice. Let $X = 1$ if first Heads, $X = 0$ if first Tails, and Y be the number of Heads.

s	HH	HT	TH	TT
X	1	1	0	0
Y	2	1	1	0
P	1/4	1/4	1/4	1/4

Example 4.2. Take joint density $f(x, y) = cx^2 y$ if $x^2 \le y \le 1$ (and 0 for other x and y) for some constant c. Find c and $\mathbf{P}(X \ge Y)$. Solution:

$$\begin{aligned}
1 &= c \int_{-1}^{1} \int_{x^2} 1 x^2 y\, \mathrm{d}y\, \mathrm{d}x = c \int_{-1}^{1} \frac{x^2 y^2}{2}\bigg|_{y=x^2}^{y=1} \mathrm{d}x \\
&= c \int_{-1}^{1} \left(x^2 - x^6\right)\, \mathrm{d}x = \frac{c}{2} \left(\frac{x^3}{3} - \frac{x^7}{7}\right)\bigg|_{x=-1}^{x=1} = \frac{4c}{21}; \\
\mathbf{P}(X \le Y) &= \frac{21}{4} \int_0^1 \int_{x^2}^{x} x^2 y\, \mathrm{d}y\, \mathrm{d}x = \frac{21}{4} \int_0^1 x^2\, \frac{y^2}{2}\bigg|_{y=x^2}^{y=x} \mathrm{d}x \\
&= \frac{21}{4} \int_0^1 \frac{x^2}{2} \left(x^2 - x^4\right)\, \mathrm{d}x = \frac{21}{8} \int_0^1 \left(x^4 - x^6\right)\, \mathrm{d}x = \frac{21}{8} \cdot \frac{2}{35} = \frac{3}{20}.
\end{aligned}$$

Theorem 4.1. *For any function $g(x, y)$, we have:*

$$\mathbb{E}g(X, Y) = \begin{cases} \sum_{x,y} g(x, y) \mathbf{P}(X = x,\, Y = y), \\ \iint_{\mathbb{R}^2} g(x, y) f(x, y)\, \mathrm{d}y\, \mathrm{d}x. \end{cases}$$

if (X, Y) are jointly discrete or jointly continuous with density f.

Example 4.3. For $f(x, y) = x + y$, $0 \le x \le 1$, $0 \le y \le 1$, we get:

$$\begin{aligned}
\mathbb{E}[XY] &= \int_0^1 \int_0^1 xy(x + y)\, \mathrm{d}y\, \mathrm{d}x = \int_0^1 \left(\frac{x^2 y^2}{2} + \frac{xy^3}{3}\right)\bigg|_{y=0}^{y=1} \mathrm{d}x \\
&= \int_0^1 \left(\frac{x^2}{2} + \frac{x}{3}\right)\, \mathrm{d}x = \frac{1}{3}.
\end{aligned}$$

4.2. **Independent random variables.** Take random variables X and Y.

Definition 4.1. If X and Y are discrete, they are called *independent* if for all values x and y which they assume, events $\{X = x\}$ and $\{Y = y\}$ are independent:

$$\mathbf{P}(X = x,\, Y = y) = \mathbf{P}(X = x) \cdot \mathbf{P}(Y = y).$$

If they are continuous, they are called *independent* if for all a, b, c, d, the following events are independent: $\{a \le X \le b\}$ and $\{c \le Y \le d\}$.

The latter is equivalent to joint density having product form: $f(x,y) = f_X(x)f_Y(y)$, where f_X and f_Y are marginal densities of X and Y, respectively.

Theorem 4.2. *For independent X and Y, we have: $\mathbb{E}[XY] = \mathbb{E}X \cdot \mathbb{E}Y$.*

Proof. Say X and Y are discrete. Then apply the above to $g(x,y) = xy$.

$$\mathbb{E}[XY] = \sum_{x,y} xy\mathbf{P}(X = x,\, Y = y) = \sum_{x,y} xy \cdot \mathbf{P}(X = x) \cdot \mathbf{P}(Y = y)$$
$$= \sum_x x \cdot \mathbf{P}(X = x) \cdot \sum_y y\mathbf{P}(Y = y) = \mathbb{E}X \cdot \mathbb{E}Y.$$

Similarly it can be proved for continuous X and Y. □

Example 4.4. For joint density $f(x,y) = 6xy^2$, $0 \le x \le 1$, $0 \le y \le 1$. This is product form: $2x \cdot 3y^2$. Thus X and Y are independent. We have:

$$\mathbb{E}[XY] = \mathbb{E}X \cdot \mathbb{E}Y = \int_0^1 2x \cdot x\,\mathrm{d}x \cdot \int_0^1 3y^2 \cdot y\,\mathrm{d}y = \frac{2}{3} \cdot \frac{3}{4} = \frac{1}{2}.$$

Example 4.5. For $f(x,y) = x + y$, $0 \le x \le 1$, $0 \le y \le 1$, random variables X and Y are not independent, because the density is not product form.

Example 4.6. For $f(x) = 2$, $0 \le x \le 1$, $x \le y \le 1$, the density is not product form, because the effective domain is not a rectangle.

4.3. **Covariance and correlation.** These are measures of dependency between two random variables.

Definition 4.2. For two random variables X and Y, their *covariance* is defined as

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}X \cdot \mathbb{E}Y.$$

Their *correlation* is defined as

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}\,X}\sqrt{\mathrm{Var}\,Y}}.$$

Positive covariance/correlation implies that on average, if X increases, Y also increases. Negative covariance/correlation implies the reverse. Correlation is always between -1 and $+1$. If it is $+1$, then Y is a linear function of X: $Y = aX + b$, with $a > 0$. If it is -1, then $Y = aX + b$ with $a < 0$.

Example 4.7. Take random variables X and Y with joint distribution as in Table below.

X	1	2	3	4
Y	0	1	1	2
P	0.2	0.3	0.2	0.3

Then we have:

$$\mathbb{E}[X] = 1 \cdot 0.2 + 2 \cdot 0.3 + 3 \cdot 0.2 + 4 \cdot 0.3 = 2.6,$$
$$\mathbb{E}[Y] = 0 \cdot 0.2 + 1 \cdot 0.3 + 1 \cdot 0.2 + 2 \cdot 0.3 = 1.1,$$
$$\mathbb{E}[XY] = 1 \cdot 0 \cdot 0.2 + 2 \cdot 1 \cdot 0.3 + 3 \cdot 1 \cdot 0.2 + 4 \cdot 2 \cdot 0.3 = 3.6.$$

Thus $\mathrm{Cov}(X,Y) = 3.6 - 2.6 \cdot 1.1 = 0.74$. Next,

$$\mathrm{Var}\,X = \mathbb{E}[X^2] - (\mathbb{E}X)^2 = 0.2 \cdot 1^2 + 0.3 \cdot 2^2 + 0.2 \cdot 3^2 + 0.3 \cdot 4^2 - 2.6^2 = 1.24;$$
$$\mathrm{Var}\,Y = \mathbb{E}[Y^2] - (\mathbb{E}Y)^2 = 1.7 - 1.1^2 = 0.49.$$

Thus the correlation is almost perfect:

$$\rho(X,Y) = \frac{0.74}{\sqrt{1.24}\sqrt{0.49}} = 0.95.$$

If X and Y are independent then they are *uncorrelated:* $\rho(X,Y) = 0$. But the converse is not true: $\mathrm{Cov}(X,Y) = 0$ does not mean X and Y are independent.

Example 4.8. Take (X,Y) uniformly distributed on four points: $(0,1), (1,0), (0,-1), (-1,0)$. Then (because $XY = 0$ always)

$$\mathbb{E}[X] = \mathbb{E}[Y] = 0, \quad \mathbb{E}[XY] = 0.$$

Thus $\mathrm{Cov}(X,Y) = 0 - 0 \cdot 0 = 0$. But X and Y are dependent: If $X = 1$, then must be $Y = 0$.

X	0	1	1
Y	0	0	1
P	0.2	0.4	0.4

Theorem 4.3. *For all (not necessarily independent) X and Y,*

$$\operatorname{Var}(X+Y) = \operatorname{Var} X + \operatorname{Var} Y + 2\operatorname{Cov}(X,Y).$$

In particular, for independent X and Y we have:

$$\operatorname{Var}(X+Y) = \operatorname{Var} X + \operatorname{Var} Y.$$

Proof. If $\mathbb{E}X = \mathbb{E}Y = 0$, we have:

$$\operatorname{Var}(X+Y) = \mathbb{E}[(X+Y)^2] = \mathbb{E}(X^2 + 2XY + Y^2) = \mathbb{E}[X^2] + \mathbb{E}[Y^2] + 2\mathbb{E}[XY].$$

But $\operatorname{Var} X = \mathbb{E}[X^2]$, $\operatorname{Var} Y = \mathbb{E}[Y^2]$, and $\mathbb{E}[XY] = \operatorname{Cov}(X,Y)$. This completes the proof. The case of general $\mathbb{E}X$ and $\mathbb{E}Y$ is similar. □

Example 4.9. Similarly we can get a more general linear combination: $\operatorname{Var}(2X - 3Y + 1) = \operatorname{Var}(2X) + \operatorname{Var}(-3Y) + 2\operatorname{Cov}(2X, -3Y) = 4\operatorname{Var}(X) + 9\operatorname{Var}(Y) - 12\operatorname{Cov}(X,Y)$.

Example 4.10. Invest in 3 independent stocks with returns S_1, S_2, S_3 such that

Measure	1	2	3
$\mathbb{E}$	2	1	1
Var	4	3	2

Invest in these stock in proportion $2:1:1$: total return is

$$\begin{aligned} R &= \frac{1}{2}S_1 + \frac{1}{4}S_2 + \frac{1}{4}S_3; \\ \mathbb{E}R &= \frac{1}{2}\cdot 2 + \frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 1 = \frac{3}{2}; \\ \operatorname{Var}(R) &= \frac{1}{4}\cdot 4 + \frac{1}{16}3 + \frac{1}{16}2 = \frac{21}{16}. \end{aligned}$$

Example 4.11. Two stocks with returns X and Y, with $\mathbb{E}X = 2$ and $\operatorname{Var} X = 3$, $\mathbb{E}Y = 1$ and $\operatorname{Var} Y = 1$, and $\rho(X,Y) = 0.5$. Invest in them equally, then total return is $R = (X+Y)/2$. We get:

$$\begin{aligned} \mathbb{E}R &= \frac{1}{2}(2+1) = 1.5, \\ \operatorname{Var} R &= \frac{1}{4}\operatorname{Var}(X+Y) = \frac{1}{4}(\operatorname{Var} X + \operatorname{Var} Y + 2\operatorname{Cov}(X,Y)) = \frac{1}{4}(3+1+2\cdot\sqrt{3}\sqrt{1}\cdot 0.5) = 1 + \frac{\sqrt{3}}{4} \approx 1.4. \end{aligned}$$

Thus diversification decreased variance and with it risk.

4.4. **Marginal distributions.** If we know the joint distribution of (X,Y), then the distribution of only X or only Y is called a *marginal distribution*. If (X,Y) are jointly discrete, then the marginal distribution of X can be found by summing the probabilities over all Y, and vice versa:

$$\mathbf{P}(X=x) = \sum_y \mathbf{P}(X=x,\, Y=y); \quad \mathbf{P}(Y=y) = \sum_x \mathbf{P}(X=x,\, Y=y).$$

If (X,Y) are jointly continuous with density f, we get:

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}y; \quad f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}x.$$

Example 4.12. Consider (X,Y) with joint distributions Then the marginal distributions of X and Y are:

$$\mathbf{P}(X=0) = 0.2,\, \mathbf{P}(X=1) = 0.8; \quad \mathbf{P}(Y=0) = 0.6,\, \mathbf{P}(Y=1) = 0.4.$$

Example 4.13. Consider (X,Y) with joint density $f(x,y) = x+y$, $0 \le x \le 1$, $0 \le y \le 1$. Then density of Y is

$$f_Y(y) = \int_0^1 (x+y)\,\mathrm{d}x = y + \frac{1}{2},\ 0 \le y \le 1.$$

Example 4.14. For $f(x,y) = 2$, $0 \le x \le 1$, $x \le y \le 1$, we have:

$$f_X(x) = \int_x^1 2\,\mathrm{d}y = 2(1-x),\ 0 \le x \le 1; \quad f_Y(y) = \int_0^y 2\,\mathrm{d}x = 2y,\ 0 \le y \le 1.$$

4.5. **Order statistics.** Take i.i.d. random variables $X_1, \dots, X_n$ with CDF $F_X(x) = \mathbb{P}(X \le x)$. Find the cdf of the minimum $m_n = \min(X_1, \dots, X_n)$ and the maximum $M_N = \max(X_1, \dots, X_n)$.

For the maximum, we can say that $M_n \le x$ is equivalent to $X_1 \le x$ and $X_2 \le x$ and etc. and $X_n \le x$. The CDF of M_N is given by

$$F_n(x) = \mathbb{P}(M_N \le x) = \mathbb{P}(X_1 \le x, \dots, X_n \le x) = \mathbb{P}(X_1 \le x) \cdot \mathbb{P}(X_n \le x) = F_X^n(x).$$

By differentiating this CDF, if this is a continuous random variable, we can get density.

Example 4.15. If $X, Y \sim \text{Exp}(\lambda)$ i.i.d. (waiting time), then $F_X(x) = 1 - e^{-\lambda x}$ for $x > 0$, and $\max(X, Y)$ has CDF $(1 - e^{-\lambda x})^2$ for $x > 0$. Thus density is

$$2\lambda e^{-\lambda x}(1 - e^{-\lambda x}),\ x > 0.$$

For the minimum, we cannot say this because $m_n \le x$ is equivalent to $X_1 \le x$ or $X_2 \le x$ etc. Here we have OR instead of AND. Thus we need to invert: $m_n > x$ is equivalent to $X_1 > x$ and $X_2 > x$ and etc. We can compute tail functions instead of CDFs: $1 - F_X(x) = \mathbb{P}(X > x)$.

$$G_n(x) = \mathbb{P}(m_n \le x) = 1 - \mathbb{P}(m_n > x) = 1 - \mathbb{P}(X_1 > x, \dots, X_n > x) = 1 - (1 - F_X(x))^n.$$

Thus $1 - G_n(x) = (1 - F_X(x))^n$.

Example 4.16. If $X, Y \sim \text{Exp}(\lambda)$ i.i.d. (waiting time), then $1 - F_X(x) = e^{-\lambda x}$ for $x > 0$, and $\max(X, Y)$ has tail function $(e^{-\lambda x})^2 = e^{-2\lambda x}$ for $x > 0$. Thus density is minus derivative of tail function:

$$2\lambda e^{-2\lambda x}, \quad x > 0.$$

5. Conditional Distributions

5.1. **Conditional probability, discrete case.** Recall how to find conditional probability from Section 2.

Example 5.1. Toss a fair coin twice. Let A be the event that there are two Heads. Let $Y = 1$ if the first toss is H, 0 otherwise. Then

$$\mathbf{P}(A \mid Y = 1) = \frac{1}{2},$$

because this is the probability that the second toss is H. Next, if the first toss is not H, then we cannot have both H; thus,

$$\mathbf{P}(A \mid Y = 0) = 0.$$

We can write this as

$$\mathbf{P}(A \mid Y) = \begin{cases} \mathbf{P}(A \mid Y = 0),\ Y = 0; \\ \mathbf{P}(A \mid Y = 1),\ Y = 1; \end{cases} = \begin{cases} 0,\ Y = 0; \\ \frac{1}{2},\ Y = 1; \end{cases} = \frac{Y}{2}.$$

The random variable $\mathbf{P}(A \mid Y)$ is called the *conditional probability of A given Y*. We can also have conditional probability depending on many random variables.

Example 5.2. Let X, Y, Z be random variables with the following joint distribution:

X	Y	Z	Prob.
0	0	0	1/8
1	0	0	1/8
0	1	2	1/8
−1	1	1	1/8
0	1	1	1/2

Find $\mathbf{P}(X = 0 \mid Y, Z)$. We have:

$$\mathbf{P}(X = 0 \mid Y = Z = 0) = \frac{1/8}{1/8 + 1/8} = \frac{1}{2}.$$

$$\mathbf{P}(X = 0 \mid Y = 1,\ Z = 2) = 1.$$

$$\mathbf{P}(X = 0 \mid Y = Z = 1) = \frac{1/2}{1/2 + 1/8} = \frac{4}{5}.$$

In other words,

$$\mathbf{P}(X = 0 \mid Y, Z) = \begin{cases} \frac{1}{2},\ Y = Z = 0; \\ 1,\ Y = 1,\ Z = 2; \\ \frac{4}{5},\ Y = Z = 1; \end{cases}$$

5.2. **Conditional expectation, discrete case.** This is defined similarly to the conditional probability.

Example 5.3. In the previous example, let us find $\mathbf{E}(X \mid Y, Z)$. We have:

$$\mathbf{E}(X \mid Y = Z = 0) = \frac{(1/8)\cdot 0 + (1/8)\cdot 1}{(1/8)+(1/8)} = \frac{1}{2},$$
$$\mathbf{E}(X \mid Y = 1,\, Z = 2) = 0,$$
$$\mathbf{E}(X \mid Y = Z = 1) = \frac{(-1)\cdot(1/8) + 0\cdot(1/2)}{(1/8)+(1/2)} = -\frac{1}{5}.$$

Example 5.4. Toss a fair coin twice, and let X be the number of Heads, let $Y = 1$ if the first toss is H, $Y = 0$ otherwise. If $Y = 0$, then $X = 0$ or $X = 1$ with probability $1/2$. If $Y = 1$, then $X = 1$ or $X = 2$ with probabilty $1/2$. Then

$$\mathbf{P}(X \mid Y = 0) = 0\cdot\frac{1}{2} + 1\cdot\frac{1}{2} = \frac{1}{2},$$
$$\mathbf{P}(X \mid Y = 1) = 1\cdot\frac{1}{2} + 2\cdot\frac{1}{2} = \frac{3}{2}.$$

We can represent this as

$$\mathbf{P}(X \mid Y) = \begin{cases} \frac{1}{2},\ Y = 0; \\ \frac{3}{2},\ Y = 1; \end{cases} \quad = \frac{1}{2} + Y.$$

5.3. **Properties of a conditional expectation.** Take random variables $X, Y, Z_1, \ldots, Z_n$.

(a) $\mathbf{E}(X + Y \mid Z_1, \ldots, Z_n) = \mathbf{E}(X \mid Z_1, \ldots, Z_n) + \mathbf{E}(Y \mid Z_1, \ldots, Z_n)$;

(b) If X is a function of $Z_1, \ldots, Z_n$, then $\mathbf{E}(X \mid Z_1, \ldots, Z_n) = X$, because if you know $Z_1, \ldots, Z_n$, you already know the exact value of X.

(c) If X is independent of $Z_1, \ldots, Z_n$, then $\mathbf{E}(X \mid Z_1, \ldots, Z_n) = \mathbf{E}X$, because knowledge of $Z_1, \ldots, Z_n$ does not give us any additional information about the distribution of X.

(d) If X is a function of $Z_1, \ldots, Z_n$, then $\mathbf{E}(XY \mid Z_1, \ldots, Z_n) = X\mathbf{E}(Y \mid Z_1, \ldots, Z_n)$. Indeed, when you try to predict the value of XY based on the information from $Z_1, \ldots, Z_n$, then you already know the value of X, and you can assume X is just a constant, and put it outside of the expectation sign.

(e) $\mathbf{E}(\mathbf{E}(X \mid Z_1, \ldots, Z_n)) = \mathbf{E}X$. The prediction of X given $Z_1, \ldots, Z_n$ is itself a random variable, which is a function of $Z_1, \ldots, Z_n$. If you average over all possible values of $Z_1, \ldots, Z_n$, you give up your knowledge of $Z_1, \ldots, Z_n$, and arrive and the original situation, when you did not know anything. There, the best prediction for X is $\mathbf{E}X$.

Example 5.5. Take i.i.d. $X, Y, Z \sim \mathrm{Poi}(1)$. Then $\mathbf{E}X = \mathrm{Var}\, X = 1 \ \Rightarrow\ \mathbf{E}X^2 = (\mathbf{E}X)^2 + \mathrm{Var}\, X = 2$, same for Y, Z. Therefore,

$$\mathbf{E}(X + 2Y + Z \mid Z) = \mathbf{E}(X \mid Z) + 2\mathbf{E}(Y \mid Z) = \mathbf{E}(Z \mid Z) = \mathbf{E}X + 2\mathbf{E}Y + Z = \boxed{3 + Z}$$

$$\begin{aligned} \mathbf{E}(2X + Z - 1)^2 \mid Z) &= \mathbf{E}(4X^2 + 4XZ - 4X + Z^2 - 2Z + 1 \mid Z) \\ &= 4\mathbf{E}(X^2 \mid Z) + 4\mathbf{E}(XZ \mid Z) - 4\mathbf{E}(X \mid Z) + \mathbf{E}(Z^2 - 2Z + 1 \mid Z) \\ &= 4\mathbf{E}X^2 + 4Z\mathbf{E}X - 4\mathbf{E}X + Z^2 - 2Z + 1 \\ &= 8 + 4Z - 4 + Z^2 - 2Z + 1 = \boxed{Z^2 + 2Z + 5} \end{aligned}$$

5.4. **Conditional distribution.** This is the distribution of a random variable X given $Z_1, \ldots, Z_n$, which consists of all probabilities $\mathbf{P}(X = x \mid Z_1, \ldots, Z_n)$, where x is a possible value of the random variable X.

Example 5.6. Toss two fair coins, and let

$$X = \begin{cases} 1,\ \text{first H}; \\ 0,\ \text{first T}, \end{cases} \qquad Y = \text{number of Heads} = \begin{cases} 2,\ \text{HH}; \\ 1,\ \text{HT, TH}; \\ 0,\ \text{TT}. \end{cases}$$

- if $Y = 2$, then $\mathbf{P}(X = 1 \mid Y) = \mathbf{P}(X = 1 \mid Y) = 1$, $\mathbf{E}(X \mid Y) = 1$, and $\mathrm{Var}(X \mid Y) = 0$, because a constant random variable X has zero variance;
- if $Y = 1$, then $\mathbf{P}(X = 1 \mid Y) = \mathbf{P}(X = 0 \mid Y) = 0.5$, $\mathbf{E}(X \mid Y) = 0.5$, $\mathbf{E}(X^2 \mid Y) = 1^2 \cdot 0.5 + 0^2 \cdot 0.5 = 0.5$, and $\mathrm{Var}(X \mid Y) = \mathbf{E}(X^2 \mid Y) - (\mathbf{E}(X \mid Y))^2 = 0.5 - 0.5^2 = 0.25$;
- if $Y = 0$, then $\mathbf{P}(X = 0 \mid Y) = 1$, $\mathbf{E}(X \mid Y) = 0$, and $\mathrm{Var}(X \mid Y) = 0$.

Example 5.7. Let $S = \{1, 2, 3, 4, 5\}$, and $p(1) = 0.5,\ p(2) = 0.2,\ p(3) = p(4) = p(5) = 0.1$. Let

$$X(\omega) = \begin{cases} 1, & \omega = 1; \\ 0, & \omega = 2, 3, 4, 5; \end{cases} \qquad Y(\omega) = \begin{cases} 1, & \omega = 1, 2, 3; \\ 0, & \omega = 4, 5; \end{cases} \qquad Z(\omega) = \begin{cases} 1, & \omega = 1, 2, 3, 4; \\ 0, & \omega = 5. \end{cases}$$

Let us find the law (the joint distribution) of (X, Y) given Z. If $Z = 1$, then this event has probability

$$\mathbf{P}(Z = 1) = p(1) + p(2) + p(3) + p(4) = 0.9,$$

and therefore we can calculate

$$\mathbf{P}(X = 1, Y = 1 \mid Z = 1) = \frac{p(1)}{\mathbf{P}(Z = 1)} = \frac{0.5}{0.9} = \frac{5}{9},$$

$$\mathbf{P}(X = 0, Y = 1 \mid Z = 1) = \frac{p(2) + p(3)}{\mathbf{P}(Z = 1)} = \frac{0.3}{0.9} = \frac{1}{3},$$

$$\mathbf{P}(X = 0, Y = 0 \mid Z = 1) = \frac{p(4)}{\mathbf{P}(Z = 1)} = \frac{0.1}{0.9} = \frac{1}{9}.$$

Also, if $Z = 0$, then $X = Y = 0$ with probability 1, but

$$\mathbf{P}(X = 0, Y = 0 \mid Z = 0) = 1.$$

5.5. **Continuous distributions.** Assume (X, Y) have joint density $f(x, y)$. The marginal density of X is given by $f_X(x) = \int f(x, y)\mathrm{d}y$, and the marginal density of Y is given by $f_Y(y) = \int f(x, y)\mathrm{d}x$. If we know that $X = x$, then Y has *conditional density*

$$f_{Y|X}(y \mid x) = \frac{f(x, y)}{f_X(x)}.$$

Similarly, if $Y = y$, then X has *conditional density*

$$f_{X|Y}(x \mid y) = \frac{f(x, y)}{f_Y(y)}.$$

We can also calculate the conditional expectation:

$$\mathbf{E}(Y \mid X = x) = \int y f_{Y|X}(y \mid x)\mathrm{d}y.$$

Example 5.8. Take the density $f(x, y) = x + y$ for $0 \le x \le 1$ and $0 \le y \le 1$. Then

$$f_X(x) = \int_0^1 (x + y)\mathrm{d}y = \left(xy + \frac{y^2}{2}\right)\Big|_{y=0}^{y=1} = x + \frac{1}{2},$$

$$f_{Y|X}(y \mid x) = \frac{x + y}{x + 1/2},\ 0 \le y \le 1.$$

For example, if $x = 1/4$, then

$$f_{Y|X}(y \mid x) = \frac{1/4 + y}{3/4} = \frac{1}{3} + \frac{4}{3}y,\ 0 \le y \le 1,$$

$$\mathbf{E}\left(Y \mid X = \frac{1}{4}\right) = \int_0^1 \left(\frac{1}{3} + \frac{4}{3}y\right) y\mathrm{d}y = \left(\frac{1}{6}y^2 + \frac{4}{9}y^3\right)\Big|_{y=0}^{y=1} = \frac{1}{6} + \frac{4}{9} = \frac{11}{18}.$$

6. Moment Generating Functions

Definition 6.1. A moment generating function (MGF) for a random variable X is defined as

$$M_X(t) = \mathbb{E}\left[e^{tX}\right].$$

Example 6.1. For the random variable which assumes values ± 1 with equal probability $1/2$, this MGF is given by

$$M_X(t) = \frac{1}{2} \cdot e^{1t} + \frac{1}{2} \cdot e^{-1t} = \frac{1}{2}\left(e^t + e^{-t}\right).$$

Example 6.2. If X is the number of Heads in two coin tosses, then the law of X is

$$\mathbb{P}(X = 0) = \frac{1}{4}, \quad \mathbb{P}(X = 1) = \frac{1}{2}, \quad \mathbb{P}(X = 0) = \frac{1}{4}.$$

Thus $\mathbb{E}\left[e^{tX}\right] = \frac{1}{4}e^{0t} + \frac{1}{2}e^{1t} + \frac{1}{4}e^{2t} = \frac{1}{4}(1 + 2e^t + e^{2t})$.

Example 6.3. Let X be the number of tosses needed to get one Head. Then

$$\mathbb{P}(X=1)=\mathbb{P}(H)=\frac{1}{2},\quad \mathbb{P}(X=2)=\mathbb{P}(TH)=\frac{1}{4},\dots,\mathbb{P}(X=n)=2^{-n}.$$

Recall: $\sum_{n=1}^{\infty} a^n = a/(1-a)$. Thus the MGF is

$$\mathbb{E}\left[e^{tX}\right]=\sum_{n=1}^{\infty}2^{-n}e^{tn}=\sum_{n=1}^{\infty}(e^t/2)^n=\frac{e^t/2}{1-e^t/2}.$$

Example 6.4. X has density $f(x)=2e^{-2x},\ x>0$. Thus

$$M_X(t)=\mathbb{E}\left[e^{tX}\right]=\int_0^{\infty}f(x)e^{tx}\,\mathrm{d}x=2\int_0^{\infty}e^{-(2-t)x}\,\mathrm{d}x=\frac{2}{2-t}.$$

Recall that $\int_0^{\infty}e^{-ax}\,\mathrm{d}x=\frac{1}{a}$ for $a>0$.

Theorem 6.1. *We have:* $\mathbb{E}[X]=M_X'(0)$*, and for other* $k=1,2,\dots$ *we have:* $\mathbb{E}\left[X^k\right]=M_X^{(k)}(0)$.

Proof. $M_X'(t)=\mathbb{E}\left[(e^{tX})'\right]=\mathbb{E}\left[Xe^{tX}\right]$. Letting $t=0$, we get: $M_X'(0)=\mathbb{E}\left[Xe^0\right]=\mathbb{E}[X]$. Similarly for $k\geq 2$. □

Example 6.5. For the previous example,

$$\mathbb{E}[X]=M_X'(0)=\left(\frac{2}{2-t}\right)'\Big|_{t=0}=\frac{2}{(2-t)^2}\Big|_{t=0}=\frac{1}{2}.$$

This is easier to compute than integrate by parts

$$\mathbb{E}[X]=\int_0^{\infty}2xe^{-2x}\,\mathrm{d}x.$$

Next, the second moment is

$$\mathbb{E}\left[X^2\right]=M_X''(0)=\frac{2\cdot 2}{(2-t)^3}\Big|_{t=0}=\frac{1}{2}.$$

Thus $\operatorname{Var}X=\mathbb{E}\left[X^2\right]-(\mathbb{E}X)^2=1/4$.

Example 6.6. For $M_X(t)=\frac{2}{3}e^{-2t}+\frac{1}{3}e^t$, we have: $\mathbb{P}(X=1)=1/3$ and $\mathbb{P}(X=-2)=2/3$.

Theorem 6.2. *For two independent random variables* X *and* Y*,* $M_{X+Y}(t)=M_X(t)M_Y(t)$.

Proof. We have: $\mathbb{E}\left[e^{t(X+Y)}\right]=\mathbb{E}\left[e^{tX}e^{tY}\right]=\mathbb{E}\left[e^{tX}\right]\mathbb{E}\left[e^{tY}\right]$. □

Example 6.7. For sum of two independent random variables with $f(x)=2e^{-2x}$, $x>0$, we get: $M(t)=(2/(2-t))^2$.

7. Special Distributions

7.1. **Binomial distribution.** Consider a sequence of Bernoulli trials: Each results in success or failure with probabilities p and q, with $p+q=1$. All trials are independent.

Example 7.1. Toss a coin 10 times. Then Heads is success, Tails is failure, and $p=q=1/2$.

Example 7.2. Roll a die. Success is 6, all others are failures. Then $p=1/6$ and $q=5/6$.

Example 7.3. A stock market, measured by Standard & Poor 500, rises or falls each day. Assume daily movements are independent. Then $p=53.29\%$ and $q=46.71\%$ are estimated from daily movements, 1996–2016.

Definition 7.1. A *binomial random variable* is the number of successes in N trials, with probability p of success. The distribution is denoted by $\mathrm{Bin}(N,p)$.

Theorem 7.1. *The distribution of* $X\sim\mathrm{Bin}(N,p)$ *is: Can take values* $0,1,\dots,N$*, with*

$$\mathbb{P}(X=k)=\binom{N}{k}p^kq^{N-k},\quad k=0,\dots,N.$$

Proof. The binomial coefficient $\binom{N}{k}$ is the number of choices which k of N trials will be successful. In other words, this is the quantity of subsets of k elements in $\{1,2,\dots,N\}$. Each such choice has probability p^kq^{N-k}. □

Example 7.4. $N=4$ and $k=2$. Then there are 6 choices: SSFF, SFSF, SFFS, FSSF, FSFS, FFSS. Each choice has probability p^2q^2. Thus $\mathbb{P}(X=2)=6p^2q^2$.

Theorem 7.2. *For* $X\sim\mathrm{Bin}(N,p)$*, we have:* $\mathbb{E}X=Np$ *and* $\operatorname{Var}(X)=Npq$.

Proof. Let $X = X_1 + \ldots + X_N$ with $X_1 = 1$ if the first toss resulted in H, $X_1 = 0$ if the first toss resulted in T. Same for $X_2, X_3, \ldots$ Then $\mathbb{E}X_i = 1 \cdot p + 0 \cdot q = p$ and $\mathbb{E}[X_i^2] = 1^2 \cdot p + 0^2 \cdot q = p$, thus $\operatorname{Var} X_i = p - p^2 = pq$. Therefore,

$$\mathbf{E}X = \mathbf{E}X_1 + \ldots + \mathbf{E}X_N = p + p + \ldots + p = N.$$

Since $X_1, \ldots, X_N$ are independent, we have:

$$\operatorname{Var} X = \operatorname{Var} X_1 + \ldots + \operatorname{Var} X_N = pq + \ldots + pq = Npq.$$

□

Example 7.5. Let $X \sim \text{Bin}(4, 0.3)$. Then $\mathbb{E}X = 4 \cdot 0.3 = 1.2$ and $\operatorname{Var} X = 4 \cdot 0.3 \cdot 0.7 = 0.84$.

Theorem 7.3. *The MGF for $X \sim \text{Bin}(N, p)$ is $M_X(t) = (pe^t + q)^N$.*

Proof. We have: $M_{X_i}(t) = \mathbb{E}\left[e^{tX_i}\right] = pe^{1t} + qe^{0t} = pe^t + q$. Thus by independence

$$M_X(t) = M_{X_1}(t) \ldots M_{X_N}(t) = (pe^t + q)^N.$$

□

Example 7.6. An IT center uses 9 disk drivers. Each disk is out of servic, independently of others, with probability 6%. The instruction stipulates that at least 7 drives must be available. What is the probability of this? The number of working drives is $X \sim \text{Bin}(9, 0.94)$. Thus

$$\mathbb{P}(X \geq 7) = \mathbb{P}(X = 7) + \mathbb{P}(X = 8) + \mathbb{P}(X = 9) = \binom{9}{7} 0.94^7 0.06^2 + \binom{9}{8} 0.94^8 0.06 + \binom{9}{9} 0.94^9 \approx 0.986.$$

Example 7.7. An insurance company has 10 clients, each has an accident independently of others, with probability 20%. Then the number of accidents is $X \sim \text{Bin}(10, 0.2)$, and $\mathbb{P}(X = 2) = \binom{10}{2} 0.2^2 0.8^8 \approx 0.302$.

7.2. **Geometric distribution.**

Definition 7.2. For a sequence of independent Bernoulli trials with probabilities p and q of success and failure, let X be the number of trials until the first success. Then $X \sim \text{Geo}(p)$.

Theorem 7.4. *For such X, $\mathbb{P}(X = n) = pq^{n-1}$, $n = 1, 2, \ldots$*

Proof. $X = n$ means first $n - 1$ failures and 1 success: FF...FS, which has probability $q \cdot q \cdot \ldots \cdot q \cdot p = pq^{n-1}$. □

Example 7.8. The probability that your first Heads will be n is equal to $(1/2)^n$, since $p = q = 1/2$.

Theorem 7.5. *The Moment Generating Function is*

$$M_X(t) = \frac{pe^t}{1 - qe^t}.$$

Proof. We have: $M_X(t) = \sum_{n \geq 1} pq^{n-1} e^{tn} = pe^t \sum_{n \geq 1} (qe^t)^{n-1}$, and now recall that $\sum_{n=0}^{\infty} a^n = 1/(1 - a)$. □

Theorem 7.6. *For $X \sim \text{Geo}(p)$, we have $\mathbb{E}[X] = 1/p$ and $\operatorname{Var} X = q/p^2$.*

Theorem 7.7. *Take $M_X'(0) = \mathbb{E}[X] = 1/p$ and $M_X''(0) = \mathbb{E}[X^2] = (1 + q)/p^2$. Thus $\operatorname{Var} X = q/p^2$.*

Example 7.9. For X the number of tosses until first Heads, then $X \sim \text{Geo}(1/2)$. Then

$$M_X(t) = \frac{e^t/2}{1 - e^t/2}, \quad \mathbb{E}X = 2, \quad \operatorname{Var} X = 2.$$

7.3. **Negative binomial distribution.** How many times X do you need to toss a coin to get r Heads, where r is a fixed parameter? This is called *negative binomial distribution* $\text{NB}(r, p)$. For example, if $r = 3$ and the sequence is TTTHTHH, then $X = 7$. What is $\mathbf{P}(X = n)$? If the nth toss resulted in rth Heads, another way to say this is the first $n - 1$ tosses contain $r - 1$ Heads and $n - r$ Tails, and the last, nth toss resulted in Heads. The probability that the first $n - 1$ tosses resulted in $r - 1$ Heads (and $n - r$ Tails) is

$$\binom{n-1}{r-1} p^{r-1} q^{n-r}.$$

Indeed, there are

$$\binom{n-1}{r-1}$$

choices for slots in the first $n-1$ tosses occupied by Heads. Each of these choices has probability $p^{r-1}q^{n-r}$. Finally, the probability of the last toss being Heads is p. So

$$\boxed{\mathbf{P}(N=n)=\binom{n-1}{r-1}p^r q^{n-r}}$$

This is true for $n=r, r+1, r+2, \ldots$. The random variable $X \backsim \mathrm{NB}(r,p)$ can take only values greater than or equal to r, because to get r Heads, you need at least r tosses. One can alternatively describe this distribution as the number of Bernoulli trials one needs to get to the rth success. For $r=1$, this becomes the Geometric distribution with parameter p. [1]

Example 7.10. Let $p=1/2$, $r=2$. Then

$$\mathbf{P}(X=3)=\binom{3-1}{2-1}\left(\frac{1}{2}\right)^2\left(1-\frac{1}{2}\right)=2\cdot\frac{1}{4}\cdot\frac{1}{2}=\frac{1}{4}.$$

We can equivalently get this as follows: to get two Heads in exactly three tosses, we need either THH or HTH. Each of these sequences has probability $1/8$. So the resulting probability is $(1/8)+(1/8)=1/4$.

For $X \backsim \mathrm{NB}(r,p)$, we have:

$$X=X_1+\ldots+X_r, \quad X_1,\ldots,X_r \backsim \mathrm{Geo}(p) \text{ i.i.d.}$$

Recall that i.i.d. means independent identically distributed. Indeed, to get to the rth Heads, we need to get the first Heads, which required geometric number X_1 of tosses; then we need to get from first to second Heads, which required also geometric number X_2 of tosses. This second number is independent of the first one, because the coin does not have a memory, etc. Therefore,

$$\mathbf{E}X=\mathbf{E}X_1+\ldots+\mathbf{E}X_r=\frac{r}{p},$$

and using independence, we get:

$$\operatorname{Var}X=\operatorname{Var}X_1+\ldots+\operatorname{Var}X_r=\frac{qr}{p^2}.$$

The negative binomial distribution and the binomial distribution answer somewhat opposite questions. For the binomial distribution, you have fixed number of tries, and you ask how many successful trials you have. For the negative binomial distribution, you have fixed number of successful trials, and you ask how many trials you need.

7.4. **Poisson distribution.** Take a parameter $\lambda>0$.

Definition 7.3. A *Poisson distribution* $X \sim \mathrm{Poi}(\lambda)$ is given by

$$\mathbf{P}(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}, \quad k=0,1,2,\ldots$$

Example 7.11. For $X \sim \mathrm{Poi}(2)$, we have: $\mathbf{P}(X\geq 1)=1-\mathbf{P}(X=0)=1-e^{-2}$; $\mathbf{P}(X\geq 2)=1-\mathbf{P}(X=0)-\mathbf{P}(X=1)-1-e^{-2}-2e^{-2}=1-3e^{-2}$.

Theorem 7.8. *Consider a binomial distribution* $\mathrm{Bin}(N,p)$. *If* $N\to\infty$, *and* $Np=\lambda$, *then* $\mathrm{Bin}(N,p)\to\mathrm{Poi}(\lambda)$. *More precisely, for every fixed* $k=0,1,2,\ldots$, *we have as* $N\to\infty$:

$$\binom{N}{k}p^k q^{N-k}\to\frac{\lambda^k}{k!}e^{-\lambda}.$$

Proof. Note that $p=\lambda/N$, $q=1-\lambda/N$, and

$$\binom{N}{k}=\frac{N!}{k!(N-k)!}=\frac{N(N-1)\ldots(N-k+1)}{N^k}.$$

Therefore,

$$\begin{aligned}\binom{N}{k}p^k q^{N-k}&=\frac{N(N-1)\ldots(N-k+1)}{N^k}\left(\frac{\lambda}{N}\right)^k\left(1-\frac{\lambda}{N}\right)^{N-k}\\&=\frac{\lambda^k}{k!}\cdot\frac{N(N-1)\ldots(N-k+1)}{N^k}\cdot\left(1-\frac{\lambda}{N}\right)^{N-k}.\end{aligned}$$

[1] Similarly to the geometric distribution, there is ambiguity in the definition of the negative binomial distribution. Sometimes it is defined as the quantity of T, as opposed to tosses, until your first H. In this case, it is shifted by r down. Thus defined negative binomial random variable takes values $0,1,2,\ldots$ We define this distribution as the number of tosses until your rth H.

The first factor $\lambda^k/k!$ is constant (does not depend on N); the second factor is

$$\frac{N(N-1)\dots(N-k+1)}{N^k} = \frac{N}{N}\cdot\frac{N-1}{N}\cdot\ldots\cdot\frac{N-k+1}{N} \to 1,$$

because each of these fractions tends to one: indeed, $N\to\infty$, and k is fixed. Finally, the third factor tends to $e^{-\lambda}$, because

$$\log\left(1-\frac{\lambda}{N}\right)^{N-k} = (N-k)\log\left(1-\frac{\lambda}{N}\right) \backsim (N-k)\left(-\frac{\lambda}{N}\right) = -\lambda\frac{N-k}{N}\to -\lambda,$$

and then we exponentiate both sides. We use the tangent line approximation:

$$\log(1+x)\backsim x,\ x\to 0,$$

which is true: let $f(x)=\log(1+x)$; then $f'(x)=1/(1+x)$, and so $f(0)=0$, $f'(0)=1$, and

$$f(x)\backsim f(0)+f'(0)x = x,\quad x\to 0.$$

□

This is the meaning of a Poisson distribution: this is approximately the quantity of many events, each of which is very rare.

Example 7.12. A company which sells flood insurance has $N=20000$ clients, and each client has probability of flood $p=0.01\%$, independently of others. What is the distribution of the number X of floods? Approximately it is $X\backsim \mathrm{Bin}(N,p)$, but it is hard for calculations: for example,

$$\mathbf{P}(X=10) = \binom{20000}{10}\left(\frac{1}{10000}\right)^{10}\left(\frac{9999}{10000}\right)^{19990}.$$

Use Poisson approximation: X is approximately distributed as $\mathrm{Poi}(2)$, because $\lambda = Np = 2$. So

$$\mathbf{P}(X=10) = \frac{2^{10}}{10!}e^{-2}.$$

Also,

$$\mathbf{P}(X\ge 3) = 1-\mathbf{P}(X=0)-\mathbf{P}(X=1)-\mathbf{P}(X=2)$$
$$= 1-e^{-2}-\frac{2^1}{1!}e^{-2}-\frac{2^2}{2!}e^{-2} = 1-e^{-2}-2e^{-2}-2e^{-2} = 1-5e^{-2}.$$

A slightly more general statement is true. If we have N (a large number of) independent events, occuring with probabilities $p_1,\dots,p_N$, then the quantity of events which happened is approximately $\mathrm{Poi}(\lambda)$ with $\lambda = p_1+\ldots+p_N$. So the Poisson distribution is the *law of rare events*: if we have many events, each of which is very unlikely to occur, then the number of events is Poisson.

Example 7.13. In one city, independently of each other, $N_1=1000$ houses have fire with probability $p_1=.2\%$ and $N_2=500$ houses have fire with probability $p_2=.6\%$. Then the number of fires is approximately $\mathrm{Poi}(\lambda)$ with $\lambda = N_1p_1+N_2p_2 = 5$.

Another application is a sequence of events, with next event happens with probability $\lambda\triangle t$ on $[t,t+\triangle t]$, independently of other time intervals. Then the number of events during $[0,T]$ is $\sim \mathrm{Poi}(\lambda T)$. Indeed, split $[0,T]$ into N equal subintervals of length T/N, and apply the above law of rare events to $p=\lambda(T/N)$. Then $Np=\lambda T$.

Example 7.14. The rate for α-particle emission is $\lambda\approx 3.87$ (with unit of time 1/8 of minute). Then during 2 minutes we have $\mathrm{Poi}(3.87\cdot 16)$ particles.

7.5. **Normal distribution.** We define first the standard normal distribution and then the general normal distribution.

Definition 7.4. A *standard normal* random variable has density $f_Z(z) = (2\pi)^{-1/2}\exp(-z^2/2)$. For a standard normal Z, we define normal $X\sim\mathcal{N}(\mu,\sigma^2)$ as $X=\mu+\sigma Z$.

Theorem 7.9. *The MGF for $Z\sim\mathcal{N}(0,1)$ is given by*

$$M_Z(t) = \int_{-\infty}^{+\infty} e^{tz}f_Z(z)\,\mathrm{d}z = e^{t^2/2}.$$

The MGF for $X\sim\mathcal{N}(\mu,\sigma^2)$ is given by

Proof. First, we compute MGF for standard normal law:

$$F_Z(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tz}e^{-z^2/2}\mathrm{d}z = \frac{1}{\sqrt{2\pi}}e^{t^2/2}\int_{-\infty}^{\infty} e^{-t^2/2+tz-z^2/2}\mathrm{d}z$$
$$= \frac{e^{t^2/2}}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(z-t)^2/2}\mathrm{d}z = \frac{e^{t^2/2}}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-u^2}\mathrm{d}u = \frac{e^{t^2/2}}{\sqrt{2\pi}}\sqrt{2\pi} = e^{t^2/2}.$$

(we changed variables $u = z - t$). Next, we can represent $X = \mu + \sigma Z$, and

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = \mathbb{E}\left[e^{t\mu+t\sigma Z}\right] = e^{t\mu}\cdot\mathbb{E}\left[e^{t\sigma Z}\right] = e^{t\mu}e^{(t\sigma)^2/2} = \exp(t\mu + t^2\sigma^2/2).$$

□

Example 7.15. For $X \sim \mathcal{N}(-2, 2)$ we get: $M_X(t) = e^{-2t+t^2}$.

Theorem 7.10. *For $X \sim \mathcal{N}(\mu, \sigma^2)$, we have: $\mathbb{E}X = \mu$ and $\operatorname{Var} X = \sigma^2$.*

Proof. We compute for standard normal Z: $\mathbb{E}Z = M_Z'(0) = 0$, $\mathbb{E}Z^2 = M_Z''(0) = 1$, and thus $\operatorname{Var} Z = 1$. Therefore $\mathbb{E}[X] = \mu + \sigma\mathbb{E}[Z] = \mu$, and $\operatorname{Var}(X) = \sigma^2\operatorname{Var}(Z) = \sigma^2$. □

Theorem 7.11. *The density of $\mathcal{N}(\mu, \sigma^2)$ is given by*

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right).$$

Proof. Indeed, for all a and b,

$$\mathbf{P}(a \le Y \le b) = \mathbf{P}(a \le \mu + \sigma X \le b) = \mathbf{P}\left(\frac{a-\mu}{\sigma} \le X \le \frac{b-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}}\int\limits_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} e^{-x^2/2}\,dx.$$

Now let us change variables:

$$x = \frac{y-\mu}{\sigma};\quad dx = \frac{dy}{\sigma};\quad a \le y \le b.$$

Therefore,

$$\frac{1}{\sqrt{2\pi}}\int\limits_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} e^{-x^2/2}\,dx = \frac{1}{\sqrt{2\pi}\sigma}\int\limits_{a}^{b} e^{-(y-\mu)^2/(2\sigma^2)}\,dy = \int_a^b p(y)\,dy.$$

This proves that Y has density $f(y)$, as shown above. □

Example 7.16. For $X \sim \mathcal{N}(1, 4)$, find $\mathbf{E}\left[(X-2)^3\right]$: $X = 1 + 2Z$ with $Z \sim \mathcal{N}(0, 1)$, and $(X-2)^3 = (2Z-1)^3 = 8Z^3 - 12Z^2 + 6Z - 1$. Taking the expected value and using $\mathbf{E}Z = \mathbf{E}Z^3 = 0$ (by symmetry of the standard normal distribution), $\mathbf{E}Z^2 = 1$, we get: $\mathbf{E}\left[(X-2)^3\right] = -12 - 1 = \boxed{-13}$

Example 7.17. For $X \sim \mathcal{N}(1, 4)$, find $\mathbf{P}(X \le 3)$: $X = 1 + 2Z$ with $Z \sim \mathcal{N}(0, 1)$, and $\mathbf{P}(X \le 3) = \mathbf{P}(1 + 2Z \le 3) = \mathbf{P}(Z \le 1) = 0.8413$

Example 7.18. Find a normal distribution with mean 1 and 99% quantile 4.5. We can represent this random variable as $X = 1 + \sigma Z$, where σ needs to be found. Because

$$0.99 = \mathbf{P}(X \le 4.5) = \mathbf{P}(1 + \sigma Z \le 4.5) = \mathbf{P}(Z \le \sigma^{-1}3.5),$$

from the Z-table at the end of these lecture notes we get: $\sigma^{-1}3.5 = x_{99\%} = 2.326$, so $\sigma = 3.5/2.326 = 1.505$, and $X \sim \boxed{\mathcal{N}(1, 1.505^2)}$

7.6. **Exponential distribution.** Sometimes it is used to model the lifespan of a device, when the remaining lifespan is independent of how long it has worked so far. Take a $\lambda > 0$.

Definition 7.5. A random variable X with density $f(x) = \lambda e^{-\lambda x}$, $x > 0$, is called *exponential* with *rate* λ.

This distribution has the following properties: for $a > 0$,

$$\mathbf{P}(X \ge a) = \int_a^\infty \lambda e^{-\lambda x}\,dx = \left(-e^{-\lambda x}\right)\Big|_{x=a}^{x=\infty} = e^{-\lambda a},$$

and thus the CDF is given by

$$F_X(u) = \begin{cases} 1 - e^{-\lambda u},\ u > 0; \\ 0,\ u \le 0. \end{cases}$$

Lemma 7.12. *This distribution has memoryless property: for $s, t > 0$, $\mathbf{P}(X > t + s \mid X > s) = \mathbf{P}(X > t)$.*

Proof. Indeed, for any $t, s > 0$, we have:

$$\mathbf{P}(X > t + s \mid X > s) = \frac{\mathbf{P}(X > t + s, X > s)}{\mathbf{P}(X > s)} = \frac{\mathbf{P}(X > t + s)}{\mathbf{P}(X > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = \mathbf{P}(X > t).$$

□

Let $X \sim \mathrm{Exp}(\lambda)$. Then we have: $f_X(x) = \lambda e^{-\lambda x}$ for $x > 0$. Thus for $t < \lambda$,

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = \int_0^\infty e^{tx} f_X(x)\,\mathrm{d}x = \lambda \int_0^\infty e^{-(\lambda - t)x}\,\mathrm{d}x = \frac{\lambda}{\lambda - t}.$$

Thus we can find

$$\mathbb{E}[X] = M_X'(0) = \left(\frac{\lambda}{\lambda - t}\right)'\Big|_{t=0} = \frac{\lambda}{(\lambda - t)^2}\Big|_{t=0} = \frac{1}{\lambda},$$

and similarly $\mathbb{E}\left[X^2\right] = M_X''(0) = 2/\lambda^2$, $\mathrm{Var}\, X = \lambda^{-2}$.

Example 7.19. Find an exponential distribution such that $\mathbf{P}(X \geq 2) = 0.4$. We have: $e^{-2\lambda} = 0.4 \Leftrightarrow \lambda = -\ln(0.4)/2 = 0.46$.

Example 7.20. Calculate the CDF of $X \sim \mathrm{Exp}(2)$. For $x \leq 0$, we have: $F(x) = 0$, because $X \geq 0$ always. For $x > 0$, we have: $F(x) = \int_0^x 2e^{-2x}\,\mathrm{d}x = (-e^{-2x})\big|_{x=0}^{x=\infty} = 1 - e^{-2x}$. Therefore,

$$F(x) = \begin{cases} 0, \, x \leq 0; \\ 1 - e^{-2x}, \, x \geq 0. \end{cases}$$

7.7. **Gamma distribution.** Take n i.i.d. $X_1, \ldots, X_n \sim \mathrm{Exp}(\lambda)$. Then $X = X_1 + \ldots + X_n$ has *Gamma distribution* and is denoted as $\Gamma(n, \lambda)$.

Theorem 7.13. *Its MGF is*

$$M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^n.$$

Its mean and variance are

$$\mathbb{E}[X] = \frac{n}{\lambda}, \quad \mathrm{Var}(X) = \frac{n}{\lambda^2}.$$

Proof. MGF is computed as $M_X(t) = M_{X_1}(t) \cdot M_{X_n}(t)$. Mean and variance are computed as sums of means and variances of $X_1, \ldots, X_n$. □

Theorem 7.14. *The density of $\Gamma(n, \lambda)$ is equal to*

$$f(x) = \frac{\lambda^n}{(n-1)!} x^{n-1} e^{-\lambda x}, \quad x > 0.$$

Proof. Note that

$$n! = I_n := \int_0^\infty x^n e^{-x} \mathrm{d}x.$$

Why? Integrate I_n by parts:

$$I_n = -\int_0^\infty x^n \mathrm{d}e^{-x} = -\,x^n e^{-x}\big|_{x=0}^{x=\infty} + \int_0^\infty e^{-x} \mathrm{d}\left(x^n\right) = 0 + n\int_0^\infty x^{n-1} e^{-x} \mathrm{d}x = nI_{n-1}.$$

Combining that with $I_0 = \int_0^\infty e^{-x} \mathrm{d}x = 1$, we have: $I_n = n!$. Doing change of variables $\lambda x = y$:

$$\int_0^\infty x^{n-1} e^{-\lambda x}\,\mathrm{d}x = \frac{(n-1)!}{\lambda^n}.$$

Applying this to $\lambda - t$ instead of λ, we get:

$$\int_0^\infty f(x) e^{tx}\,\mathrm{d}x = \frac{\lambda^n}{(n-1)!} \int_0^\infty x^{n-1} e^{tx} e^{-\lambda x}\,\mathrm{d}x$$
$$= \frac{\lambda^n}{(n-1)!} \frac{(n-1)!}{(\lambda - t)^n} = \left(\frac{\lambda}{\lambda - t}\right)^n.$$

□

8. Convergence Modes

We compare three modes: almost surely, in probability, and in law (distribution).

Definition 8.1. Take random variables $Z_0, Z_1, Z_2, \dots$ Then $Z_n \to Z$ in probability if for all $\varepsilon > 0$ we have: $\mathbb{P}(|Z_n - Z_0| > \varepsilon) \to 0$ as $n \to \infty$.

But there is another mode of convergence:

Definition 8.2. $Z_n \to Z_0$ *in law*, or *in distribution*, if $F_{Z_n}(x) = \mathbb{P}(Z_n \le x) \to F_{Z_0}(x) = \mathbb{P}(Z_0 \le x)$ for all x where F_{Z_0} is continuous.

In particular, if Z_0 is continuous, then $F_{Z_n}(x) \to F_{Z_0}(x)$ as $n \to \infty$ for all x. Thus for all $a < b$,

$$\mathbb{P}(a \le Z_n \le b) = F_{Z_n}(b) - F_{Z_n}(a) \to \mathbb{P}(a \le Z_0 \le b) = F_{Z_0}(b) - F_{Z_0}(a).$$

Example 8.1. $Z_n = 1/n$ and $Z_0 = 0$ (constant random variables). Then $Z_n \to Z_0$ in law, because

$$F_{Z_n}(x) = \begin{cases} 1, & x \ge 1/n; \\ 0, & x < 1/n; \end{cases} \qquad F_{Z_0}(x) = \begin{cases} 1, & x \ge 0; \\ 0, & x < 0. \end{cases}$$

Thus $F_{Z_n}(x) \to F_{Z_0}(x)$ for $x \ne 0$. Indeed, for $x > 0$, we have $x \ge 1/n$ for sufficiently large n, and $F_{Z_n}(x) = 1 = F_{Z_0}(x)$. And for $x < 0$, $F_{Z_n}(x) = 0 = F_{Z_0}(x)$. But $F_{Z_n}(0) = 0 \nrightarrow 1 = F_{Z_0}(0)$. This is why we need to exclude discontinuity.

Theorem 8.1. *Convergence in probability implies convergence in distribution, but not vice versa.*

Example 8.2. Toss a fair coin. Let $Z_n = 1$ if Heads, $Z_n = 0$ if Tails, for odd n; vice versa when n is even. Then all Z_n have the same distribution: $\mathbb{P}(Z_n = 0) = \mathbb{P}(Z_n = 1) = 1/2$. Thus $Z_n \to Z_0$ in law. But not in probability: $|Z_n - Z_0| = 1$ for n odd.

The third convergence mode is almost surely: $\mathbb{P}(Z_n \to Z_0) = 1$. This mode is stronger than in probability.

9. Probabilistic Inequalities

9.1. **Markov's inequality.** Take a nonnegative random variable $X \ge 0$.

Theorem 9.1. *For every $a > 0$, we have:*

$$\mathbf{P}(X \ge a) \le \frac{\mathbf{E}X}{a}.$$

Proof. Consider another random variable:

$$Y = \begin{cases} a, & X \ge a; \\ 0, & 0 \le X < a. \end{cases}$$

Then always $Y \le X$, so $\mathbf{E}Y \le \mathbf{E}X$. But $\mathbf{E}Y = a\mathbf{P}(X \ge a) + 0\mathbf{P}(0 \le X < a) = a\mathbf{P}(X \ge a)$. This gives us

$$a\mathbf{P}(X \ge a) \le \mathbf{E}X \quad \Rightarrow \quad \mathbf{P}(X \ge a) \le \frac{\mathbf{E}X}{a}.$$

□

Example 9.1. Consider the exponential random variable $X \backsim \mathrm{Exp}(1)$. Then $\mathbf{P}(X \ge 10) \le \frac{\mathbf{E}X}{10} = \frac{1}{10}$, because $\mathbf{E}X = 1$. In reality, $\mathbf{P}(X \ge 10) = e^{-10}$, which is much smaller than $1/10$. Markov's inequality gives us very rough and crude estimates, which are one-size-fits-all solutions. But they can be applied to any variable.

9.2. **Chebyshev's inequality.** This measures large deviations of X from its mean.

Theorem 9.2. *If $\mathbf{E}X = \mu$, then for every $b > 0$ we have:*

$$\mathbf{P}(|X - \mu| \ge b) \le \frac{\operatorname{Var} X}{b^2}.$$

Proof. Apply Markov's inequality to $(X - \mu)^2$ instead of X and b^2 instead of a. We get:

$$\mathbf{P}(|X - \mu| \ge b) = \mathbf{P}((X - \mu)^2 \ge b^2) \le \frac{\mathbf{E}(X - \mu)^2}{b^2} = \frac{\operatorname{Var} X}{b^2}.$$

□

Example 9.2. For $X \sim N(4, 1)$, $\mathbb{E}X = 4$, $\operatorname{Var} X = 1$, and thus $\mathbf{P}(|X - 4| > 2) \le 1/4$. This is equivalent to $X < 2$ or $X > 6$.

9.3. **Law of large numbers.** Let $X_1, X_2, \ldots$ be iid (independent identically distributed) random variables with

$$\mathbf{E}X_1 = \mathbf{E}X_2 = \ldots = \mu, \quad \operatorname{Var} X_1 = \operatorname{Var} X_2 = \ldots = \sigma^2.$$

Theorem 9.3. *Let $S_N = X_1 + \ldots + X_N$. Then we have convergence in probability:*

$$\frac{S_N}{N} \to \mu, \ N \to \infty,$$

in the sense that for every $\varepsilon > 0$ we have:

$$\mathbf{P}\left(\left|\frac{S_N}{N} - \mu\right| \geq \varepsilon\right) \to 0.$$

Proof. Indeed, apply Chebyshev's inequality to S_N/N. We have:

$$\mathbf{E}\frac{S_N}{N} = \frac{1}{N}(\mathbf{E}X_1 + \ldots + \mathbf{E}X_N) = \frac{1}{N}(\mu + \ldots + \mu) = \frac{\mu N}{N} = \mu;$$
$$\operatorname{Var}\frac{S_N}{N} = \frac{1}{N^2}\operatorname{Var} S_N = \frac{1}{N^2}\left(\operatorname{Var} X_1 + \ldots + \operatorname{Var} X_N\right) = \frac{1}{N^2}\left(\sigma^2 + \ldots + \sigma^2\right) = \frac{\sigma^2}{N}.$$

Therefore, as $N \to \infty$, we have:

$$\mathbf{P}\left(\left|\frac{S_N}{N} - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2/N}{\varepsilon^2} \to 0.$$

□

Actually it is possible to prove stronger convergence almost surely, but this is harder.

Example 9.3. If $X \sim \text{Exp}(2)$, then $\mu = 1/2$ and $\sigma^2 = 1/4$, and

$$\mathbf{P}(|\overline{X}_n - \frac{1}{2}| \geq 0.1) \leq \frac{1/4}{0.1^2 \cdot n} = \frac{25}{n}.$$

For $n = 10$ this is still greater than 1, and this estimate is useless. But for $n = 1000$ this is useful: this probability is 2.5%.

9.4. **Large deviations.** Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbf{E}X_1 = \mu$. Then by the Law of Large Numbers for $S_N = X_1 + \ldots + X_N$ we have: $S_N/N \to \mu$. Large Deviations refer to estimating the following probabilities for $\varepsilon > 0$:

$$\mathbf{P}\left(\frac{S_N}{N} - \mu \geq \varepsilon\right) \text{ and } \mathbf{P}\left(\frac{S_N}{N} - \mu \leq -\varepsilon\right).$$

This can be done by combining Markov's inequality and moment generating functions. Instead of formulating general theory, we do a few examples.

Example 9.4. Take $X \sim \mathcal{N}(2, 1)$. Then $\mathbf{E}X = 2$. Let us estimate $\mathbf{P}(X \geq 3)$: For $t > 0$, a parameter to be determined later. Then $\mathbf{P}(X \geq 3) = \mathbf{P}\left(e^{tX} \geq e^{3t}\right)$. By Markov's inequality,

$$\mathbf{P}\left(e^{tX} \geq e^{3t}\right) \leq \frac{\mathbf{E}e^{tX}}{e^{3t}}.$$

But the moment generating function for $X \sim \mathcal{N}(2, 1)$ is given by $\mathbf{E}e^{tX} = e^{2t + t^2/2}$. Choose t to minimize

$$\frac{e^{2t+t^2/2}}{e^{3t}} = \exp\left(\frac{1}{2}t^2 - t\right).$$

To minimize $t^2/2 - t$, take the derivative with respect to t: $(t^2/2 - t)' = t - 1 = 0 \ \Rightarrow \ t = 1$. This gives us

$$\exp\left(\frac{1}{2}t^2 - t\right) = e^{-0.5}.$$

Therefore, $\mathbf{P}(X \geq 3) \leq \boxed{e^{-0.5}}$ Of course, in this simple case we might as well directly calculate this probability. But is more complicated settings, this direct calculation is impossible.

Example 9.5. Take $X \sim \mathcal{N}(2, 1)$, and estimate $\mathbf{P}(X \leq 1)$. For $t > 0$, we have:

$$\mathbf{P}(X \leq 1) = \mathbf{P}\left(e^{-tX} \geq e^{-t}\right) = \frac{\mathbf{E}e^{-tX}}{e^{-t}} = \frac{e^{-2t+t^2/2}}{e^{-t}} = \exp\left(\frac{t^2}{2} - t\right),$$

and this problem is solved simlarly: $\mathbf{P}(X \leq 1) \leq \boxed{e^{-0.5}}$

This method works particularly well for *sums of independent random variables.*

Example 9.6. Let $X_1, \dots, X_{10} \sim \text{Exp}(3)$ be i.i.d. random variables, and let $S := X_1 + \dots + X_{10}$. Then $\mathbf{E}S = \frac{10}{3}$. Let us estmate $\mathbf{P}(S \le 2)$. For $t > 0$,

$$\mathbf{P}(S \le 2) = \mathbf{P}\left(e^{-tS} \le e^{-2t}\right) \le \frac{\mathbf{E}e^{-tS}}{e^{-2t}}.$$

Since $X_1, \dots, X_{10}$ are independent,

$$\begin{aligned}\mathbf{E}e^{-tS} &= \mathbf{E}\left[e^{-tX_1 - \dots - tX_{10}}\right] = \mathbf{E}\left[e^{-tX_1} \dots e^{-tX_{10}}\right] = \mathbf{E}\left[e^{-tX_1}\right] \cdot \dots \cdot \mathbf{E}\left[e^{-tX_{10}}\right] \\ &= \left(\mathbf{E}e^{-tX_1}\right)^{10} = \left(\frac{3}{3+t}\right)^{10}.\end{aligned}$$

We used expression for moment generating function of the exponential random variable: $\mathbf{E}e^{-tX_1} = 3/(3+t)$. Therefore, we need to minimize

$$F(t) = \frac{\left(\frac{3}{3+t}\right)^{10}}{e^{-2t}} = e^{2t}\left(\frac{3}{3+t}\right)^{10}$$

Take the logarithm:

$$\ln F(t) = 2t - 10\ln(3+t) \;\Rightarrow\; (\ln F(t))' = 2 - \frac{10}{3+t} = 0 \;\Rightarrow\; t = 2.$$

Therefore, the minimal value is $\mathbf{P}(S \le 2) \le F(2) = e^{2\cdot 2}(3/5)^{10} = \boxed{0.33}$

9.5. **Chernov's inequality.** This is a special case of large deviations above. Let $X_1, \dots, X_N$ be independent Bernoulli random variables with $\mathbf{P}(X_i = 1) = p_i$, $\mathbf{P}(X_i = 0) = 1 - p_i$. Then $S = X_1 + \dots + X_N$ has mean $\mu = \mathbf{E}S = \mathbf{E}X_1 + \dots + \mathbf{E}X_N = p_1 + \dots + p_N$. Fix a $\delta \in (0,1)$ and estimate

$$\mathbf{P}(S \ge \mu + \delta\mu).$$

Take a parameter $t > 0$:

$$\mathbf{P}(S \ge \mu + \delta\mu) \le \mathbf{P}\left(e^{tS} \ge e^{t\mu(1+\delta)}\right) \le \frac{\mathbf{E}e^{tS}}{e^{t\mu(1+\delta)}}. \tag{1}$$

Since $X_1, \dots, X_N$ are independent,

$$\mathbf{E}e^{tS} = \mathbf{E}\left[e^{tX_1 + \dots + tX_N}\right] = \mathbf{E}e^{tX_1} \cdot \dots \cdot \mathbf{E}e^{tX_N}. \tag{2}$$

For each i, we have: $\mathbf{E}e^{tX_i} = p_i e^{t\cdot 1} + (1-p_i)e^{t\cdot 0} = p_i e^t + (1-p_i) = 1 + p_i(e^t - 1)$. Now comes the main step. Recall that

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} \ge 1 + x \text{ for } x \ge 0. \tag{3}$$

Apply inequality (3) to $x = p_i(e^t - 1)$. Then

$$\mathbf{E}e^{tX_i} \le \exp\left(p_i(e^t - 1)\right). \tag{4}$$

Substituting (4) into (2), we get:

$$\mathbf{E}e^{tS} \le \exp\left(p_1(e^t-1)\right) \cdot \dots \cdot \exp\left(p_N(e^t-1)\right) = \exp\left((p_1 + \dots + p_N)(e^t - 1)\right) = \exp(\mu(e^t - 1)). \tag{5}$$

The right-hand side of (5) depends only on μ, regardless of individual p_i. Plugging (5) into (1), we get:

$$\mathbf{P}(S \ge \mu + \delta\mu) \le \exp\left(\mu(e^t - 1 - t(1+\delta))\right) = \exp(\mu F(t)),\; F(t) := e^t - 1 - (1+\delta)t.$$

Minimize $F(t)$ with respect to t: $F'(t) = e^t - (1+\delta) = 0 \;\Rightarrow\; t = \ln(1+\delta)$. The minimal value is $F(\delta) = e^{\ln(1+\delta)} - 1 - (1+\delta)\ln(1+\delta) = \delta - (1+\delta)\ln(1+\delta)$. It is possible to show that this is less than $-\delta^2/3$ for $\delta \in (0,1)$. Thus,

$$\boxed{\mathbf{P}(S \ge \mu + \delta\mu) \le \exp\left(-\frac{\delta^2\mu}{3}\right)}$$

Similarly, it is possible to show that

$$\boxed{\mathbf{P}(S \le \mu - \delta\mu) \le \exp\left(-\frac{\delta^2\mu}{2}\right)}$$

Example 9.7. $S \sim \text{Bin}(20, 0.4)$ can be represented as $S = X_1 + \ldots + X_{20}$ with i.i.d. Bernoulli X_i with $p_i = 0.4$. Then $\mu = p_1 + \ldots + p_{20} = 0.4 \cdot 20 = 8$. Let us estimate $\mathbf{P}(S \geq 12)$: this corresponds to $8(1+\delta) = 12 \Rightarrow \delta = 0.5$, and $\mathbf{P}(S \geq 12) \leq \exp\left(-\frac{0.5^2 \cdot 8}{2}\right) = \boxed{e^{-1}}$

Example 9.8. Take $S_1 \sim \text{Bin}(10, 0.4)$ and $S_2 \sim \text{Bin}(20, 0.6)$. Then $S_1 = X_1 + \ldots + X_{10}$ with i.i.d. Bernoulli $X_1, \ldots, X_{10}$ with $p_1 = \ldots = p_{10} = 0.4$, and $S_2 = X_{11} + \ldots + X_{30}$ with i.i.d. Bernoulli $X_{11}, \ldots, X_{30}$ with $p_{11} = \ldots = p_{30} = 0.6$. Then $\mathbf{E}S = \mu = p_1 + \ldots + p_{30} = 10 \cdot 0.4 + 20 \cdot 0.6 = 16$. Therefore, $\mathbf{P}(S \leq 10) \leq \mathbf{P}(S \leq \mu(1-\delta))$ for $\delta = 3/8$, and is estimated as $\exp(-0.5 \cdot (3/8)^2 \cdot 16) = \boxed{0.32}$

10. Central Limit Theorem

10.1. **Statement of the CLT.** The LLN states: For $X_1, X_2, \ldots$ i.i.d. with mean $\mathbb{E}X_i = \mu$ and $\text{Var}\, X_i = \sigma^2$, if $S_n = X_1 + \ldots + X_n$, then $S_n/n \to \mu$ in probability. But how to estimate the speed of convergence? Note that

$$\mathbb{E}S_n = \mathbb{E}X_1 + \ldots + \mathbb{E}X_n = n\mu,$$

and by independence,

$$\text{Var}(S_n) = \text{Var}(X_1) + \ldots + \text{Var}(X_n) = n\sigma^2.$$

Following is the statement of the CLT.

Theorem 10.1. *We have the following convergence in law, as $n \to \infty$:*

$$\frac{S_n - \mathbb{E}S_n}{\sqrt{\text{Var}\, S_n}} = \frac{S_n - n\mu}{\sqrt{n}\sigma} \to \mathcal{N}(0,1).$$

Since normal distribution is continuous,

$$\mathbb{P}\left(a \leq \frac{S_n - n\mu}{\sqrt{n}\sigma} \leq b\right) \to \frac{1}{\sqrt{2\pi}} \int_a^b e^{-z^2/2}\, \mathrm{d}z = \Phi(b) - \Phi(a),$$

where Φ is the CDF of $\mathcal{N}(0,1)$:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-u^2/2}\, \mathrm{d}u.$$

Proof. See the book, Chapter 4, Appendix B. Compute MGF for $(S_n - n\mu)/\sqrt{n}\sigma$ and prove that as $n \to \infty$, it converges to $e^{t^2/2}$, which is MGF for $\mathcal{N}(0,1)$. □

In practice, these integrals are found from the Z-table.

Example 10.1. Suppose we have a plane of $N = 200$ people. Each customer, independently of others, chooses chicken or pasta, with probability 50%. How many chicken and pasta is needed to satisfy chicken-lovers with probability 95% and pasta-lovers with probability 95% (so that everybody will be satisfied with probability 90%)?

If we prepare only 100 chicken and 100 pasta, we will likely run out ofeither chicken or pasta: it is very unlikely that *exactly* 100 people will choose chicken. The other extreme is to prepare 200 chicken and 200 pasta. This guarantees that every client will be pleased, but a lot of food (200 meals) will be thrown away. Which is not good, because airlines have tight budgets. Let us find a compromise. Let $X_k = 1$ if the kth person chooses chicken, $X_k = 0$ if he chooses pasta. Then $S_N = X_1 + \ldots + X_N$ is the total number of chicken required. The random variable X_k is Bernoulli, with $p = 1/2$, so

$$\mu = \frac{1}{2}, \quad \sigma^2 = \frac{1}{4}, \quad \sigma = \frac{1}{2};$$

$$\frac{S_N - N\mu}{\sigma\sqrt{N}} = \frac{S_N - 100}{\sqrt{200}/2} \approx \mathcal{N}(0,1).$$

We can find $x_{95\%} = 1.645$ such that

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-u^2/2}\, du = 95\%.$$

Therefore, approximately

$$\mathbf{P}\left(\frac{S_N - 100}{\sqrt{200}/2} \leq 1.645\right) \approx 95\%.$$

With this high probability 95%,

$$\frac{S_N - 100}{\sqrt{200}/2} \leq 1.645 \quad \Leftrightarrow \quad S_N \leq 100 + \frac{1}{2}\sqrt{200} \cdot 1.645 = 112.$$

So we need only 112 chicken: 100 chicken according to the mean, and 12 additional, as a buffer. The same applies to pasta. We will throw away $12 + 12 = 24$ meals after the trip, which is not a lot. Or maybe just give them to pilots or flight attendants.

Example 10.2. A research sample $X_1, \dots, X_n$ of size $n = 100$ is drawn from distribution with density $f(x) = 3(1-x)^2,\ 0 \le x \le 1$. Let

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

Estimate $\mathbb{P}(0.2 \le \overline{X} \le 0.3)$. Solution: This distribution has mean and variance (after computation)

$$\mu = \int_0^1 x f(x)\,\mathrm{d}x = \frac{1}{4},\ \sigma^2 = \int_0^1 x^2 f(x)\,\mathrm{d}x - \mu^2 = \frac{3}{80}.$$

Thus $0.2 \le \overline{X} \le 0.3$ is equivalent to $0.2n \le S_n \le 0.3n$, that is,

$$\frac{0.2n - n\mu}{\sqrt{n}\sigma} \le \frac{S_n - n\mu}{\sqrt{n}\sigma} \le \frac{0.3n - n\mu}{\sqrt{n}\sigma}.$$

The lower limit is $a = -0.26$ and the upper limit is $b = 0.26$. By CLT, the probability of this event is approximately

$$\Phi(0.26) - \Phi(-0.26) = 0.6026 - (1 - 0.6026) = 0.2502.$$

10.2. **Normal and Poisson approximations.** Suppose you have a large quantity N of independent events, each of which happens with probability p. The number of these events is $\mathrm{Bin}(N, p)$. We have studied two approximations:

- Poisson, when $\mathrm{Bin}(N, p) \to \mathrm{Poi}(\lambda)$ if $p = \lambda/N$, as $N \to \infty$.
- Normal, when for $X \backsim \mathrm{Bin}(N, p)$, as $N \to \infty$, for constant p (independent of N):

$$\frac{X - pN}{\sqrt{pqN}} \to \mathcal{N}(0, 1).$$

The first approximation is used when the events are rare, the second - when they are "usual", not rare. In practice, if you have something like $N = 1000$ and $p = 0.1\%$, you should use Poisson, and if $N = 1000$ but $p = 10\%$, you should use Normal.

Example 10.3. An accountant lists each of $N = 100$ expenses with precision up to 1 dollar. That is, the rounding error is $\mathrm{Uni}[-0.5, 0.5]$. Find the probability that the total error exceeds 5. Let $X_k =$ error in the kth expense. The total error is $S = X_1 + \ldots + X_N$. We need $|S| \ge 5$. After computations, we get: $\mathbb{E}X_k = 0$ and $\operatorname{Var} X_k = 1/12$. Thus $\mathbb{E}S = 0$, $\operatorname{Var} S = N/12$. By the Central Limit Theorem,

$$\mathbf{P}(S \ge 5) = \mathbf{P}\left(\frac{S - \mathbb{E}S}{\sqrt{\operatorname{Var} S}} \ge \frac{5 - 0}{\sqrt{100/12}}\right) \approx 1 - \Phi(\frac{5 - 0}{\sqrt{100/12}}) = \ldots$$

Similarly, $\mathbf{P}(S \le -5) = \mathbf{P}(S \ge 5)$ because S is symmetric.

10.3. **Extensions of the Central Limit Theorem.** The CLT is valid not only for iid (independent identically distributed) random variables. It is (sometimes) true if variables are independent but not identically distributed, or if they have slight dependence. The search for the conditions under which CLT is valid is still an active research area. For the purposes of this class, you can assume it is true if $\operatorname{Var} S_N$ is large. This is how it is formulated in the general form.

Theorem 10.2. *For independent (but not identically distributed) $X_1, X_2, \ldots$, convergence in law:*

$$\frac{S_N - \mathbf{E}S_N}{\sqrt{\operatorname{Var} S_N}} \to \mathcal{N}(0, 1), \quad N \to \infty.$$

Example 10.4. An insurance company has $N = 10000$ clients. Half of them file a claim which is distributed as $\mathrm{Exp}(1)$. The other half file a claim $\mathrm{Exp}(2)$; all claims are independent. Then the total amount of claims is

$$S_N = X_1 + \ldots + X_{5000} + Y_1 + \ldots + Y_{5000},$$

where $X_i \backsim \mathrm{Exp}(1)$, $Y_i \backsim \mathrm{Exp}(2)$. Thus

$$\mathbf{E}X_i = 1,\ \operatorname{Var} X_i = 1,\ \mathbf{E}Y_i = 1/2,\ \operatorname{Var} Y_i = 1/4.$$
$$\mathbf{E}S_N = 5000 \cdot 1 + 5000 \cdot (1/2) = 7500, \quad \operatorname{Var} S_N = 5000 \cdot 1 + 5000 \cdot (1/4) = 6250.$$

By the Central Limit Theorem,

$$\frac{S_N - 7500}{\sqrt{6250}} \approx \mathcal{N}(0, 1).$$

11. Discrete-Time Markov Chains with Applications

11.1. **Definitions.** Assume there was only one fast-food restaurant in a small town: McDonalds. Every day customers went to this restaurant. But then another fast-food place, a Subway, opened there. People started to switch between these two places. Each next day, 20% of customers who went to McDonalds the day before, went to Subway. Conversely, 10% of customers who went to Subway the day before switched to McDonalds.

On the 0th day, before Subway opened, 100% of customers went to McDonalds and 0% to Subway. On the 1st day, 80% of customers went to McDonalds and 20% went to Subway. On the 2nd day, $0.8 \cdot 0.8 + 0.2 \cdot 0.1 = 0.66 = 66\%$ of customers went to McDonalds and $0.8 \cdot 0.2 + 0.2 \cdot 0.9 = 0.34 = 34\%$ of customers went to Subway. If at day n the share mc_n of customers went to McDonalds and the share sub_n went to Subway, then

$$\begin{cases} \mathrm{mc}_{n+1} = 0.8\,\mathrm{mc}_n + 0.1\,\mathrm{sub}_n, \\ \mathrm{sub}_{n+1} = 0.2\,\mathrm{mc}_n + 0.9\,\mathrm{sub}_n\,. \end{cases}$$

We can write this as

$$\begin{bmatrix} \mathrm{mc}_{n+1} & \mathrm{sub}_{n+1} \end{bmatrix} = \begin{bmatrix} \mathrm{mc}_n & \mathrm{sub}_n \end{bmatrix} A, \quad \text{where} \quad A = \begin{bmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \end{bmatrix}$$

is called the *transition matrix*. Consider a sequence of random variables

$$X = (X_n)_{n \geq 0} = (X_0, X_1, X_2, \ldots)$$

which evolves according to the following laws. Each random variable takes values M and S. The *initial distribution* is given by

$$\mathbf{P}(X_0 = M) = 1, \ \mathbf{P}(X_0 = S) = 0.$$

The distribution of X_{n+1} given X_n is given by

$$\begin{cases} \mathbf{P}(X_{n+1} = M \mid X_n = M) = 0.8, \\ \mathbf{P}(X_{n+1} = S \mid X_n = M) = 0.2, \\ \mathbf{P}(X_{n+1} = M \mid X_n = S) = 0.1, \\ \mathbf{P}(X_{n+1} = S \mid X_n = S) = 0.9. \end{cases}$$

Then we have:

$$\mathbf{P}(X_n = M) = \mathrm{mc}_n, \ \mathbf{P}(X_n = S) = \mathrm{sub}_n\,.$$

In other words, the vector $x(n) = \begin{bmatrix} \mathrm{mc}_n & \mathrm{sub}_n \end{bmatrix}$ is the *distribution* of X_n at time n. The random process $X = (X_0, X_1, \ldots)$ is called a *Markov chain*.

11.2. **Stationary and limiting distributions.** Does $x(n)$ converge to something as $n \to \infty$? Actually, it does. To find this limit, let us find nonzero vectors $v = \begin{bmatrix} v_1 & v_2 \end{bmatrix}$ such that $vA = \lambda v$ for some real number λ. These λ and v are called *eigenvalues* and *eigenvectors*. Usually, it is formulated in terms of column vectors rather than row vectors, but here it is convenient for us to study row vectors, and it does not really make any substantial difference.

We can rewrite this as $v(A - \lambda I_2) = 0$, where I_2 is the identity 2×2 matrix. Since we multiply a matrix by a nonzero vector and get zero vector, the matrix $A - \lambda I_2$ must be *nonsingular*: $\det(A - \lambda I_2) = 0$. But

$$\det(A - \lambda I_2) = \begin{vmatrix} 0.8 - \lambda & 0.2 \\ 0.1 & 0.9 - \lambda \end{vmatrix} = 0 \quad \Rightarrow \quad (0.8 - \lambda)(0.9 - \lambda) - 0.1 \cdot 0.2 = 0 \ \Rightarrow \ \lambda = 1, 0.7$$

These are called *eigenvalues*. Let us find a vector v corresponding to each of these eigenvalues. For $\lambda = 1$, we have:

$$v = vA \ \Rightarrow \ \begin{cases} v_1 = 0.8v_1 + 0.1v_2 \\ v_2 = 0.2v_1 + 0.9v_2 \end{cases} \quad \Rightarrow \quad 2v_1 = v_2 \ \Rightarrow \ v = \begin{bmatrix} 1 & 2 \end{bmatrix}$$

For $\lambda = 0.7$, we have:

$$0.7w = wA \ \Rightarrow \ \begin{bmatrix} 0.7w_1 = 0.8w_1 + 0.1w_2 \\ 0.7w_2 = 0.2w_1 + 0.9w_2 \end{bmatrix} \ \Rightarrow \ w_1 + w_2 = 0 \ \Rightarrow \ w = \begin{bmatrix} 1 & -1 \end{bmatrix}$$

Every vector $x(0)$ can be decomposed as a linear combination of these two vectors. For example, if $x(0) = \begin{bmatrix} 1 & 0 \end{bmatrix}$ (initially all customers went to McDonalds), then

$$x(0) = c_1 v + c_2 w.$$

How do we find c_1 and c_2? Solve the system of equations

$$\begin{cases} 1 = c_1 + c_2 \\ 0 = 2c_1 - c_2 \end{cases} \quad \Rightarrow \quad c_1 = \frac{1}{3}, \ \ c_2 = \frac{2}{3}.$$

Therefore, $x(0) = \frac{1}{3}v + \frac{2}{3}w$, and

$$x(n) = x(0)A^n = \frac{2}{3}0.7^n w + \frac{1}{3}v \to \frac{1}{3}v = \begin{bmatrix}\frac{1}{3} & \frac{2}{3}\end{bmatrix} =: p.$$

This is the *limiting distribution*. Also, it is a *stationary distribution*: if $x(0) = p$, then $x(1) = p$, because $p = pA$. If we start from this distribution, then we forever remain in this distribution. Any stationary distribution must be a solution to this equation $p = pA$, and the sum $p_1 + p_2 = 1$ (because total probability is 1). Here, p is a unique solution to this problem, that is, a unique probability distribution.

Actually, p is a limit regardless of the initial distribution. Indeed, suppose that $x(0)$ were different. Then this would change only c_1 and c_2: the matrix A would remain the same, with its eigenvalues and eigenvectors $\lambda_1, \lambda_2, v, w$. So $x(n) \to c_1 v$. But $x(n)$ is a probability distribution for each n, so its components sum up to 1. The same must be true for $c_1 v$, because it is the limit of $x(n)$. However, $v_1 = \begin{bmatrix}1 & 2\end{bmatrix}$, so we must have $c_1 = 1/(1+2) = 1/3$ and $c_1 v = \begin{bmatrix}1/3 & 2/3\end{bmatrix} = p$.

You can see that in the long run (actually, in a few weeks, because 0.7^{10} is already quite small) Subway will have $2/3$ of customers, and McDonalds will have only $1/3$ of them. In addition, each customer will spend approximately $2/3$ of days in the long-term in Subway, and $1/3$ in McDonalds. Say, among the first 1500 days approximately 1000 will be spent in Subway, and 500 in McDonalds. This type of statement is true if a Markov chain has a unique stationary distribution.

Assume a person always buys coffee with the meal. The coffee at McDonalds costs 1\$, and at Subway it costs 2\$. Then the *long-term average* cost of coffee is given by

$$1 \cdot p_1 + 2 \cdot p_2 = \frac{5}{3}.$$

This is the approximate average cost of coffee during the first N days, when N is large.

Not all Markov chains have a unique stationary distribution. For example,

$$A = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$$

then every distribution is stationary: $x(1) = x(0)A = x(0)$, and so $x(n) = x(0)$. The limit coincides with the stationary distribution! This corresponds to the case where all customers are completely loyal: if they went to McDonalds yesterday, they are going to make it there today, and same for Subway.

11.3. **General finite-state Markov chains.** Such a chain has *state space* $\{1, \dots, N\}$ of N elements; if p_{ij} is the probability of moving from state i to state j, then

$$p_{i1} + \dots + p_{iN} = 1,\ i = 1, \dots, N.$$

We can write these *transition probabilities* in the form of an $N \times N$ transition matrix

$$P = (p_{ij})_{i,j=1,\dots,N}.$$

The Markov chain is a collection $X = (X_n)_{n \geq 0} = (X_0, X_1, X_2, \dots)$ of random variables X_n, each of which takes values in the state space $\{1, \dots, N\}$. If $X_n = i$, then $X_{n+1} = j$ with probability p_{ij}. We can write this as

$$\mathbf{P}(X_{n+1} = j \mid X_n = i) = p_{ij}.$$

We denote by $x(n)$ the distribution of X_n:

$$x(n) = \begin{bmatrix}x_1(n) & \dots & x_N(n)\end{bmatrix},\ x_i(n) = \mathbf{P}(X_n = i).$$

These components $x_1(n), \dots, x_N(n)$ must satisfy

$$x_1(n) + \dots + x_N(n) = 1.$$

Also, we have the following matrix multiplication formula:

$$x(n+1) = x(n)P.$$

The difference between X_n and $x(n)$ is that X_n are random variables, and $x(n)$ are vectors.

Every Markov chain (with a finite state space) has a stationary distribution (at least one, possibly more than one). But it does not necessarily converge to this stationary distribution. However, if it converges, then each probability p_i in this stationary distribution corresponds to the long-term proportion of time spent in this state i.

11.4. **Rate of convergence.** Consider the transition matrix

$$A = \begin{bmatrix} 0 & 0.3 & 0.7 \\ 0.2 & 0.8 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

It has a unique stationary distribution $p = \begin{bmatrix} p_1 & p_2 & p_3 \end{bmatrix}$, which can be found from

$$p = pA, \quad p_1 + p_2 + p_3 = 1.$$

Let us solve this system of equations:

$$p_1 = 0.2p_2 + p_3, \quad p_2 = 0.3p_1 + 0.8p_2, \quad p_3 = 0.7p_1.$$

Therefore, from the second equation we get:

$$0.2p_2 = 0.3p_1 \ \Rightarrow \ p_2 = 1.5p_1.$$

And $p_3 = 0.7p_1$, so $1 = p_1 + p_2 + p_3 = (1.5 + 1 + 0.7)p_1 = 3.2p_1$, and $p_1 = 1/3.2 = 5/16$, $p_2 = 15/32$, $p_3 = 7/32$. What is the rate of convergence of $x(n) = x(0)A^n$, the distribution at the nth step, to this distribution? Let us find eigenvalues and eigenvectors of A. Eigenvalues:

$$\lambda_1 = 1, \quad \lambda_2 = \frac{1}{10}\left(\sqrt{57} - 1\right), \quad \lambda_3 = \frac{1}{10}\left(-\sqrt{57} - 1\right).$$

We have: $|\lambda_2| < |\lambda_3| < 1$. Let v_1, v_2, v_3 be eigenvectors corresponding to $\lambda_1, \lambda_2, \lambda_3$:

$$v_1\lambda_1 = v_1A, \quad v_2\lambda_2 = v_2A, \quad v_3\lambda_3 = v_3A.$$

We can take $v_1 = p$ because $\lambda_1 = 1$. Then for any initial distribution $x(0)$ we can decompose

$$x(0) = c_1v_1 + c_2v_2 + c_3v_3,$$

for some numbers c_1, c_2, c_3, and

$$x(n) = x(0)A^n = c_1v_1A^n + c_2v_2A^n + c_3v_3A^n = c_1p + c_2\lambda_2^n v_2 + c_3\lambda_3^n v_3.$$

As $n \to \infty$, $x(n) \to c_1p$. But we know that $x(n) \to p$, so $c_1 = 1$. Therefore,

$$x(n) - p = c_2\lambda_2^n v_2 + c_3\lambda_3^n v_3,$$

$$|x(n) - p| \le c_2|\lambda_2|^n|v_2| + |c_3||\lambda_3|^n|v_3|.$$

As $n \to \infty$, $|\lambda_2|^n$ converges to zero faster than $|\lambda_3|^n$, because $|\lambda_2| < |\lambda_3|$. Therefore, the whole expression

$$c_2|\lambda_2|^n|v_2| + |c_3||\lambda_3|^n|v_3|$$

converges to zero with the same rate as $|c_3||\lambda_3|^n|v_3|$. We say that the rate of convergence is $|\lambda_3|^n$ (because the rest are just constants, independent of n).

11.5. **Recurrence and transience.** Let $f_i := \mathbf{P}(S \text{ ever returns to the state } i \mid S_0 = i)$ be the probability that the Markov chain returns to the position i if it started from i. The state i is called *recurrent* if $f_i = 1$, and *transient* $f_i < 1$. If the state i is recurrent, then the process returns to i. But then, starting again from there, it again returns to i, and again, etc. So the process returns to the recurrent state infinitely often. If N_i is the number of times when the process returns to i, then $N_i = \infty$, and $\mathbf{E}N_i = \infty$.

For a transient state i, with probability $1 - f_i$ the process never returns there. With probability f_i, it does return there. After this, with probability $f_i(1 - f_i)$, it never returns there again. With probability f_i^2, it returns there for the second time. With probability $f_i^2(1 - f_i)$, it never returns there again, etc. So

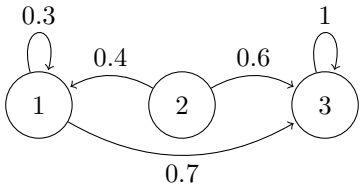
$$\mathbf{P}(N_i = k) = f_i^k(1 - f_i), \quad k = 0, 1, 2, \dots$$

Therefore, N_i has geometric distribution with parameter $1 - f_i$, and $\mathbf{E}N_i = (1 - f_i)^{-1} < \infty$. If we start from a state and then return there infinitely many times, this state is recurrent; otherwise, it is transient.

Example 11.1. Consider the following Markov chain:

$$A = \begin{bmatrix} 0.3 & 0 & 0.7 \\ 0.4 & 0 & 0.6 \\ 0 & 0 & 1 \end{bmatrix}$$

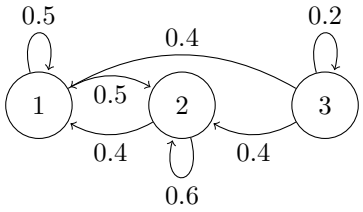
We can draw the diagram



Then you will eventually get to state 3; states 1 and 2 are transient, state 3 is recurrent. This chain is irreducible and aperiodic; therefore, it is ergodic. Let $p = \begin{bmatrix} p_1 & p_2 & p_3 \end{bmatrix}$ be the stationary distribution. We can solve for the stationary distribution: $p = pA$, and find $p = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$.

Example 11.2. Consider the following Markov chain:

$$A = \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0.4 & 0.6 & 0 \\ 0.4 & 0.4 & 0.2 \end{bmatrix}$$



Stationary distribution $p = \begin{bmatrix} p_1 & p_2 & p_3 \end{bmatrix}$ is determined from the equation $p = pA$ and the additional condition $p_1 + p_2 + p_3 = 1$. But the state 3 is transient, therefore $p_3 = 0$ (see the previous example), and so we have the system of equations

$$\begin{cases} p_1 = 0.5p_1 + 0.4p_2 + 0.4p_3 \\ p_2 = 0.5p_1 + 0.6p_2 + 0.4p_3 \\ p_3 = 0.2p_3 \end{cases} \Rightarrow\ 0.5p_1 = 0.4p_2.$$

Because $p_1 + p_2 + p_3 = 1$, we have: $p_1 + p_2 = 1$. Therefore, $p_1 = 4/9$, and $p_2 = 5/9$.

11.6. **Irreducibility.** Take a Markov chain and ignore its transient states, because the Markov chain will eventually leave them forever. In other words, suppose all states are recurrent. Take two states i, j. Suppose we can get from i to j with positive probability (in a certain number of steps): we write this as $i \to j$. Then we can also get from j to i with positive probability: $j \to i$. Otherwise, j would serve as a "sink": you leave i for j and you do not return to i with positive probability. Then the probability of return to i, starting from i, is less than 1, and this means i is transient. So if $i \to j$, then $j \to i$; we write this as $i \leftrightarrow j$ and call these states *communicating*.

We can split the state space into a few classes of communicating states. If there is more than one class, this chain is called *reducible*, and if there is only one class, it is called *irreducible*. Each such class can be considered as a Markov chain of its own. [2]

If the chain is reducible, there is more than one stationary distribution. Indeed, suppose the Markov chain has a state space $\{1, 2, 3, 4, 5\}$, where $\{1, 3, 5\}$ form an irreducible class with stationary distribution

$$\pi_1 = \frac{1}{2},\ \pi_3 = \frac{3}{8},\ \pi_5 = \frac{1}{8},$$

and $\{2, 4\}$ form an irreducible class with stationary distribution

$$\pi_2 = \frac{1}{3},\ \pi_4 = \frac{2}{3}.$$

Then there are many stationary distributions for the whole chain: add the first distribution multiplied by p_1 and the second distribution multiplied by p_2, where $p_1 + p_2 = 1$, $p_1, p_2 \geq 0$. Therefore,

$$\pi = \begin{bmatrix} \frac{1}{2}p_1 & \frac{1}{3}p_2 & \frac{3}{8}p_1 & \frac{2}{3}p_2 & \frac{1}{8}p_1 \end{bmatrix}$$

You can view it as follows: we toss a coin which has Heads with probability p_1 and Tails with probability p_2. If Heads, we move to the class $\{1, 3, 5\}$ and start from distribution $\begin{bmatrix} \pi_1 & \pi_3 & \pi_5 \end{bmatrix}$, and forever remain in this distribution

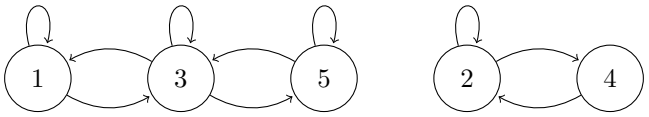
[2] Some textbooks use a different definition of irreducibility, without removal of transient states. However, we shall utilize the one given above, because it is related to stationary distributions, as shown below.

(because it is stationary). If Tails, we move to the class $\{2,4\}$ and start from the distribution $\begin{bmatrix}\pi_2 & \pi_4\end{bmatrix}$, and forever remain in this distribution. For example, take $p_1 = 3/4$ and $p_2 = 1/4$; then

$$\pi = \begin{bmatrix}\frac{3}{8} & \frac{1}{12} & \frac{9}{32} & \frac{1}{6} & \frac{3}{32}\end{bmatrix}$$

We might also take $p_1 = 0$ and $p_2 = 1$; then we only move in the class $\{2,4\}$, and

$$\pi = \begin{bmatrix}0 & \frac{1}{3} & 0 & \frac{2}{3} & 0\end{bmatrix}$$



We showed that if a Markov chain is reducible, then there are multiple stationary distributions. (Because there exists a stationary distribution for each communicating class, and we can combine them as above in many ways.) We can also show the following theorem.

If a Markov chain is irreducible, then there is only one stationary distribution.

11.7. **Aperiodicity.** Again, ignore transient states, and suppose the Markov chain is irreducible (or, in case it is reducible, move to any class of communicating states). Start with some element i and find the number of steps needed to get there; let $\varphi_1, \varphi_2, \dots$ be numbers of steps for different paths. Take d, the greatest common divisor of $\varphi_1, \varphi_2, \dots$. This is called the *period* of this Markov chain. It does not depend on i. This means we can split the state space into d subclasses such that if you start in the 1st subclass, at the next step you get to the 2nd one, etc. to the dth one, and then to the first one. If $d = 1$, the Markov chain is called *aperiodic*, and if $d \geq 2$, it is called *periodic*. If there is a loop (there exists a state from which you can return to itself at the next step), then $\varphi_1 = 1$, and $d = 1$.

As mentioned before, there exists a unique stationary distribution p. However, if $d \geq 2$ (the chain is periodic), then it might not converge to this stationary distribution p. That is, not necessarily $x(n) \to p$ as $n \to \infty$. [3]

Example 11.3. Consider a Markov chain with the state space $\{1,2\}$ and the transition matrix

$$A = \begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}$$

This chain is 2-periodic, and its stationary distribution p satisfies

$$p = pA \;\Rightarrow\; \begin{cases}p_1 = p_2\\ p_2 = p_1\end{cases} \;\Rightarrow\; p_1 = p_2 = \frac{1}{2}.$$

But if $x(0) = \begin{bmatrix}1/3 & 2/3\end{bmatrix}$, then $x(1) = x(0)A = \begin{bmatrix}2/3 & 1/3\end{bmatrix}$, $x(2) = x(1)A = \begin{bmatrix}1/3 & 2/3\end{bmatrix}$, etc. So

$$x(n) = x(0) \;\text{ for even }\; n, \quad x(n) = x(1) \;\text{ for odd }\; n.$$

Therefore, $x(n)$ does not have a limit. The distribution switches between $x(0)$ and $x(1)$.

If the Markov chain is irreducible and aperiodic, then $x(n) \to p$ for $n \to \infty$, regardless of the initial distribution $x(0)$. Moreover, $p_i > 0$ for each state i. In this case, the Markov chain is called *ergodic*.

11.8. **Summary.** When we analyze a Markov chain X, the following steps are required:

Identify all transient and all recurrent states: For every state, you should find whether it is transient or recurrent. Then remove all transient states. You get a certain Markov chain Y, with all states being recurrent.

Is the resulting Markov chain Y irreducible? That is, can you get from any state to any state? If yes, the Markov chain is irreducible. If not, it splits into more than one communicating classes $C^1, \dots, C^r$. Each such class is by itself an irreducible Markov chain, with all its states being recurrent.

Each such communicating class has a unique stationary distribution. In fact, an irreducible Markov chain has a unique stationary distribution. If the class C^i has a stationary distribution π^i, then we can weigh them (as explained earlier in this section) with certain weight coefficients $p_1, \dots, p_r$ which satisfy

$$p_1, \dots, p_r \geq 0, \; p_1 + \dots + p_r = 1,$$

[3] Some textbooks use another definition of aperiodicity, without removal of transient states. However, we shall remove transient states first, because the connection to stationary distributions becomes clearer.

and get a stationary distribution π for the Markov chain Y. If a Markov chain is reducible (that is, if it consists of more than one communicating class), then it has more than one (in fact, infinitely many) stationary distributions, because $p_1, \dots, p_r$ can be chosen in multiple ways.

Every stationary distribution of Y corresponds to a stationary distribution of X: add zeroes in the places corresponding to the removed transient states. Every stationary distribution has zeroes in components corresponding to every transient state.

The stationary distribution for an irreducible Markov chain $Z = (Z_n)_{n\geq 0}$ has the meaning of long-term time proportions spent in this state. Take the state i, and let π_i be the corresponding component of the stationary distribution π. Let $N_i(n)$ be the number of steps $k = 1, \dots, n$ spent in i: $Z_k = i$. Then we have:

$$\frac{N_i(n)}{n} \to \pi_i \text{ a.s. as } n \to \infty.$$

In particular, if i is a transient state, then the Markov chain spends only finitely many steps in i. Say it spends M_i steps there. Then $N_i(n) = M_i$ for large enough n, so $N_i(n)$ does not grow as $n \to \infty$. Thus,

$$\frac{N_i(n)}{n} \to 0 \text{ a.s. as } n \to \infty.$$

However, if i is a recurrent state, then $\pi_i > 0$. In other words, the long-term proportion of steps spent at i is positive. During the first n steps for a large n, the Markov chain spends approximately $\pi_i n$ steps in the state i.

Finally, fix a communicating class; or, alternatively, assume the Markov chain is irreducible and does not contain any transient states. We can find its period d by fixing the state i, and taking the gcd of lengths of all paths returning from i to i, as explained above. This number is the same for all states i. Fix a state i; it is in a subclass D_0. All states j where you can go from i in one step belong to the next subclass D_1. All states where you can get in one step from some states in D_1 form the subclass D_2, etc. The whole Markov chain now is split into d subclasses $D_0, \dots, D_{d-1}$, and the dynamics is as follows:

$$D_0 \Rightarrow D_1 \Rightarrow D_2 \Rightarrow \dots \Rightarrow D_{d-1} \Rightarrow D_0.$$

If an irreducible Markov chain is aperiodic ($d = 1$), then for every initial distribution $x(0)$, we have convergence to the stationary distribution:

$$x(n) \to \pi,\ n \to \infty.$$

However, if this Markov chain is periodic ($d \geq 2$), then for some initial distribution $x(0)$ there is no such convergence. For example, consider the initial distribution concentrated on D_0. That is, $X_0 \in D_0$: the initial state is among the states from the subclass D_0. Then at the next step the process goes to the subclass D_1: $X_1 \in D_1$. Next,

$$X_2 \in D_2,\ X_3 \in D_3,\ \dots, X_{d-1} \in D_{d-1},\ X_d \in D_0,\ X_{d+1} \in D_1, \dots$$

For each step n, only some of the components of the distribution $x(n)$ are positive: those who correspond to the respective class. Others are zero. But for the stationary distribution π, all components are positive. So we cannot have convergence $x(n) \to \pi$ as $n \to \infty$.

Every comunicating class has its own period; they might be different.

The overall scheme is as follows:

(1) For each state, find whether it is recurrent or transient. Remove all transient states.
(2) Split the resulting Markov chain (with only recurrent states left) into communicating classes. If there is only one, the chain is irreducible; otherwise, it is reducible.
(3) For each communicating class, find its period and split into subclasses. If the period is 1, this class is aperiodic; otherwise, it is periodic.

11.9. **Time spent in transient states.** Assume we have a Markov chain

$$\begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \end{bmatrix}$$

States 1 and 2 are transient. If we start from state 1, then every attempt to escape succeeds with probability 0.4 and fails with probability 0.6. And once it leaves state 1, it never returns there. Therefore, time T_1 spent in 1 is distributed geometrically with parameter 0.4. In particular,

$$\mathbf{P}(T_1 = n) = 0.4 \cdot 0.6^{n-1},\ n = 1, 2, \dots; \quad \mathbf{E}T_1 = \frac{1}{0.4} = 1.25.$$

Similarly, if we start from state 2, then time T_2 spent in 2 is Geo(0.5), and $\mathbf{E}T_2 = 1/0.5 = 2$.

11.10. **Comparison of hitting times.** Consider the Markov chain with four states $1, 2, 3, 4$ and transition matrix

$$\begin{bmatrix} 0.5 & 0.3 & 0.2 & 0 \\ 0.3 & 0.4 & 0.2 & 0.1 \\ * & * & * & * \\ * & * & * & * \end{bmatrix}$$

Let p_x be the probability that, starting from x, the process will reach 3 before it reaches 4. By definition,

$$p_3 = 1, \quad p_4 = 0.$$

$$\begin{cases} p_1 = 0.5p_1 + 0.3p_2 + 0.2p_3; \\ p_2 = 0.3p_1 + 0.4p_2 + 0.2p_3 + 0.1p_4 \end{cases}$$

Derive the first equation: When we start from 1, we can move back to 1 with probability 0.5, and then we start from 1 again; or we can move to 2 with probability 0.3, then we start from 2, etc. We can rewrite these equations as

$$0.5p_1 = 0.3p_2 + 0.2, \quad 0.6p_2 = 0.3p_1 + 0.2.$$

It suffices to solve them and find p_1 and p_2.

11.11. **Random walk on graphs.** Consider a graph with finitely many vertices $1, 2, \ldots, N$. A Markov chain starts from some vertex and chooses its neighbors with equal probability. We write $i \leftrightarrow j$ if the vertices i and j are connected. Let $\deg i$ be the degree of the vertex i (the number of neighbors for i). Therefore, the transition matrix is given by

$$A_{ij} = \begin{cases} (\deg i)^{-1}, & i \leftrightarrow j; \\ 0, & \text{otherwise.} \end{cases}$$

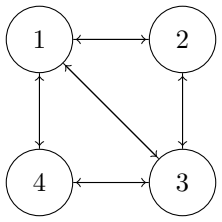
Let us take the stationary distribution

$$\pi_i := \frac{\deg i}{2M}, \ i = 1, \ldots, N.$$

where M is the total number of edges. Let us check that this is indeed a stationary distribution. For each vertex j, we have:

$$(\pi A)_j = \sum_i \pi_i A_{ij} = \sum_{i \leftrightarrow j} \pi_i A_{ij} = \sum_{i \leftrightarrow j} \frac{\deg i}{2M} \frac{1}{\deg i} = \frac{1}{2M} \sum_{i \leftrightarrow j} 1 = \frac{\deg i}{2M}.$$

For example, the following graph:



has $M = 5$, and stationary distribution

$$p_1 = p_3 = \frac{3}{10}, \ p_2 = p_4 = \frac{2}{10} = \frac{1}{5}.$$

This Markov chain (or, rather, its analogue for directed graph) is important for Google PageRank search algorithm: see section on simulation.

If the graph is connected, then the random walk on this graph is irreducible and has a unique stationary distribution. However, if the graph is disconnected, then the random walk is reducible and has more than one stationary distribution.

11.12. **Application: a Bonus-Malus system.** An insurance company classifies car drivers into three tiers: Tier 0 consists of the best drivers, tier 1 consists of not-so-good drivers, and tier 2 consists of bad drivers. This is the table which shows the change between current year and next year, depending on the number of accidents this year:

Tier	0	1	2	≥ 3
0	0	1	1	2
1	0	1	2	2
2	1	2	2	2

Say, if a driver from the first tier had no accidents this year, this driver moves to the zeroth tier next year. Assume now that the number of accidents is always $X \sim \text{Poi}(2)$ for the first two tiers and $Y \sim \text{Poi}(3)$ for the worst tier. Then

$$\mathbf{P}(X=0) = e^{-2},\ \mathbf{P}(X=1) = \mathbf{P}(X=2) = 2e^{-2},\ \mathbf{P}(X \geq 3) = 1 - 5e^{-2},$$
$$\mathbf{P}(Y=0) = e^{-3},\ \mathbf{P}(Y \geq 1) = 1 - e^{-3}.$$

The transition matrix is

$$P = \begin{bmatrix} e^{-2} & 4e^{-2} & 1-5e^{-2} \\ e^{-2} & 2e^{-2} & 1-3e^{-2} \\ 0 & e^{-3} & 1-e^{-3} \end{bmatrix}$$

We can find the stationary distribution for this Markov chain: $\begin{bmatrix} p_0 & p_1 & p_2 \end{bmatrix}$ by solving the system of equations $pP = p$. If r_k is the premium for tier k, then the long-term average premium is $p_0 r_0 + p_1 r_1 + p_2 r_2$.

12. Random Walk and the Binomial Model in Finance

12.1. **Construction.** Consider a sequence $X_1, X_2, \ldots$ of i.i.d. (independent identically distributed) random variables with distribution $\mathbf{P}(X_i = 1) = p$ and $\mathbf{P}(X_i = -1) = q$, $p + q = 1$. Take

$$S_0 := x, \text{ and } S_n = S_{n-1} + X_n,\ n = 1, 2, \ldots$$
$$S_1 = x + X_1,\ S_2 = x + X_1 + X_2,\ S_3 = x + X_1 + X_2 + X_3, \ldots$$

Each time, the particle moves either one step to the right (with probability p) or one step to the left (with probability q). Each next step is independent of the past. An example of the trajectory starting from $x = 0$ is:

$$(S_0 = 0,\ S_1 = 1,\ S_2 = 2,\ S_3 = 1,\ S_4 = 2,\ S_5 = 3, \ldots)$$

This process $(S_0, S_1, \ldots) = (S_n)_{n \geq 0}$ is called a *random walk*. When $p = q = 1/2$, this is called *simple* or *symmetric random walk*.

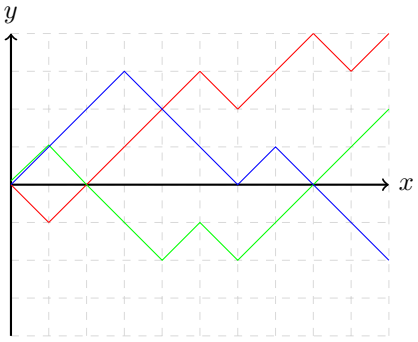
Example 12.1. Consider a random walk starting from $S_0 = -1$ with $p = 2/3$ and $q = 1/3$. Find $\mathbf{E}S_{12}$ and $\operatorname{Var} S_{12}$.

$$\mathbf{E}X_k = 1 \cdot p + (-1) \cdot q = \frac{1}{3},$$

$$\mathbf{E}X_k^2 = 1^2 \cdot p + (-1)^2 \cdot q = 1,\ \operatorname{Var} X_k = \mathbf{E}X_k^2 - (\mathbf{E}X_k)^2 = 1 - \left(\frac{1}{3}\right)^2 = \frac{8}{9}.$$

$$\mathbf{E}S_{12} = \mathbf{E}\left[-1 + X_1 + \ldots + X_{12}\right] = -1 + 12 \cdot \frac{1}{3} = \boxed{3}$$

$$\operatorname{Var} S_{12} = \operatorname{Var} X_1 + \ldots + \operatorname{Var} X_{12} = 12 \cdot \frac{8}{9} = \boxed{\frac{32}{3}}$$



12.2. **Transition probabilities.** The *transition probabilities* are $\mathbf{P}(S_n = z \mid S_m = y)$. Let us give an example on how to calculate them: take $\mathbf{P}(S_{10} = 4 \mid S_5 = 3)$. The quantity of paths from $(5, 3)$ to $(10, 4)$: suppose the path has a steps up and b steps down. The total number of steps is $a + b = 5$, and $a - b = 1$ (because the total ascent from 3 to 4 is 1). So $a = (5+1)/2 = 3$, and $b = (5-1)/2 = 2$. The path made 3 steps up and 2 steps down. There are five possible slots for upward steps; we must choose three upward steps, the rest will be occupied by downward steps. The number of ways to choose this is $\binom{5}{3} = (5 \cdot 4)/2 = 10$. Each such path has probability $p^3 q^2$, and so the total probability is

$$\mathbf{P}(S_{10} = 4 \mid S_5 = 3) = 10 p^3 q^2.$$

12.3. **Recurrence and transience.** Recall the definition of recurrent and transient states for random walk. Let N_i be the number of times when the process returns to this state. If $N_i = \infty$, and therefore $\mathbf{E}N_i = \infty$, this state is recurrent. If N_i is finite (Geometric distribution) and therefore $\mathbf{E}N_i < \infty$, it is transient. Now, let us find which of two cases holds for the random walk. Since all states are alike, let us take the state $i = 0$. Let 1_A be the indicator of the event A:

$$1_A = \begin{cases} 1, \text{ if } A \text{ happened;} \\ 0, \text{ if } A \text{ did not happend} \end{cases} \qquad \text{then } \mathbf{E}1_A = \mathbf{P}(A).$$

Because the random walk can return to zero only in an even number $n = 2k$ of steps.

$$N_0 = \sum_{n=1}^{\infty} 1_{\{S_n = 0\}}, \quad \mathbf{E}N_0 = \sum_{n=1}^{\infty} \mathbf{P}(S_n = 0) = \sum_{k=1}^{\infty} \mathbf{P}(S_{2k} = 0).$$

Let us calculate $\mathbf{P}(S_{2k} = 0)$. To get in $2k$ steps from 0 to 0, the process has to make k steps up and k steps down. The number of ways to choose k steps upward is $\binom{2k}{k}$. Each such path has probability $p^k q^k$. Therefore,

$$\mathbf{P}(S_{2k} = 0) = \binom{2k}{k} p^k q^k.$$

Now, we need to find when the series

$$\sum_{k=1}^{\infty} \binom{2k}{k} (pq)^k$$

converges or diverges. By the Stirling formula, we have: $k! \sim \sqrt{2\pi k}(k/e)^k$. Therefore,

$$\binom{2k}{k} = \frac{(2k)!}{k!^2} \sim \frac{\sqrt{2\pi \cdot 2k}(2k/e)^{2k}}{\left(\sqrt{2\pi k}(k/e)^k\right)^2} = \frac{2^{2k}}{\sqrt{\pi k}}.$$

The series

$$\sum_{k=1}^{\infty} \frac{2^{2k}}{\sqrt{\pi k}} (pq)^k = \sum_{k=1}^{\infty} \frac{(4pq)^k}{\sqrt{\pi k}}$$

converges if $4pq < 1$, because of the geometric series:

$$\sum_{k=1}^{\infty} \frac{(4pq)^k}{\sqrt{\pi k}} \le \sum_{k=1}^{\infty} (4pq)^k < \infty.$$

And it diverges if $4pq = 1$, because

$$\sum_{k=1}^{\infty} \frac{1}{\sqrt{\pi k}} = \infty.$$

Now, if $p = q = 1/2$, then $4pq = 1$, and otherwise $4pq = 4p(1-p) < 1$. Indeed,

$$1 - 4p(1-p) = 1 - 4p + 4p^2 = (1-2p)^2 \ge 0,$$

and this equals to zero only if $p = 1/2$. Result: **for the symmetric random walk, every state is recurrent; otherwise, every state is transient**.

12.4. **Gambler's ruin.** Consider a random walk $X_0, X_1, X_2, \ldots$, with probability p of going up and probability q of going down. Suppose $X_0 = n$. What is the probability q_n that this random walk hits N before hitting zero? We can restate it: if a gambler wants to get N dollars, starting from n dollars, but does not want to go bankrupt before that, what is the probability that he can pull this out?

Let us deduce this for $N = 3$, $p = .25$, $q = .75$. We will simulateously deduce this for $n = 0, 1, 2, 3$. At the first step, starting from n, the random walk can either go up (with probability p) and reach $n + 1$ or go down (with probability q) and reach $n - 1$. Therefore,

$$q_n = pq_{n+1} + qq_{n-1}, \quad n = 1, \ldots, N-1. \tag{6}$$

Also, $q_0 = 0$, because if the gambler already had nothing in the beginning, he is bankrupt. And $q_N = 1$, because if the gambler started with N dollars, he already achieved this goal. Plugging $n = 1$ and $n = 2$ in (6), we get:

$$q_2 = .25 \cdot q_3 + .75 \cdot q_1 = .25 + .75 \cdot q_1; \quad q_1 = .25 \cdot q_2 + .75 \cdot q_0 = .25 \cdot q_2.$$

Solving this system, we get: $q_2 = .25 + .75 \cdot .25 \cdot q_2 = 1/4 + (3/16)q_2$, therefore $q_2 = 4/13$, and $q_1 = .25q_2 = 1/13$. In general case, we get with $\alpha = q/p$:

$$q_k = \begin{cases} \frac{1-\alpha^k}{1-\alpha^N}, \ \alpha \neq 1; \\ \frac{k}{N}, \ \alpha = 1. \end{cases}$$

If $p > q$, then $\alpha < 1$, and as $N \to \infty$, we have:

$$q_k \to 1 - \alpha^k > 0.$$

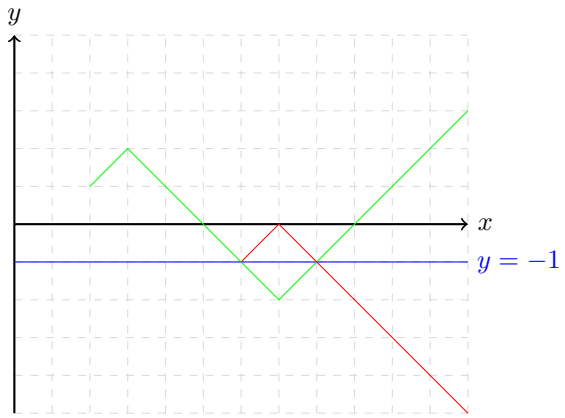
Therefore, $1 - \alpha^k$ is the probability of never becoming bankrupt for a good gambler: with $\alpha < 1$, who is more likely to win than lose at each step, who is playing against a bank with an infinite amount of money.

12.5. **Reflection principle.** Consider $(S_n)_{n \geq 0}$, a random walk. Let us find the quantity of paths from $S_2 = 1$ to $S_{12} = 3$ such that they do not hit or cross the line $y = -1$.

First, let us find the quantity of all paths from $S_2 = 1$ to $S_{12} = 3$, that is, from $(2, 1)$ to $(12, 3)$, regardless of whether they cross or hit this line. Let n and m be the quantity of up and down steps. Then the total number of steps is $n + m = 12 - 2 = 10$, and the change in altitude is $n - m = 3 - 1 = 2$. Therefore, $n = 6$ and $m = 4$. There are $\binom{10}{6}$ ways to choose 6 upward steps out of 10, so $\binom{10}{6}$ such random walk paths.

Then, let us find the quantity of these paths that do hit or cross this line $y = -1$. Starting from this hitting point, reflect them across this line. Then we get a new random walk path, which goes from $(2, 1)$ to $(12, -5)$ (the point symmetric to $(12, 3)$ with respect to the line $y = -1$). Every path from $(2, 1)$ to $(12, 3)$ which hits $y = -1$ corresponds to a path from $(2, 1)$ to $(12, -5)$, and vice versa. Let us find the quantity of paths from $(2, 1)$ to $(12, -5)$. If it has n steps up and m steps down, then $n + m = 12 - 2 = 10$ and $n - m = (-5) - 1 = -6$. Therefore, $n = 2$ and $m = 8$. The number of such paths is $\binom{10}{8}$

Thus, the quantity of paths from $(2, 1)$ to $(12, 3)$ which do not cross the line $y = -1$ is $\binom{10}{6} - \binom{10}{8} = \frac{10 \cdot 9 \cdot 8 \cdot 7}{4!} - \binom{10 \cdot 9}{2} = 210 - 45 = 165$.



Another way to say that the path $(S_n)_{n \geq 0}$ of random walk from $(2, 1)$ to $(12, 3)$ does not cross or hit $y = -1$ is

$$S_{12} = 3, \ S_n \geq 0, \ n = 2, \ldots, 12.$$

Assume we have a random walk with $p = 0.6$ and $q = 0.4$. Every such path has 6 steps up and 4 steps down (see above), and the probability of each such path is $p^6 q^4$. Therefore, the probability of this event is

$$\mathbf{P}\left(S_{12} = 3, \ S_n \geq 0, \ n = 2, \ldots, 12 \mid S_2 = 1\right) = \boxed{165 \cdot p^6 q^4}$$

12.6. **The ballot problem.** Assume there were 1000 voters and two candidates, A and B, who got 600 and 400 votes, respectively. Therefore A won. We start counting ballots.

How many different paths of counting ballots, that is, random walks $(S_n)_{n \geq 0}$ from $(0, 0)$ to $(1000, 200)$? Here, S_n is the ballots cast for A minus ballots cast for B when nth ballot has been counted. Therefore, $S_0 = 0$ and $S_{1000} = 600 - 400 = 200$. There has to be 600 ballots for A, and there are 1000 total ballots. The number of ways to choose them is $\binom{1000}{600}$.

How many such paths when A is always ahead of B? The first ballot must be for A, otherwise B immediately comes ahead. Next, how many paths from $(1, 1)$ to $(1000, 200)$ which do not hit the line $y = 0$? Similarly to the previous subsection, we can find: Each such path has n steps up and m steps down, with $n + m = 1000 - 1 = 999$, and $n - m = 200 - 1 = 199$. Therefore, $n = 599$, $m = 400$, and there are $\binom{999}{400}$ such paths.

How many such paths do cross or hit this line? Reflect it, starting from the first hitting point, across this line $y = 0$. Then we get a random walk path from $(1, 1)$ to $(1000, -200)$. It has n steps up and m steps down, with $n + m = 1000 - 1 = 999$, and $n - m = -200 - 1 = -201$. Therefore, $n = 399$ and $m = 600$. There are $\binom{999}{399}$ such paths. Therefore, the number of ways to count ballots such that A is always ahead of B is:

$$\binom{999}{400} - \binom{999}{399}$$

We can rewrite this as (because $1000! = 1000 \cdot 999!$ and $600! = 600 \cdot 599!$)

$$\frac{999!}{599!400!} - \frac{999!}{399!600!} = \frac{600}{1000}\frac{1000!}{600!400!} - \frac{400}{1000}\frac{1000!}{600!400!} = \frac{600-400}{1000}\binom{1000}{600}.$$

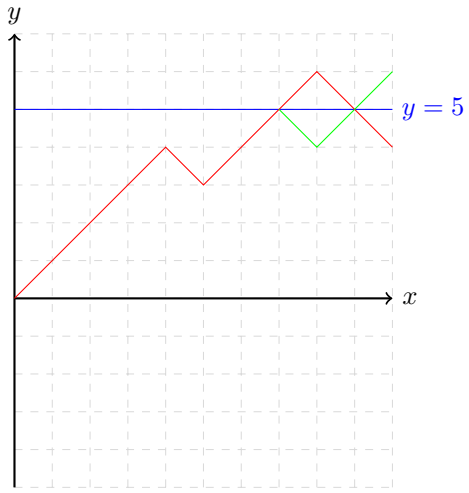
The probability that A was always ahead of B is

$$\frac{600-400}{1000} = \boxed{\frac{1}{5}}$$

12.7. **The maximum of a random walk.** Using the reflection principle, we can find the distribution of the maximum

$$M_n := \max(S_0, S_1, \dots, S_n)$$

of the symmetric random walk, with $p = q = 0.5$. For example, let us find $\mathbf{P}(M_{10} \geq 5)$. This event $\{M_{10} \geq 5\}$ can happen in two cases: (a) $S_{10} \geq 5$; (b) $M_{10} \geq 5$ but $S_{10} < 5$. By reflection principle, the probability of the second event is equal to $\mathbf{P}(S_{10} > 5)$.



Indeed, starting from the first hitting point of the line $y = 5$, we reflect this trajectory of a random walk across this line. Thus,

$$\mathbf{P}(M_{10} \geq 5) = \mathbf{P}(S_{10} \geq 5) + \mathbf{P}(S_{10} > 5) = \mathbf{P}(S_{10} \geq 5) + \mathbf{P}(S_{10} \geq 6). \tag{7}$$

Similarly, substituting 6 instead of 5 and 7 instead of 6, we get:

$$\mathbf{P}(M_{10} \geq 6) = \mathbf{P}(S_{10} \geq 6) + \mathbf{P}(S_{10} \geq 7). \tag{8}$$

Subtract (8) from (7) and get:

$$\mathbf{P}(M_{10} = 5) = \mathbf{P}(S_{10} = 5) + \mathbf{P}(S_{10} = 6). \tag{9}$$

But S_{10} can take only even values, therefore the first term in the right-hand side of (9) is zero, and (9) takes the form

$$\mathbf{P}(M_{10} = 5) = \mathbf{P}(S_{10} = 6).$$

This we can find, as explained earlier: If the path with $S_{10} = 6$ takes a steps up and b steps down, then $a + b = 10$ and $a - b = 6$, so $a = 8$ and $b = 2$. Each given path has probability 2^{-10}. Thus

$$\mathbf{P}(S_{10} = 6) = 2^{-10}\binom{10}{2} = \frac{10 \cdot 9}{2 \cdot 1024} = 0.0439.$$

12.8. **Financial modeling.** Random walk is used for elementary modeling of movements of the stock. Assume P_n is the price of a stock at day n. We can model it as a *geometric random walk:* $P_n = P_0 Z_1 \dots Z_n$, where $Z_1, \dots, Z_n$ are i.i.d. positive random variables. Take a random walk $S = (S_n)_{n\geq 0}$ starting from $S_0 = x$, with probabilities p and q of moving up and down. Then we let

$$P_n = \exp(\sigma S_n + bn), \ n = 0, 1, 2, \dots$$

for some coefficients $\sigma > 0$ and b. The coefficient σ is called *volatility*, since it is responsible for random fluctuations. The coefficient b is called *drift*. We can alternatively write this as follows:

$$P_n = \exp\left(\sigma \sum_{k=1}^{n} X_k + bn\right) = \exp\left(\sum_{k=1}^{n} Y_k\right),$$

where $Y_k = \sigma X_k + b$. All $Y_1, Y_2, \dots$ are i.i.d. and have distribution

$$\mathbf{P}(Y_k = b + \sigma) = p, \ \mathbf{P}(Y_k = b - \sigma) = q.$$

Example 12.2. For $p = q = 0.5$ (symmetric random walk), drift $b = -2$ and volatility $\sigma = 3$, find $\mathbf{E}P_{10}$ if $P_0 = 2$. We have:

$$\mathbf{E}e^{Y_k} = 0.5e^{-2+3} + 0.5e^{-2-3} = 0.5\left(e + e^{-5}\right) = 1.363.$$

Therefore,

$$\mathbf{E}P_{10} = \mathbf{E}\left[P_0 e^{Y_1} e^{Y_2} \dots e^{Y_{10}}\right] = 2\mathbf{E}\left[e^{Y_1}\right] \cdot \dots \cdot \mathbf{E}\left[e^{Y_{10}}\right] = 2 \cdot 1.363^{10} = 44.1.$$

Example 12.3. For $p = 0.75$, $q = 0.25$, $b = 1$ and $\sigma = 2$, given that $P_0 = 0.5$, what are the possible values of P_3, and what is the distribution of P_3? Each random variable Y_k takes values 3 and -1 with probabilities p and q, respectively. Therefore, the possible values of P_3 are

$$P_3 = e^{3+3+3}P_0 = \frac{e^9}{2}, \text{ with probability } p^3;$$

$$P_3 = e^{3+3-1}P_0 = \frac{e^5}{2}, \text{ with probability } 3p^2q;$$

$$P_3 = e^{3-1-1}P_0 = \frac{e}{2}, \text{ with probability } 3pq^2;$$

$$P_3 = e^{-1-1-1}P_0 = \frac{e^{-3}}{2}, \text{ with probability } q^3.$$

12.9. **Options and other derivatives.** Assume the price of a stock is modeled by $(P_n)_{n\geq 0}$, as above. A *European option call* is the right to buy this stock at a given price K, called the *strike*, at some future time T, called the *maturity*. A *European option put* is the right to sell this stock at the price K at time T. Let us find the values of these options.

If the stock price P_T at maturity is less than K, then it makes no sense to exercise the European option call: We can just buy the stock on the free market. Therefore, the price of the option is zero. However, the European option put is valuable, and its value is $K - P_T$. For the case $P_T > K$, the situation is reversed: The value of the European option call is $P_T - K$, and the European option put has value 0. Thus, if we denote $a_+ := \max(a, 0)$ for every real number a, then the value of the European option call is $(P_T - K)_+$, and the value of the European option put is $(K - P_T)_+$.

An *American option call/put* can be exercised at *any* time until maturity. Therefore, such option is generally more valuable than the corresponding European option. We shall discuss valuation of the American options later.

European options are particular cases of *European derivatives*, which are worth $f(P_T)$, where $f(x)$ is a certain function, and T is the maturity. For example, $f(x) = 1(x \geq a)$ is the *binary option:* pay 1 if the price at maturity P_T exceeded a, and pay nothing otherwise. A *forward contract* corresponds to $f(x) = x - K$, which is the *obligation* to buy the stock at time T for the price K. Such contracts are used by farmers and ranchers, who want to lock in good prices (to protect themselves from downside risk), and international companies (airlines, oil, steel), who want to lock in a certain exchange rate (say, be able to buy a euro next year using the current exchange rate, rather than the future exchange rate).

12.10. **Hedging a derivative.** This means buying a certain amount of this stock to be able to *replicate this derivative:* exactly match the value of this derivative at maturity. Let us illustrate this using a *one-step model:* $T = 1$, and

$$P_0 = 1, \quad P_1 = \begin{cases} 1.2 \text{ with probability } p = 40\%; \\ 0.6 \text{ with probability } q = 60\%. \end{cases}$$

We sell a European option call with maturity $T = 1$ and strike $K = 0.9$. Assume v is its current *fair value.* Then we need to use this money v to buy s stocks and c cash, and hedge the option. At time $t = 0$, the stock price is 1; therefore,

$$v = s + c. \tag{10}$$

Next, at time $t = 1$, if the stock price is 1.2, then we need that our portfolio (consisting of cash c and s shares) has value $1.2s + c$, equal to our obligation (which is the option value):

$$1.2s + c = (1.2 - 0.9)_+ = 0.3. \tag{11}$$

Similarly, if the stock price is 0.6, then we need:

$$0.6s + c = (0.6 - 0.9)_+ = 0. \tag{12}$$

Solve these two equations: $s = 0.5$, $c = -0.3$. Thus, we need to borrow 0.3 in cash and buy 0.6 shares of stock. The fair value is $v = -0.3 + 0.5 = 0.2$. Note that this is *not* an expected value $\mathbf{E}(P_1 - K)_+$, which is equal to $0.3 \cdot 40\% + 0 \cdot 60\% = 0.12$. Rather, this is an expected value $\mathbf{E}_0(P_1 - K)_+$ with respect to the *risk-neutral probability* p_0, q_0, such that $\mathbf{E}_0 P_1 = P_0$. Let us find them:

$$\begin{cases} p_0 \cdot 1.2 + q_0 \cdot 0.6 = 1 \\ p_0 + q_0 = 1 \end{cases} \Rightarrow p_0 = \frac{2}{3}, \ q_0 = \frac{1}{3} \ \Rightarrow \ \mathbf{E}_0(P_1 - K)_+ = \frac{2}{3} \cdot 0.3 + \frac{1}{3} \cdot 0 = 0.2.$$

Indeed, if we wish to get (10) as follows: multiply (11) by p_0 and (12) by q_0, and sum them; then to make the coefficient at s be equal to 1, we get: $p_0 \cdot 1.2 + q_0 \cdot 0.6 = 1$, which is the same as $\mathbf{E}_0 P_1 = P_0$. Then $v = p_0 \cdot 0.3 + q_0 \cdot 0 = \mathbf{E}_0(P_1 - K)_+$. The same can be done with any other derivative with maturity $T = 1$. Actually, the same result is true for any maturity.

This works if P_1 has only two values. If it has three or more values, or has continuous distribution (such as normal), then the system of equations in place of (11) and (12) has three or more equations, but two variables, and it does not have a solution. Therefore, not all derivatives can be replicated. Such market is called *incomplete.* There is more than one risk-neutral probabilities, because there are two equations for them, but more than two variables.

For more than one step: $T \geq 2$, we hedge for every step, starting from the end: Assume $P_n = P_0 Z_1 \dots Z_n$, where Z_i are i.i.d. with values 2 and 0.5 with some (irrelevant) probabilities. Let $P_0 = 1$. Consider the European option call with maturity $T = 2$ and strike $K = .5$. We have three cases:

$$\begin{cases} Z_1 = Z_2 = 2: \text{ then } \quad P_2 = 4, \ (P_2 - K)_+ = 3.5; \\ Z_1 = 2, \ Z_2 = 0.5 \text{ or } Z_1 = 0.5, \ Z_2 = 2: \text{ then } \quad P_2 = 1, \ (P_2 - K)_+ = 0.5; \\ Z_1 = Z_2 = 0.5: P_2 = 0.25, \text{ then } \quad (P_2 - K)_+ = 0. \end{cases}$$

Assume we are at $P_1 = 0.5$, then hedge:

$$c + 0.5s = v, \ c + s = 0.5, \ c + 0.25s = 0 \ \Rightarrow \ v = \frac{0.5}{3}.$$

The risk-neutral probability in this case is $1/3$ and $2/3$ corresponding to the jumps up and down, because

$$0.5 = \frac{1}{3} \cdot 1 + \frac{2}{3} \cdot 0.25.$$

Similarly, if $P_1 = 2$, then

$$c + 2s = v, \ c + s = 0.5, \ c + 4s = 3.5 \ \Rightarrow \ v = \frac{1}{3} \cdot 3.5 + \frac{2}{3} \cdot 0.5 = 1.5.$$

We have fair values corresponding to $P_1 = 0.5$ ($V_1 = 1/6$) and $P_2 = 2$ ($V_2 = 1.5$), then we can hedge on the first step, with $P_0 = 1$:

$$c + s = v, \ c + 0.5s = \frac{1}{6}, \ c + 2s = 1.5.$$

At every step, we take the expectation with respect to risk-neutral probability, when $\mathbf{E}Z_1 = \mathbf{E}Z_2 = 1$, and therefore $\mathbf{E}(P_2 \mid P_1) = \mathbf{E}(P_1 Z_2 \mid P_1) = P_1 \mathbf{E} Z_2 = P_1$, and $\mathbf{E}(P_1 \mid P_0) = \mathbf{E} P_1 = \mathbf{E} Z_1 = 1$. We say that under this measure, the process $(P_n)_{n \geq 0}$ is a *martingale* (see Section 13).

There is a general result (see more details in courses on stochastic finance): If there exists a unique risk-neutral probability, then the market is complete: Every European derivative is replicable, and its fair price is the expectation with respect to this risk-neutral probability.

13. Discrete-Time Martingales

13.1. **Definitions.** A process $X = (X_n)_{n\geq 0}$ is called a *martingale* if for each $n = 0, 1, 2, \ldots$, we have:

$$\mathbf{E}(X_{n+1} \mid X_0, \ldots, X_n) = X_n. \tag{13}$$

That is, if the best prediction of the next value, given all history, is the current value. If we have

$$\mathbf{E}(X_{n+1} \mid X_0, \ldots, X_n) \geq X_n, \tag{14}$$

then the process is called a *submartingale*. If we have:

$$\mathbf{E}(X_{n+1} \mid X_0, \ldots, X_n) \leq X_n,$$

then the process is called a *supermartingale*.

Example 13.1. Take independent random variables $Z_1, Z_2, \ldots$. Let

$$X_0 := 0, \quad \text{and} \quad X_n := Z_1 + \ldots + Z_n \quad \text{for} \quad n = 1, 2, \ldots$$

Then $X_{n+1} = X_n + Z_{n+1}$. Therefore,

$$\mathbf{E}(X_{n+1} \mid X_0, \ldots, X_n) = \mathbf{E}(X_n \mid X_0, \ldots, X_n) + \mathbf{E}(Z_{n+1} \mid X_0, \ldots, X_n) = X_n + \mathbf{E}Z_{n+1}.$$

because Z_{n+1} is independent of $X_0, \ldots, X_n$. Thus, $(X_n)_{n\geq 0}$ is a martingale if and only if all $\mathbf{E}Z_1 = \mathbf{E}Z_2 = \ldots = 0$. It is a submartingale if all $\mathbf{E}Z_k \geq 0$, and a supermartingale if all $\mathbf{E}Z_k \leq 0$.

Example 13.2. Again, take independent random variables $Z_1, Z_2, \ldots$. Let $Y_n := e^{X_n}$. Then $Y_{n+1} = Y_n e^{Z_{n+1}}$. Therefore,

$$\mathbf{E}(Y_{n+1} \mid Y_0, \ldots, Y_n) = \mathbf{E}(Y_n e^{Z_{n+1}} \mid Y_0, \ldots, Y_n) = Y_n \mathbf{E}(e^{Z_{n+1}} \mid Y_0, \ldots, Y_n) = Y_n \mathbf{E}\left[e^{Z_{n+1}}\right],$$

because Z_{n+1} is independent of $Y_0, \ldots, Y_n$. Thus, $(X_n)_{n\geq 0}$ is a martingale if and only if all $\mathbf{E}e^{Z_1} = \mathbf{E}e^{Z_2} = \ldots = 1$. It is a submartingale if all $\mathbf{E}e^{Z_k} \geq 1$, and a supermartingale if all $\mathbf{E}e^{Z_k} \leq 1$.

For a submartingale $X = (X_n)_{n\geq 0}$, taking expectations in (14), we get: $\mathbf{E}(X_{n+1}) \geq \mathbf{E}X_n$. Therefore,

$$\mathbf{E}X_0 \leq \mathbf{E}X_1 \leq \mathbf{E}X_2 \leq \ldots$$

For a martingale, we have the equality. Our goal now is to show this is true for random times τ instead of fixed times n.

13.2. **Stopping times.** A random variable τ which takes values $0, 1, 2, \ldots$ is called a *stopping time* if the event $\{\tau = n\}$ depends only on $X_0, \ldots, X_n$. If X_n is the price of your stock on day n, then τ can be the moment when you decide to sell your stock. But you can decide whether to do this right now (on day n) or not based only on the current information, that is, on $X_0, \ldots, X_n$; not on X_{n+1}, for example.

Example 13.3. The moment $\tau := \min\{n : X_n > 1\}$ is a stopping time, because $\{\tau = n\}$ means that $X_0 \leq 1, X_1 \leq 1, \ldots, X_{n-1} \leq 1, X_n > 1$. So $\{\tau = n\}$ depends only on $X_0, \ldots, X_n$.

Example 13.4. Similarly, $\tau = \min\{n : a \leq X_n \leq b\}$ is a stopping time, for fixed a and b. More generally, for every subset $D \subseteq \mathbb{R}$, $\tau = \min\{n : X_n \in D\}$ is a stopping time.

Example 13.5. However, $\tau = \max\{n : X_n > 1\}$ is not a stopping time. Indeed, consider the event $\{\tau = 1\}$. This means that $X_2 \leq 1$, so it depends on the future values of X.

13.3. **Optional stopping theorem.** Consider a stopping time τ whch is bounded from above by some constant N: $\tau \le N$. If $X = (X_n)_{n\ge 0}$ is a martingale, then

$$\mathbf{E}X_\tau = \mathbf{E}X_N = \mathbf{E}X_0. \tag{15}$$

Indeed, τ can take values $0, 1, \dots, N-1, N$. Therefore,

$$\mathbf{E}X_N = \sum_{n=0}^{N} \mathbf{E}\left[X_N 1_{\{\tau=n\}}\right], \text{ and } \mathbf{E}X_\tau = \sum_{n=0}^{N} \mathbf{E}\left[X_\tau 1_{\{\tau=n\}}\right]. \tag{16}$$

But $\{\tau = n\}$ depends only on $X_0, \dots, X_n$, and $\mathbf{E}[X_N \mid X_0, \dots, X_n] = X_n$. Therefore

$$\mathbf{E}\left[X_N 1_{\{\tau=n\}}\right] = \mathbf{E}\left[\mathbf{E}\left[X_N 1_{\{\tau=n\}} \mid X_0, \dots, X_n\right]\right] = \mathbf{E}\left[1_{\{\tau=n\}}\mathbf{E}\left[X_N \mid X_0, \dots, X_n\right]\right] = \mathbf{E}\left[1_{\{\tau=n\}}X_n\right]. \tag{17}$$

Summing (17) over $n = 0, \dots, N$, and using (16), we complete the proof of (15).

For a submartingale, we have $\mathbf{E}X_0 \le \mathbf{E}X_\tau \le \mathbf{E}X_N$, and for a supermartingale, these inequalities are reversed. Assume X is the price of a stock. If it is a martingale, then the optional stopping theorem means the following: Suppose you need to sell the stock by day N. Then you cannot devise a strategy (a rule) which will make you more (expected) profit than if you simply sell the stock at the initial day. No matter how you observe and analyze the behavior of this stock, you will not gain extra profit. However, if the price is a submartingale, then you should hold it and sell it at the last day. If the price is a supermartingale, you should sell it immediately.

The condition that τ is bounded is important. Consider the symmetric random walk from Section 12:

$$X_n = Z_1 + \ldots + Z_n, \ \mathbf{P}(Z_n = \pm 1) = \frac{1}{2}.$$

Since $\mathbf{E}Z_n = 1 \cdot \frac{1}{2} + (-1) \cdot \frac{1}{2} = 0$, the process $X = (X_n)_{n\ge 0}$ is a martingale. Let $\tau = \min\{n : X_n = 1\}$. As noted in Section 12, this stopping time is well defined, since this random walk will eventually hit level 1. But this hitting can happen very late; this stopping time τ is unbounded. By construction, $X_\tau = 1$, so $\mathbf{E}X_\tau = 1 \ne 0 = \mathbf{E}X_0$. If you are an investor trying to sell the stock with price X_n (forgetting for a minute about the fact that stock prices cannot go negative), then your strategy τ is to sell the stock when it hits level 1. But until then, your investment might go far into the negative terrirory. You will need to borrow potentially unlimited amount of money, and no lender will agree to do this.

13.4. **Jensen's inequality.** Recall that $g : \mathbb{R} \to \mathbb{R}$ is a *convex function* if

$$g(\lambda x + (1-\lambda)y) \le \lambda g(x) + (1-\lambda)g(y).$$

For example, $g(x) = x$ and $g(x) = x^2$ are convex functions, while $g(x) = \sin x$ is not. Equivalently, if you connect any two points on the graph of g by a segment, then it lies above the graph. For a twice differentiable function g, it is convex if and only if its second derivative is nonnegative: $g''(x) \ge 0$ for all x. Jensen's inequality says that if g is a convex function, then

$$\mathbf{E}g(Z) \ge g(\mathbf{E}Z). \tag{18}$$

Indeed, let $m = \mathbf{E}Z$. Since g is convex, there exists a real-valued a such that for all real x we have: $g(x) - g(m) \ge a(x-m)$. (The graph of g lies above the tangent line at point $x = m$.) Plugging in $x = Z$, we have: $g(Z) - g(m) \ge a(Z-m)$. Take the expectations:

$$\mathbf{E}g(Z) - g(m) \ge a\mathbf{E}(Z-m) = 0.$$

Therefore,

$$\mathbf{E}g(Z) \ge g(m) = g(\mathbf{E}Z).$$

One example of this is a well-known fact that $\mathbf{E}Z^2 \ge (\mathbf{E}Z)^2$. This is true, and $\mathbf{E}Z^2 - (\mathbf{E}Z)^2 = \operatorname{Var} Z \ge 0$. This immediately follows from (18): jsut apply $g(x) = x^2$.

Similarly, we can show (18) for conditional expectation instead of the unconditional:

$$\mathbf{E}\left[g(Z) \mid Y_1, \dots, Y_n\right] \ge g\left(\mathbf{E}\left[Z \mid Y_1, \dots, Y_n\right]\right). \tag{19}$$

13.5. **Preservation of the martingale property.** Take a martingale $X = (X_n)_{n\geq 0}$. Apply a convex function g to (13). By Jensen's inequality (19), we have:

$$\mathbf{E}\left[g(X_{n+1}) \mid X_0, \ldots, X_n\right] \geq g\left(\mathbf{E}\left[X_{n+1} \mid X_0, \ldots, X_n\right]\right) = g(X_n).$$

Therefore, $g(X) = (g(X_n))_{n\geq 0}$ is a submartingale.

Let us apply this to option pricing. We already discussed European options and other European derivatives in Section 12. Recall that a *call option* is the right to buy a stock at a certain *strike price* K. A European call option has *maturity time* N, when you can *exercise* this option: demand to buy this stock at price K. If the market price S_N of this stock at time N is less than K, then you can just buy the stock at the market price and forget about your option. Then your option does not have value. However, if the market price $S_N \geq K$, then you should exercise this option, and its value is $S_N - K$. In general, the value of this option is $\max(S_N - K, 0) = g(S_N)$, where $g(x) = \max(x - K, 0)$.

An American call option is different from a European one in the following way: the former can be exercised at any time until maturity N, while the latter must be exercised at maturity. Therefore, let τ be the time you decide to exercise your American call option to get the best expected value $\mathbf{E}g\left(S_\tau\right)$. When is the best exercise time τ? This is a stopping time, since your decision to exercise at time n or not is based only on your observations until time n, but not on future observations. But the function g is convex (draw a graph and check). Frequently, the stock price X is modeled by a martingale. Then $g(S) = (g(S_n))_{n\geq 0}$ is a submartingale. By the optional stopping theorem,

$$\mathbf{E}g\left(S_\tau\right) \leq \mathbf{E}g\left(S_N\right).$$

Therefore, the best time to exercise your American call option is at maturity $n = N$. In fact, American and European call options are of the same value in this case. Additional freedom to choose exercise time does not give you anything.

13.6. **Doob's inequalities.** These are generalizations of Markov's and Chebyshev's inequalities. Take a nonnegative submartingale $X = (X_n)_{n\geq 0}$ and a number $\lambda > 0$. Then

$$\mathbf{P}\left(\max_{k=0,\ldots,n} X_k \geq \lambda\right) \leq \frac{\mathbf{E}X_n}{\lambda}.$$

Indeed, take the stopping time $\tau := \min\{k = 0, \ldots, n : X_k \geq \lambda\}$, with $\tau = n$ if $X_0 < \lambda, \ldots, X_n < \lambda$. If the event

$$A = \left\{\max_{k=0,\ldots,n} X_k \geq \lambda\right\}$$

has happened, then $X_\tau \geq \lambda$, and

$$\lambda\mathbf{P}(A) \leq \mathbf{E}\left(X_\tau 1_A\right).$$

But X is nonnegative, so

$$\mathbf{E}\left(X_\tau 1_A\right) \leq \mathbf{E}X_\tau 1_A + \mathbf{E}X_\tau 1_{A^c} = \mathbf{E}X_\tau.$$

Because of the optional stopping theorem, $\mathbf{E}X_\tau \leq \mathbf{E}X_n$. Comparing this, we get:

$$\lambda\mathbf{P}(A) \leq \mathbf{E}\left(X_\tau 1_A\right) \leq \mathbf{E}X_\tau \leq \mathbf{E}X_n.$$

It suffices to divide by λ. A corollary: for a nonnegative submartingale or a martingale $X = (X_n)_{n\geq 0}$, if $p \geq 1$, then

$$\mathbf{P}\left(\max_{k=0,\ldots,n} X_k \geq \lambda\right) \leq \frac{\mathbf{E}|X_n|^p}{\lambda^p}.$$

Indeed, the function $x \mapsto |x|^p$ is convex for $p \geq 1$, and for $x \geq 0$ it is equal to x^p, so it is nondecreasing.

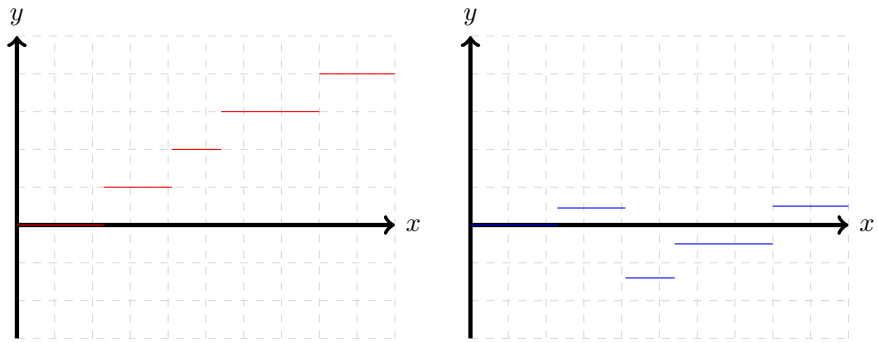
Letting $p = 2$ and recalling the very first example of a martingale: $X_n = Z_1 + \ldots + Z_n$ for independent $Z_1, \ldots, Z_n$, we have the following *Kolmogorov's inequality*:

$$\mathbf{P}\left(\max\left(X_0, \ldots, X_n\right) \geq \lambda\right) \leq \frac{1}{\lambda^2}\mathbf{E}X_n^2 = \frac{1}{\lambda^2}\sum_{k=1}^{N}\mathbf{E}Z_k^2.$$

Example 13.6. Simple random walk S_n starting from $S_0 = 0$ is a martingale. Estimate by Kolmogorov's inequality:

$$\mathbf{P}\left[\max_{0\leq n\leq 10} |S_n| \geq 5\right] \leq \frac{\mathbf{E}S_{10}^2}{5^2}. \tag{20}$$

Since $\mathbf{E}S_{10} = 0$, $\mathbf{E}S_{10}^2 = \operatorname{Var} S_{10} = \operatorname{Var}(X_1 + \ldots + X_{10}) = \operatorname{Var} X_1 + \ldots + \operatorname{Var} X_{10} = 1 + \ldots + 1 = 10$, where $X_i = \pm 1$ with equal probabilities and are independent. Thus the right-hand side in (20) is $10/25 = \boxed{0.4}$



Left picture: Poisson process with intensity $\lambda = .5$. Right picture: Compound Poisson process with the same Poisson process (and the same trajectory) as on the left picture, and with increments distributed as $\mathcal{N}(0,1)$

Example 13.7. Now, consider $M_n = e^{S_n - cn}$. Need to find n such that this is a martingale:

$$\mathbf{E}e^{X_1 - c} = 1 \;\Rightarrow\; \frac{1}{2}\left(e^1 + e^{-1}\right) = e^c \;\Rightarrow\; c = \ln\left(e^1 + e^{-1}\right) - \ln 2.$$

Apply Doob's martingale inequality with $f(x) = x^3$. For $x > 0$, this function is convex, since $f''(x) = 6x > 0$; and only this matters, because $M_n > 0$ always. Thus for $\lambda > 0$ we have:

$$\mathbf{P}\left[\max_{0 \le n \le 100} M_n \ge \lambda\right] \le \frac{\mathbf{E}\left[M_{100}^3\right]}{\lambda^3}. \tag{21}$$

Now we calculate the right-hand side of (21):

$$M_{100}^3 = e^{3S_{100}} e^{-300c} = e^{3X_1} e^{3X_2} \dots e^{3X_{100}} e^{-300c}. \tag{22}$$

All terms in the right-hand side of (22) are independent, so we can calculate the expectation:

$$\mathbf{E}\left[M_3^{100}\right] = e^{-300c}\left(\mathbf{E}\left[e^{3X_1}\right]\right)^{100} = e^{-300c}\left(\frac{1}{2}\left[e^3 + e^{-3}\right]\right)^{100}.$$

14. Poisson Process

14.1. **Poisson process.** Fix a positive real number λ. A collection $N = (N(t),\, t \ge 0)$, of random variables with values in $\mathbb{Z}_+ := \{0, 1, 2, \dots\}$ is called a *Poisson process* if: $N(0) = 0$, and for every $0 \le s < t$, $N(t) - N(s)$ is distributed as a Poisson random variable $\mathrm{Poi}(\lambda(t-s))$ and is independent of $N(u)$, $u \le s$. In particular,

$$N(t) \sim \mathrm{Poi}(\lambda t);\; \mathbf{P}(N(t) = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \quad k = 0, 1, 2, \dots$$

This process is used, among other applications, to model arrival of insurance claims, and a stock price. Since a Poisson random variable with parameter μ has mean and variance μ, we have: $\mathbf{E}N(t) = \mathrm{Var}\, N(t) = \lambda t$.

Example 14.1. Take $\lambda = 3$. What is the probability that $N(5) = 3$, given $N(2) = 1$, for the same Poisson process?

$$\mathbf{P}(N(5) = 3 \mid N(2) = 1) = \mathbf{P}(N(5) - N(2) = 2 \mid N(2) = 1) = \mathbf{P}(N(5) - N(2) = 2)$$
$$= \mathbf{P}(\mathrm{Poi}(3 \cdot 3) = 2) = \frac{9^2}{2!} e^{-9} = \frac{81}{2} e^{-9}.$$

$$\mathbf{P}(N(5) = 3,\, N(2) = 1) = \mathbf{P}(N(5) = 3 \mid N(2) = 1) \cdot \mathbf{P}(N(2) = 1) = \frac{81}{2} e^{-9} \cdot \frac{6^1}{1!} e^{-6} = \boxed{243 e^{-15}}$$

Example 14.2. Take $\lambda = 3$. Then $N(2) \sim \mathrm{Poi}(6)$, therefore $\mathbf{E}N(2) = 6$ and $\mathrm{Var}\, N(2) = 6$. Thus

$$\mathbf{E}N^2(2) = \mathrm{Var}\, N(2) + (\mathbf{E}N(2))^2 = 6 + 6^2 = \boxed{42}$$

Also, by independence of $N(2)$ and $N(5) - N(2)$, we have:

$$\mathbf{E}\left[N(2)N(5)\right] = \mathbf{E}N^2(2) + \mathbf{E}\left[N(2)(N(5) - N(2))\right] = 42 + \mathbf{E}N(2) \cdot \mathbf{E}\left[N(5) - N(2)\right] = 42 + 2\lambda \cdot (5-2)\lambda = \boxed{96}$$

14.2. **Jump times of Poisson process.** Note that Poisson random variable is nonnegative, and so for $s \le t$, we have: $N(t) - N(s) \ge 0$. Therefore, the process N is nondecreasing: $N(t) \ge N(s)$. It starts from 0, then jumps to 1 after waiting an exponential time $\tau_1 \sim \text{Exp}(\lambda)$. Why? Because the probability that this process has not yet jumped by time t is equal to the probability that $N(t) = 0$. Letting $k = 0$ into the formula above, we get: $\mathbf{P}(N(t) = 0) = e^{-\lambda t}$. Therefore,

$$\mathbf{P}(\tau_1 > t) = e^{-\lambda t}.$$

This means that $\tau_1 \sim \text{Exp}(\lambda)$. This random variable has cumulative distribution function and density

$$F(t) = \mathbf{P}(\tau_1 \le t) = 1 - e^{-\lambda t},\ p(t) = F'(t) = \lambda e^{-\lambda t},\ t > 0.$$

This random variable has $\mathbf{E}\tau_1 = \lambda^{-1}$ and $\operatorname{Var}\tau_1 = \lambda^{-2}$.

After jumping to τ_1, the process waits some random time and jumps from 1 to 2. If τ_2 is the time when it jumps from 1 to 2, then $\tau_2 - \tau_1$ is the time which it waits between its first and second jumps. It turns out that $\tau_2 - \tau_1$ is independent from τ_1 and is also distributed as $\text{Exp}(\lambda)$. The reason for this is when the process N jumps to 1, it "forgets the past" and behaves as if instead of 1 it was at 0 and nothing yet has happened. This follows from the property above: that $N(t) - N(s)$ for $t > s$ is independent of $N(u)$ for $u \le s$. This important property is called the *Markov property*, after a Russian mathematician Andrey Markov.

Similarly, if τ_k is the time of jump from $k-1$ to k, then $\tau_k - \tau_{k-1}, \tau_{k-1} - \tau_{k-2}, \dots, \tau_2 - \tau_1, \tau_1$ are i.i.d. $\text{Exp}(\lambda)$ random variables. It is possible to show that this process does not jump, say, from 0 to 2: all its jumps have magnitude 1.

Let us now find the distribution of τ_2. This is a sum of two independent random variables τ_1 and $\tau_2 - \tau_1$, both with density $p_1(t) = \lambda e^{-\lambda t}$ for $t > 0$. As mentioned in subsection 5.8, the density $p_2(x)$ of τ_2 is then the *convolution* of these exponential densities:

$$p_2(x) = \int_0^x p_1(y)p_1(x-y)\,\mathrm{d}y = \lambda^2 x e^{-\lambda x}.$$

It was calculated in subsection 5.9. More generally, the density p_k of τ_k is given by

$$p_k(x) = \frac{\lambda^k}{(k-1)!}x^{k-1}e^{-\lambda x},\ x > 0.$$

This is called the *Gamma distribution* $\Gamma(k, \lambda)$ with parameters k and λ. It has expectation and variance (because of independence):

$$\mathbf{E}\tau_k = \mathbf{E}\tau_1 + \mathbf{E}(\tau_2 - \tau_1) + \ldots + \mathbf{E}(\tau_k - \tau_{k-1}) = k\lambda^{-1},$$
$$\operatorname{Var}\tau_k = \operatorname{Var}\tau_1 + \operatorname{Var}(\tau_2 - \tau_1) + \ldots + \operatorname{Var}(\tau_k - \tau_{k-1}) = k\lambda^{-2}.$$

Example 14.3. What is the probability that the Poisson process with $\lambda = 3$, by time $t = 4$, has jumped two or more times?

$$\mathbf{P}(\tau_2 \le 4) = \mathbf{P}(N(4) \ge 2) = 1 - \mathbf{P}(N(4) = 0) - \mathbf{P}(N(4) = 1).$$

Since $N(4) \sim \text{Poi}(12)$, this becomes

$$1 - \frac{12^0}{0!}e^{-12} - \frac{12^1}{1!}e^{-12} = \boxed{1 - 13e^{-12}}$$

14.3. **Compound Poisson process.** Now, we modify the Poisson process to let it jump not only by 1 upward, but in a more general way. Define a sequence of i.i.d. (independent identically distributed) random variables $Z_1, Z_2, \dots$, independent of $N(t)$, $t \ge 0$). Then a *compound Poisson process* is defined as

$$X(t) = \sum_{k=1}^{N(t)} Z_k.$$

It starts from $X(0) = 0$, then waits time τ_1 and jumps to $X(\tau_1) = Z_1$. Next, it waits time $\tau_2 - \tau_1$ and jumps to $X(\tau_2) = Z_1 + Z_2$, then waits time $\tau_3 - \tau_2$ and jumps to $X(\tau_3) = Z_1 + Z_2 + Z_3$, etc.

This process also has the property that $X(t) - X(s)$ is independent of $X(u)$, $u \le s$ (Markov property), but it is distributed usually not as a Poisson random variable. The distribution of $X(t) - X(s)$ is the same as the distribution of $X(t-s)$, because they are sums of $N(t) - N(s)$ and, respectively, $N(t-s)$i.i.d. random variables $Z_1, Z_2, \dots$ but $N(t) - N(s)$ and $N(t-s)$ have the same distribution.

To find the distribution of $X(t)$, let us recall the theory of Section 8: random sum of random numbers. The generating function $\varphi_{N(t)}(s)$ of $N(t) \sim \text{Poi}(\lambda t)$ is given by

$$\varphi_{N(t)}(s) = \mathbf{E}s^{N(t)} = e^{\lambda t(s-1)}.$$

Assume $F_Z(y) = \mathbf{E}e^{yZ_i}$ is the moment generating function of each Z_i; it is independent of i, since $Z_1, Z_2, \dots$ have the same distribution. Then by results of Section 8 (random sum of random variables), we have: the moment generating function of $X(t)$ is equal to

$$G_t(y) := \mathbf{E}e^{yX(t)} = \varphi_{N(t)}(F_Z(y)) = \exp\left(\lambda t(F_Z(y) - 1)\right).$$

We can also find expectation and variance, using the formulas from Section 8. Assume $\mathbf{E}Z_k = \mu$, and $\operatorname{Var} Z_k = \sigma^2$. Because $\mathbf{E}N(t) = \operatorname{Var} N(t) = \lambda t$,

$$\mathbf{E}X(t) = \mathbf{E}N(t) \cdot \mathbf{E}Z_k = \lambda\mu t, \tag{23}$$

$$\operatorname{Var} X(t) = \mathbf{E}N(t) \cdot \operatorname{Var} Z_k + \operatorname{Var} N(t) \cdot (\mathbf{E}Z_k)^2 = \lambda(\sigma^2 + \mu^2)t. \tag{24}$$

14.4. **Sampling from a Poisson process.** One can consider the Poisson process $N = (N(t),\, t \geq 0)$ with intensity λ, as the process of arrival of customers with interarrival times independent $\operatorname{Exp}(\lambda)$. Now assume that "good" customers arrive with probability p; that is, each customer is "good" with probability p, independently of other customers and of the Poisson process. Then the number of "good" customers arrived by time t is

$$M(t) = \sum_{k=1}^{N(t)} Z_k, \quad Z_k = \begin{cases} 1, & \text{if } k\text{th customer good;} \\ 0, & \text{else.} \end{cases}$$

This is a compound Poisson process with $M(t)$ having moment generating function

$$G_t(y) = \mathbf{E}e^{yM(t)} = \exp\left(\lambda t(F_Z(y) - 1)\right), \quad F_Z(y) := \mathbf{E}e^{Z_k y} = pe^y + (1 - p).$$

Then we get:

$$G_t(y) = \exp\left(\lambda pt(F_Z(y) - 1)\right).$$

Therefore, $M = (M(t),\, t \geq 0)$ is itself a Poisson process, but with intensity λp.

14.5. **Sum of two Poisson processes.** Take two independent Poisson processes N_1 and N_2 with intensities λ_1 and λ_2. Then $N = N_1 + N_2$ is also a Poisson process with intensity $\lambda = \lambda_1 + \lambda_2$. Why? First, $N(0) = N_1(0) + N_2(0) = 0$. Next, for $t > s$ we have:

$$N(t) - N(s) = (N_1(t) - N_1(s)) + (N_2(t) - N_2(s)). \tag{25}$$

The expression in (25) is independent of $N(u) = N_1(u) + N_2(u)$ for $u \leq s$. Indeed, $N_1(t) - N_1(s)$ is independent of $N_1(u)$, because N_1 is a Poisson process; and $N_1(t) - N_1(s)$ is independent of N_2, because these two Poisson processes are independent. Therefore, $N_1(t) - N_1(s)$ is independent from $N(u)$. Same for $N_2(t) - N_2(s)$. Finally, from (25) we get: This is the sum of two independent Poisson random variables:

$$N_1(t) - N_1(s) \sim \operatorname{Poi}(\mu_1), \quad N_2(t) - N_2(s) \sim \operatorname{Poi}(\mu_2),$$

$$\mu_1 = \lambda_1(t - s),\ \mu_2 = \lambda_2(t - s).$$

Therefore, this sum is also a Poisson random variable with parameter $\mu_1 + \mu_2 = (t - s)(\lambda_1 + \lambda_2)$. Which completes the proof that N is a Poisson process with intensity $\lambda = \lambda_1 + \lambda_2$.

14.6. **Central Limit Theorem.** We can apply Central Limit Theorem to a sum of large quantity of i.i.d. random variables, even if the sum itself is random (that is, random quantity of summands). Then we get: For a compound Poisson process, if $\mathbf{E}Z_k = \mu$ and $\operatorname{Var} Z_k = \sigma^2$,

$$\frac{X(t) - \mathbf{E}X(t)}{\sqrt{\operatorname{Var} X(t)}} \Rightarrow \mathcal{N}(0, 1).$$

We can use this normal approximation to find distribution of $X(t)$ for large t, when it becomes inconvenient to use moment generating function.

Example 14.4. An insurance company receives claims as a Poisson process with $\lambda = 2$. Each claim has mean $\mu = 3$ and variance $\sigma^2 = 4$. What is the *value at risk* $\mathrm{VaR}_{95\%}$ for two years ($t = 2$) at confidence level 95%? That is, which capital does the company need to accumulate so that it can pay its obligations? From (23), (24), we have:

$$\mathbf{E}X(t) = \lambda\mu t = 12,\ \operatorname{Var} X(t) = \lambda(\mu^2 + \sigma^2)t = 52.$$

Therefore, from the Central Limit Theorem we get:

$$\frac{X(2) - 12}{\sqrt{52}} \approx \mathcal{N}(0, 1).$$

Looking at the table of normal distribution, the quantile corresponding to 95% is $x_{95\%} = 1.645$. Therefore, the following event happens with probability approximately 95%:

$$\frac{X(2) - 12}{\sqrt{52}} \le x_{95\%} \;\Leftrightarrow\; X(2) \le 12 + \sqrt{52} \cdot 1.645 = \boxed{23.86}$$

14.7. **Cramer-Lundberg model.** Assume the initial capital of an insurance company is u, the constant flow rate of premiums is c, and the claims arrive according to the Poisson process $N = (N(t),\, t \ge 0)$ and have i.i.d. sizes $Z_1, Z_2, \dots$ Then the capital of the company at time t is

$$X(t) = u + ct - \sum_{k=1}^{N(t)} Z_k.$$

The *ruin time* is the first jump when $X(t) < 0$. This can be infinite, when ruin actually never happens.

Example 14.5. Assume $u = 1$, $c = 2$, $\lambda = 2$, $Z_k \sim \mathrm{Exp}(1.5)$. Then

$$X(3) = 7 - \sum_{k=1}^{N(3)} Z_k,$$

Since $\mathbf{E}N(3) = \operatorname{Var} N(3) = 6$ and $\mathbf{E}Z_k = 1/1.5$, $\operatorname{Var} Z_k = (1/1.5)^2$, we can calculate the mean and variance:

$$\mathbf{E}X(3) = 7 - \mathbf{E}Z_k \cdot \mathbf{E}N(3) = 7 - \frac{1}{1.5} \cdot 2 \cdot 3 = 3,$$

$$\operatorname{Var} X(3) = \mathbf{E}N(3) \cdot \operatorname{Var} Z_k + \operatorname{Var} N(3) \cdot (\mathbf{E}Z_k)^2 = 6 \cdot \frac{1}{1.5^2} + 6 \cdot \frac{1}{1.5^2} = 16/3.$$

15. Continuous-Time Markov Chains

15.1. **Properties of exponential random variables.** Recall that $\tau \sim \mathrm{Exp}(\lambda)$ is an exponential random variable with intensity λ if it has density $\lambda e^{-\lambda x}$ for $x > 0$. Then it has the property $\mathbf{P}(\tau > t) = e^{-\lambda t}$. It is commonly used to model waiting time until something happens.

1. *Memoryless property*: for $s, t \ge 0$,

$$\mathbf{P}(\tau > t + s \mid \tau > s) = \mathbf{P}(\tau > t).$$

The meaning is that if we already waited time s, the remaining wait time is distributed in the same way as if we had not waited at all. In other words, how much more we need to wait is independent of how long we already waited. (This is the only distribution which has such property.)

This property follows from

$$\begin{aligned}\mathbf{P}(\tau > t + s \mid \tau > s) &= \frac{\mathbf{P}(\tau > t + s,\, \tau > s)}{\mathbf{P}(\tau > s)} = \frac{\mathbf{P}(\tau > t + s)}{\mathbf{P}(\tau > s)} \\ &= \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} = \mathbf{P}(\tau > t).\end{aligned}$$

2. If we have two independent exponential random variables

$$X_1 \sim \mathrm{Exp}(\lambda_1), \text{ and } X_2 \sim \mathrm{Exp}(\lambda_2),$$

then $\min(X_1, X_2) \sim \mathrm{Exp}(\lambda_1 + \lambda_2)$:

$$\begin{aligned}\mathbf{P}(\min(X_1, X_2) > t) &= \mathbf{P}(X_1 > t,\, X_2 > t) = \mathbf{P}(X_1 > t)\mathbf{P}(X_2 > t) \\ &= e^{-\lambda_1 t} e^{-\lambda_2 t} = e^{-(\lambda_1 + \lambda_2)t}.\end{aligned}$$

3. If we have two independent exponential random variables $X_1 \sim \mathrm{Exp}(\lambda_1)$ and $X_2 \sim \mathrm{Exp}(\lambda_2)$, then

$$\mathbf{P}(X_1 < X_2) = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

Indeed, we condition on the value of X_1:

$$\begin{aligned}\mathbf{P}(X_1 < X_2) &= \int_0^\infty \mathbf{P}(X_2 > x) \lambda_1 e^{-\lambda_1 x}\, \mathrm{d}x = \int_0^\infty e^{-\lambda_2 x} \lambda_1 e^{-\lambda_1 x}\, \mathrm{d}x \\ &= \lambda_1 e^{-(\lambda_1 + \lambda_2)x}\, \mathrm{d}x = \frac{\lambda_1}{\lambda_1 + \lambda_2}.\end{aligned}$$

15.2. **Definition and construction of continuous-time Markov chains.** A continuous-time Markov chain on the state space $\{1, 2, \dots, N\}$ is a process $X = (X(t),\, t \geq 0)$, where for every $t \geq 0$ $X(t)$ is a random variable which takes values $1, \dots, N$, such that the behavior of $X(t)$ for $t \geq s$ is determined only by the knowledge of $X(s)$; we do not need additional knowledge of $X(u)$ for $u < s$. This is called the *Markov property.*

If the process stays at, for example, 1, then it can stay there an exponential amount of time τ_1 (with some intensity λ_1) before jumping to some other state. This is because of the Markov property: how much more we need to wait at state 1 does not depend on how long we waited to far. Therefore, every state i has an associated waiting time $\mathrm{Exp}(\lambda_i)$: an exponential "clock". When this clock rings, we have to leave i. As we leave i, we have a choice of going to other states $j \neq i$, with the probability of going from i to j being p_{ij}. The parameter $\lambda_{ij} = p_{ij}\lambda_i$ is called the *intensity of move* from i to j. We usually write them in the form of the *generating matrix,* or simply a *generator:*

$$A = (a_{ij}),\ a_{ij} = \lambda_{ij},\ i \neq j; a_{ii} = -\lambda_i.$$

This matrix has the property that its off-diagonal elements are nonnegative, and the sum of each row is equal to zero.

We can construct this Markov chain alternatively: for each state i we can generate $N-1$ independent exponential clocks $\tau_{ij} \sim \mathrm{Exp}(\lambda_{ij})$, $j \neq i$. Whichever rings first (whichever is the minimal of them) we move to the corresponding state j. This construction gives us the same result as above, because by the property 2 above we have:

$$\sum_{j \neq i} \tau_{ij} \sim \mathrm{Exp}\left(\sum_{j \neq i} \lambda_{ij}\right) = \mathrm{Exp}(\lambda_i).$$

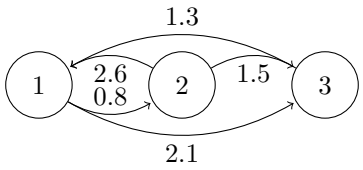
From Property 3 above, the probability that clock τ_{ij} will ring first among these clocks is

$$\mathbf{P}\left(\tau_{ij} = \min_{k \neq i} \tau_{ik}\right) = \frac{\lambda_{ij}}{\sum_{k \neq i} \lambda_{ik}} = \frac{\lambda_{ij}}{\lambda_i} = p_{ij}.$$

The parameter λ_i is called the *intensity of exit* from state i. The parameter λ_{ij} is called the *intensity of moving* from i to j.

Example 15.1. State space $\{1, 2, 3\}$, with generating matrix

$$A = \begin{bmatrix} -2.9 & 0.8 & 2.1 \\ 2.6 & -4.1 & 1.5 \\ 1.3 & 0 & -1.3 \end{bmatrix}$$



Generate two independent exponential clocks $\tau_{12} \sim \mathrm{Exp}(0.8)$, $\tau_{13} \sim \mathrm{Exp}(2.1)$. Then $\tau_1 = \min(\tau_{12}, \tau_{13}) \sim \mathrm{Exp}(2.9)$, so we will have to wait τ_1 time until leaving 1. Upon leaving 1, we will go to 2 with probability

$$\mathbf{P}(\tau_{12} < \tau_{13}) = \frac{0.8}{0.8 + 2.1} = \frac{0.8}{2.9}$$

and to 3 with probability

$$\mathbf{P}(\tau_{13} < \tau_{12}) = \frac{2.1}{0.8 + 2.1} = \frac{2.1}{2.9}.$$

Similarly with exit from 2. Exiting from 3, we will always go to 1, because the intensity of moving from 3 to 2 is 0. We can include this in our framework by letting $\tau_{32} \sim \mathrm{Exp}(0) = \infty$. (An exponential distribution $\mathrm{Exp}(\lambda)$ with intensity $\lambda = 0$ is concentrated at infinity.)

15.3. **Transition matrix.** For states i and j, let $p_{ij}(t) = \mathbf{P}(X(t) = j \mid X(0) = i)$ be the probability that the Markov chain, starting from state i, will be at time t at state j. These probabilities form the *transition matrix:* $P(t) = (p_{ij}(t))_{i,j=1,\dots,N}$.

Let the row-vector in $\mathbb{R}^N$:

$$x(t) = \begin{bmatrix} \mathbf{P}(X(t) = 1) & \mathbf{P}(X(t) = 2) & \dots & \mathbf{P}(X(t) = N) \end{bmatrix}$$

be the distribution of $X(t)$. If we start from the initial distribution $x(0) \in \mathbb{R}^N$, then the distribution $x(t)$ can be found as

$$x(t) = x(0)P(t). \tag{26}$$

Indeed, assume you want to be at state j at time t. The probability $\mathbf{P}(X(t) = j)$ of this is the jth element of the vector $x(t)$. Then you can achieve this by being at state i initially (which happens with probability $x_i(0) = \mathbf{P}(X(0) = i)$), and moving from state i to state j in time t (which happens with probability $p_{ij}(t)$). Summing over all $i = 1, \dots, N$, we have:

$$x_j(t) = \sum_{i=1}^{N} p_{ij}(t)x_i(0),$$

which is the same as (26). We can similarly show *Chapman-Kolmorogov equations:*

$$P(t+s) = P(t)P(s) \quad \text{for} \quad t, s \geq 0.$$

Note that $p_{ii}(0) = 1$ and $p_{ij}(0) = 0$ for $i \neq j$. Therefore,

$$P(0) = I_N, \tag{27}$$

which is the $N \times N$-identity matrix.

15.4. **Forward and backward Kolmorogov equations.** Taking the derivatives entrywise, we get: $P'(t) = (p'_{ij}(t))$. The following system of *Kolmogorov equations* hold true:

$$P'(t) = P(t)A \quad \text{(forward Kolmorogov equations)} \tag{28}$$

$$P'(t) = AP(t) \quad \text{(backward Kolmorogov equations)} \tag{29}$$

Indeed, how does the matrix $P(t)$ change from t to $t + \mathrm{d}t$? In other words, how can a process get from state i to state j in time $t + \mathrm{d}t$?

Case 1. It can already be at state j by time t. This happens with probability $p_{ij}(t)$. And the probability that it will stay at state j from time t to time $t + \mathrm{d}t$ is $\mathbf{P}(\tau_j > \mathrm{d}t) = e^{-\lambda_j \mathrm{d}t} \approx 1 - \lambda_j \mathrm{d}t$. This gives us the term $p_{ij}(t)(1 - \lambda_j \mathrm{d}t)$.

Case 2. It can be at a different state $k \neq j$ by time t. This happens with probability $p_{ik}(t)$. and the probability that the process will jump from k to j during time $[t, t + \mathrm{d}t]$ is $\mathbf{P}(\tau_{kj} \leq \mathrm{d}t) = 1 - e^{-\lambda_{kj}\,\mathrm{d}t} = \lambda_{kj}\,\mathrm{d}t$. This gives us the term $p_{ik}(t)\lambda_{kj}\,\mathrm{d}t$.

Summing these terms, we have:

$$p_{ij}(t + \mathrm{d}t) = p_{ij}(t)(1 - \lambda_j \mathrm{d}t) + \sum_{k \neq j} p_{ik}(t)\lambda_{kj}\,\mathrm{d}t.$$

Subtracting p_{ij}, and dividing by $\mathrm{d}t$, we get:

$$p'_{ij}(t) = \frac{p_{ij}(t + \mathrm{d}t) - p_{ij}(t)}{\mathrm{d}t} = -\lambda_j p_{ij}(t) + \sum_{k \neq j} p_{ik}(t)\lambda_{kj}(t) = \sum_{k=1}^{N} p_{ik}(t)a_{kj}.$$

This shows, by conditioning on $X(t)$, that $P'(t) = P(t)A$; this proves (28). One can also show (29) by conditioning on $X(\mathrm{d}t)$.

Finally, in light of (27) and (28), we have:

$$P'(0) = A. \tag{30}$$

15.5. **Stationary distribution.** Take a distribution π. Assume we start from there: $x(0) = \pi$. If the Markov chain remains there for each time $t \geq 0$: $x(t) = \pi$, then we have a *stationary distribution* p. In light of (26), we can rewrite this as

$$\pi = \pi P(t). \tag{31}$$

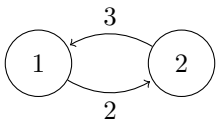
Take a derivative in (31) at $t = 0$. The left-hand side is constant, so its derivative is zero (row vector). The right-hand side by (30) has derivative πA. Therefore,

$$0 = \pi A. \tag{32}$$

A continuous-time Markov chain on the state space $\{1, \dots, N\}$ always has a stationary distribution.

Example 15.2. Take the generator

$$A = \begin{bmatrix} -2 & 2 \\ 3 & -3 \end{bmatrix}$$



Solve the system of equations for $\pi = \begin{bmatrix} \pi_1 & \pi_2 \end{bmatrix}$:

$$0 = \pi A \ \Rightarrow \ -2\pi_1 + 3\pi_2 = 0 \ \Rightarrow \ \pi_1 = \frac{3}{2}\pi_2.$$

But since this is a probability distribution, it should also satisfy

$$\pi_1 + \pi_2 = 1.$$

Therefore,

$$\boxed{\pi_1 = \frac{3}{5}, \ \pi_2 = \frac{2}{5}}$$

15.6. **Convergence.** The vector $x(t)$, which is the distribution of $X(t)$, satisfies the ODE itself: Differentiating (26), and using (28), we have:

$$x'(t) = x(0)P'(t) = x(0)P(t)A = x(t)A. \tag{33}$$

Let us solve it, if we know the initial distribution $\pi(0)$. Assume that this Markov chain has a unique stationary distribution π; in other words, it has only one eigenvector corresponding to the eigenvalue $\lambda_1 = 0$. Other eigenvalues $\lambda_2, \dots, \lambda_N$ are negative (or are complex but have negative real parts - the analysis is the same) and have eigenvectors $u_2, \dots, u_N$:

$$u_i A = \lambda_i u_i, \quad i = 2, \dots, N.$$

Then the vectors $\pi, u_2, \dots, u_N$ form a basis of $\mathbb{R}^N$. We can find coefficients $c_1(t), \dots, c_N(t)$ such that

$$x(t) = c_1(t)\pi + c_2(t)u_2 + \ldots + c_N(t)u_N. \tag{34}$$

Let us derive the ODE for each $c_i(t)$:

$$\begin{aligned} x'(t) =& c_1'(t)\pi + c_2'(t)u_2 + \ldots + c_N'(t)u_N = c_1(t)\pi A + c_2(t)u_2 A + \ldots + c_N(t)u_N A \\ &= 0 + \lambda_2 c_2(t)u_2 + \ldots + \lambda_N c_N(t)u_N. \end{aligned} \tag{35}$$

Comparing left and right-hand sides of (35), we have:

$$c_1'(t) = 0 \ \Rightarrow \ c_1(t) = c_1(0),$$

$$c_i'(t) = \lambda_i c_i(t) \ \Rightarrow \ c_i(t) = c_i(0)e^{\lambda_i t}, \ i = 2, \dots, N. \tag{36}$$

We can find $c_1(0), \dots, c_N(0)$ by decomposing the initial distribution $x(0) = c_1(0)\pi + c_2(0)u_2 + \ldots + c_N(0)u_N$ as a linear combination of the basis vectors. Then plug these initial conditions into (36). We have:

$$x(t) = c_1\pi + c_2(0)e^{\lambda_2 t}u_2 + \ldots + c_N(0)e^{\lambda_N t}u_N. \tag{37}$$

Because $\lambda_2, \dots, \lambda_N < 0$, we have: $x(t) \to c_1\pi$ as $t \to \infty$. Actually, $c_1 = 1$, because all $x(t)$ and π have the same property: sum of components equals to one. Moreover, from (37) we get:

$$|x(n) - \pi| \le c_2(0)|u_2|e^{\lambda_2 t} + \ldots + c_N(0)|u_N|e^{\lambda_N t}.$$

Each summand converges to zero with rate $e^{\lambda_i t}$. The slowest convergence is for $\lambda = \lambda_i$ with smallest absolute value (closest to zero). This gives general convergence rate $e^{\lambda t}$. To find convergence rate, take among nonzero eigenvalues the closest λ to zero, and then take $e^{\lambda t}$.

The matrix $P(t)$ can be found in the same way, because its ith row is, in fact, $x(t)$ for

$$x(0) = \begin{bmatrix} 0 & 0 & \ldots 0 & 1 & 0 & \ldots & 0 \end{bmatrix}, \tag{38}$$

where the unity is on the ith place. Indeed, $p_{ij}(t)$ is the probability of $X(t) = j$, given that $X(0) = i$; and if $X(0) = i$ with probabiltiy one, then $X(0)$ has distribution (38).

Example 15.3. A Markov chain on two states $\{1, 2\}$, with

$$A = \begin{bmatrix} -2 & 2 \\ 3 & -3 \end{bmatrix}$$

Let the initial distribution be $x(0) = \begin{bmatrix} 1 & 0 \end{bmatrix}$ Then $\pi = \begin{bmatrix} \frac{3}{5} & \frac{2}{5} \end{bmatrix}$ (found above). The eigenvalues of A: $\det(A - \lambda I_2) = 0 \Rightarrow \lambda^2 + 5\lambda = 0 \Rightarrow \lambda_{1,2} = 0, -5$. An eigenvector corresponding to $\lambda_1 = 0$ is π. An eigenvector corresponding to $\lambda_2 = -5$ can be found from

$$-5u = uA \Rightarrow \begin{cases} -5u_1 = -2u_1 + 3u_2 \\ -5u_2 = 2u_1 - 3u_2 \end{cases} \Rightarrow 3u_1 + 3u_2 = 0 \Rightarrow u = \begin{bmatrix} 1 & -1 \end{bmatrix}$$

We can decompose $x(0) = c_1\pi + c_2 u$ by comparing componentwise:

$$\begin{cases} 1 = \frac{3}{5}c_1 + c_2 \\ 0 = \frac{2}{5}c_1 - c_2 \end{cases} \Rightarrow c_1 = 1,\ c_2 = \frac{2}{5}.$$

Then

$$x(t) = \pi + \frac{2}{5}e^{-5t}u_2 \Leftrightarrow \begin{cases} x_1(t) = \frac{3}{5} + \frac{2}{5}e^{-5t} \\ x_2(t) = \frac{2}{5} - \frac{2}{5}e^{-5t} \end{cases}$$

The rate of convergence is e^{-5t}.

15.7. **Relation between discrete- and continuous-time Markov chains.** Take a continuous-time Markov chain on three states $1, 2, 3$, with generator

$$A = \begin{bmatrix} -2.9 & 0.8 & 2.1 \\ 1.3 & -3.9 & 2.6 \\ 3 & 0 & -3 \end{bmatrix}$$

Let $\tau_0 = 0 < \tau_1 < \tau_2 < \dots$ be the jump times. For example, if the Markov chain starts from 1: $X(0) = 1$, then $\tau_1 \sim \text{Exp}(2.9)$. Next, if $X(\tau_1) = 2$ (that is, if X makes its first jump from 1 to 2), then $\tau_2 - \tau_1 \sim \text{Exp}(3.9)$, etc. Let

$$Y_n := X(\tau_n),\ n = 0, 1, 2, \dots$$

Then $Y = (Y_n)_{n \geq 0} = (Y_0, Y_1, Y_2, \dots)$ is a discrete-time Markov chain, with transition matrix

$$P = \begin{bmatrix} 0 & \frac{0.8}{2.9} & \frac{2.1}{2.9} \\ \frac{1.3}{3.9} & 0 & \frac{2.6}{3.9} \\ 1 & 0 & 0 \end{bmatrix}$$

Indeed, if the continuous-time Markov chain X jumps from 1, then it goes to 2 with probability $0.8/2.9$, and to 3 with probability $2.1/2.9$.

Remark 15.1. This discrete-time Markov chain Y, by construction, cannot go in one step from a state i to the same state i. It has to jump somewhere else (although it can return to the same state in more than one step).

When we switch from a continuous-time to this discrete-time Markov chain, we lose information about *when* jumps occurred. We only know *from where to where* the process has jumped.

We can also move backward, from a discrete-time to a continuous-time Markov chain. Assume we have a discrete-time Markov chain

$$P = \begin{bmatrix} 0 & 0.7 & 0.3 \\ 0.4 & 0 & 0.6 \\ 1 & 0 & 0 \end{bmatrix}$$

It has to have zeros on the main diagonal, which corresponds to the property from Remark 15.1. Choose any intensities $\lambda_1, \lambda_2, \lambda_3 > 0$ of exiting from states $1, 2, 3$; for example,

$$\lambda_1 = 2,\ \lambda_2 = 3,\ \lambda_3 = 4.$$

We have complete freedom in choosing these intensities; by doing this, we restore the information lost when switching from continuous time to discrete time (see above). Then the corresponding continuous-time Markov chain will have generator

$$A = \begin{bmatrix} -2 & 2 \cdot 0.7 & 2 \cdot 0.3 \\ 3 \cdot 0.4 & -3 & 3 \cdot 0.6 \\ 4 & 0 & -4 \end{bmatrix}$$

There is also a relation between stationary distributions of the discrete-time and the continuous-time Markov chains: Assume

$$\pi = \begin{bmatrix} \pi_1 & \pi_2 & \pi_3 \end{bmatrix}$$

is a stationary distribution for the discrete-time Markov chain. Then we have:

$$\pi = \pi P \ \Rightarrow \ \begin{cases} \pi_1 = 0.4\pi_2 + \pi_3 \\ \pi_2 = 0.7\pi_1 \\ \pi_3 = 0.3\pi_1 + 0.6\pi_2 \end{cases}$$

Take $\rho_i = \pi_i/\lambda_i$, $i = 1, 2, 3$, and $\rho = \begin{bmatrix} \rho_1 & \rho_2 & \rho_3 \end{bmatrix}$ Then

$$\begin{cases} -2\rho_1 + 3 \cdot 0.4\rho_2 + 4\rho_3 = -\pi_1 + 0.4\pi_2 + \pi_3 = 0 \\ 2 \cdot 0.7\rho_1 - 3\rho_2 = 0.7\pi_1 - \pi_2 = 0 \\ 2 \cdot 0.3\rho_1 + 3 \cdot 0.6\rho_2 - 4\rho_3 = 0.3\pi_1 + 0.6\pi_2 - \pi_3 = 0 \end{cases} \Rightarrow \ \rho A = 0.$$

This means ρ, or, rather,

$$\rho' = \begin{bmatrix} \rho_1' & \rho_2' & \rho_3' \end{bmatrix}$$
$$\rho_i' = \frac{\rho_i}{\rho_1 + \rho_2 + \rho_3}, \ i = 1, 2, 3$$

(we divide each ρ_i to make them sum up to 1) is a stationary distribution for the continuous-time Markov chain. Conversely, if ρ is a stationary distribution for a continuous-time Markov chain, then

$$\pi = \begin{bmatrix} \lambda_1\rho_1 & \lambda_2\rho_2 & \lambda_3\rho_3 \end{bmatrix}$$

is a stationary distribution for the corresponding discrete-time Markov chain.

There is a heuristical explanation of this relation between π and ρ. If the Markov chain is ergodic, then it converges to its stationary distribution as time goes to infinity. And the long-term share of time spent at state i is equal to π_i for the discrete-time Markov chain or ρ_i' for the continuous-time Markov chain. Assume that λ_i is very large compared to other λ_j. Then, as long as the continuous-time Markov chain jumps to i, it almost immediately jumps out of i. The share of time spent at i is small. This corresponds to a small $\rho_i = \pi_i/\lambda_i$, which comes from the large λ_i.

15.8. **Reducible and irreducible Markov chains.** There is a well-developed theory of discrete-time Markov chains, see Section 11. We can apply the same theory to continuous-time Markov chains. A state i is called *recurrent* for a continuous-time Markov chain, if it is recurrent for the corresponding discrete-time Markov chain; that is, if the probability of ever returning to i if we start from i is 1. Otherwise, it is called *transient*. After removal of all transient states, if all remaining states are connected, then the Markov chain is called *irreducible*; otherwise, it is called *reducible*, and it splits into corresponding *communicating classes*. An irreducible Markov chain (both discrete and continuous time) has a unique stationary distribution, but a reducible one has infinitely many stationary distributions.

Take a continuous-time Markov chain on $\{1, 2, 3, 4\}$ with generating matrix

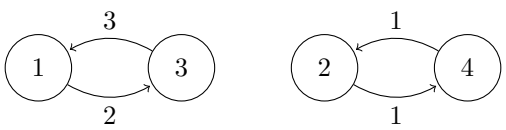
$$A = \begin{bmatrix} -2 & 0 & 2 & 0 \\ 0 & -1 & 0 & 1 \\ 3 & 0 & -3 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}$$

Then this Markov chain splits into two parts: a Markov chain with states 1 and 3, and with transition matrix

$$A_1 = \begin{bmatrix} -2 & 2 \\ 3 & -3 \end{bmatrix}$$

and a Markov chain with states 2 and 4, and with transition matrix

$$A_2 = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}$$



These two parts are not connected to each other. We can find a stationary distribution for the first part:

$$\begin{bmatrix}\pi_1 & \pi_3\end{bmatrix} A_1 = \begin{bmatrix}0 & 0\end{bmatrix} \;\Rightarrow\; -2\pi_1 + 3\pi_3 = 0 \;\Rightarrow\; \pi_1 = \frac{3}{2}\pi_3,$$

and because $\pi_1 + \pi_3 = 1$, we have: $\pi_1 = \frac{3}{5}$, $\pi_3 = \frac{2}{5}$. Similarly, the second part has the following stationary distribution:

$$\begin{bmatrix}\pi_2 & \pi_4\end{bmatrix} = \begin{bmatrix}\frac{1}{2} & \frac{1}{2}\end{bmatrix}$$

Now, let us construct a stationary distribution for the whole Markov chain. Let p_1 be the probability that we are in part $1-3$. Denote it by p_1. Let p_2 be the probability that we are in $2-4$. Then $p_1 + p_2 = 1$ and $p_1, p_2 \geq 0$. Given that we are in $1-3$, we are at 1 with probability $3/5$ (conditional probability). Therefore, the unconditional probability of being in 1 is $(3/5)p_1$. Similarly, we can find it for other states:

$$\begin{bmatrix}\frac{3}{5}p_1 & \frac{1}{2}p_2 & \frac{2}{5}p_1 & \frac{1}{2}p_2\end{bmatrix}$$

This is an infinite family of stationary distributions. Members of this family include:

$$\begin{bmatrix}\frac{3}{5} & 0 & \frac{2}{5} & 0\end{bmatrix} \quad \text{for} \quad p_1 = 1,\, p_2 = 0;$$

$$\begin{bmatrix}0 & \frac{1}{2} & 0 & \frac{1}{2}\end{bmatrix} \quad \text{for} \quad p_1 = 0,\, p_2 = 1;$$

$$\begin{bmatrix}\frac{3}{10} & \frac{1}{4} & \frac{1}{5} & \frac{1}{4}\end{bmatrix} \quad \text{for} \quad p_1 = p_2 = \frac{1}{2}.$$

Another way to find this would be simply to solve the system

$$\pi A = \begin{bmatrix}0 & 0 & 0 & 0\end{bmatrix} \quad \pi_1 + \pi_2 + \pi_3 + \pi_4 = 1.$$

This would give us a free variable:

$$-2\pi_1 + 3\pi_3 = 0,\; -\pi_2 + \pi_4 = 0, \quad \pi_1 + \pi_2 + \pi_3 + \pi_4 = 1.$$

$$\pi_1 = \frac{3}{2}\pi_3,\; \pi_4 = \pi_3,\; \frac{5}{2}\pi_3 + 2\pi_4 = 1 \;\Rightarrow\; \pi_4 = \frac{1}{2} - \frac{5}{4}\pi_3.$$

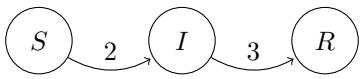
$$\begin{bmatrix}\pi_1 & \pi_2 & \pi_3 & \pi_4\end{bmatrix} = \begin{bmatrix}\frac{3}{2}\pi_3 & \frac{1}{2} - \frac{5}{4}\pi_3 & \pi_3 & \frac{1}{2} - \frac{5}{4}\pi_3\end{bmatrix}$$

$$= \pi_3 \begin{bmatrix}\frac{3}{2} & -\frac{5}{4} & 1 & -\frac{5}{4}\end{bmatrix} + \begin{bmatrix}0 & \frac{1}{2} & 0 & \frac{1}{2}\end{bmatrix}$$

This is the same answer, but in a different form.

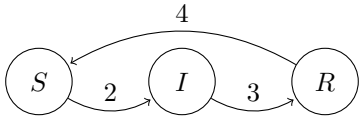
15.9. **Applications.** Consider a common infection model, with three states: S = susceptible, I = infected, R = recovered, and intensity 2 of moving from S to I, and intensity 3 from I to R; the state R is absorbing, and S, I are transient. The generator is

$$A = \begin{bmatrix}-2 & 2 & 0\\ 0 & -3 & 3\\ 0 & 0 & 0\end{bmatrix}$$



If a virus mutates and can threaten again after recovery, we can add, say, intensity 4 transition from R to S:

$$A = \begin{bmatrix}-2 & 2 & 0\\ 0 & -3 & 3\\ 4 & 0 & -4\end{bmatrix}$$



16. Queueing Theory

16.1. **M/M/1 queue.** Assume we have a server (cashier) which serves customers, one by one. Each customer is served for a random time, while others in the queue wait. This random time is independent for each customer and is distributed as $\mathrm{Exp}(\mu)$. Customers arrive (from the shop) with intensity λ; that is, the interarrival time of each next customer is distributed as $\mathrm{Exp}(\lambda)$, independently of other interarrival times, and of serving times. Let $X(t)$ be the number of customers in the queue at time t (including the customer who is currently being served). Then $X(t)$ can take values $0, 1, 2, \ldots$ Actually, this is a continuous-time Markov chain on the state space $\{0, 1, 2, \ldots\}$ with transition intensities

$$\lambda_{n,n+1} = \lambda,\ \lambda_{n,n-1} = \mu.$$

It has generating matrix

$$A = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \ldots \\ \mu & -\lambda-\mu & \lambda & 0 & \ldots \\ 0 & \mu & -\lambda-\mu & \lambda & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Let us find its stationary distribution

$$\pi = \begin{bmatrix} \pi_0 & \pi_1 & \pi_2 & \ldots \end{bmatrix}$$

It has to satisfy $\pi A = \begin{bmatrix} 0 & 0 & 0 & \ldots \end{bmatrix}$ Therefore,

$$-\lambda\pi_0 + \mu\pi_1 = 0, \tag{39}$$

$$\lambda\pi_{n-1} - (\lambda+\mu)\pi_n + \mu\pi_{n+1} = 0,\ n \geq 1. \tag{40}$$

Let us try to find a solution to this system of equations. First, consider the case $\lambda < \mu$: the arrival intensity is less than the service intensity. Try $\pi_n = c\rho^n,\ n = 0, 1, 2, \ldots$ Then plugging into the equation (40), we get:

$$\lambda \cdot c\rho^{n-1} - (\lambda+\mu)\cdot c\rho^n + \mu \cdot c\rho^{n+1} = 0.$$

Canceling $c\rho^{n-1}$, we have: $\lambda - (\lambda+\mu)\rho + \mu\rho^2 = 0$. Solving this quadratic equation, we get: $\rho_1 = 1,\ \rho_2 = \lambda/\mu$. Therefore, we get the following solutions of (40):

$$\pi_n = c_1 \cdot 1^n = c_1, \text{ and } \pi_n = c_2\left(\frac{\lambda}{\mu}\right)^n.$$

And their sum is also a solution to (40); actually, it is the most general solution:

$$\pi_n = c_1 + c_2\left(\frac{\lambda}{\mu}\right)^n. \tag{41}$$

Plug into (39) to find c_1 and c_2:

$$\pi_0 = c_1 + c_2,\ \pi_1 = c_1 + c_2\frac{\lambda}{\mu} \ \Rightarrow\ -\lambda(c_1+c_2) + \mu\left(c_1 + c_2\frac{\lambda}{\mu}\right) = (\mu-\lambda)c_1 \ \Rightarrow\ c_1 = 0.$$

Therefore, we have:

$$\pi_n = c_2\left(\frac{\lambda}{\mu}\right)^n,\quad n = 0, 1, 2, \ldots$$

Next, because of $\sum_{n=0}^{\infty}\pi_n = 1$, we have:

$$c_2\sum_{n=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^n = 1 \ \Rightarrow\ c_2\left(1-\frac{\lambda}{\mu}\right)^{-1} = 1 \ \Rightarrow\ c_2 = 1 - \frac{\lambda}{\mu}.$$

$$\boxed{\pi_n = \left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n,\quad n = 0, 1, 2, \ldots}$$

This is the distribution of $N = Z - 1$, where Z has a geometric distribution $\mathrm{Geo}(\rho)$ with parameter $\rho = 1 - \lambda/\mu \in (0, 1)$. We know mean and variance of the geometric distribution. Therefore, the distribution π has mean and variance

$$\mathbf{E}N = \mathbf{E}Z - 1 = \frac{1}{\rho} - 1,\ \mathrm{Var}\, N = \mathrm{Var}\, Z = \frac{1-\rho}{\rho^2}.$$

For the case $\lambda \geq \mu$, the intensity of arrival is greater than or equal to the intensity of service. One can show that there is no stationary distribution, and the queue, on average, grows infinitely large as time goes to infinity: $\mathbf{E}X(t) \to \infty$ as $t \to \infty$.

In the stationary distribution, the outflow of customers also forms a Poisson process with intensity λ, just like the inflow (Burke's theorem). This follows from the fact that the system is in the equilibrium.

16.2. **Finite queues.** Now, assume new people do not come when N people already are in the queue. Then $X = (X(t),\, t \geq 0)$ is a continuous-time Markov chain with state space $\{0, 1, \ldots, N-1, N\}$, with generator

$$A = \begin{bmatrix} -\lambda & \lambda & 0 & \ldots & 0 & 0 \\ \mu & -\lambda-\mu & \lambda & \ldots & 0 & 0 \\ 0 & \mu & -\lambda-\mu & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & 0 & \ldots & -\lambda-\mu & \lambda \\ 0 & 0 & 0 & \ldots & \mu & -\mu \end{bmatrix}$$

Let us find its stationary distribution $\pi = \begin{bmatrix} \pi_0 & \pi_1 & \ldots & \pi_N \end{bmatrix}$ It has to satisfy $\pi A = \begin{bmatrix} 0 & 0 & \ldots & 0 \end{bmatrix}$ We can write this as (39) and (40) for $n = 1, \ldots, N-1$; the last column gives us another boundary condition

$$\lambda \pi_{N-1} - \mu \pi_N = 0. \tag{42}$$

The general solution to the difference equation (40) is given by (41). The boundary condition (39) again gives us $c_1 = 0$, so we have

$$\pi_n = c \left(\frac{\lambda}{\mu}\right)^n,\ n = 0, \ldots, N.$$

This solution also satisfies (42). And to find c, we need to use

$$\pi_0 + \ldots + \pi_N = 1 \ \Rightarrow\ c\left(1 + \rho + \ldots + \rho^N\right) = 1, \quad \rho := \frac{\lambda}{\mu}. \tag{43}$$

Summing this finite geometric series, we get (for $\lambda \neq \mu \ \Leftrightarrow\ \rho \neq 1$):

$$c \frac{\rho^{N+1} - 1}{\rho - 1} = 1 \ \Rightarrow\ c = \frac{\rho - 1}{\rho^{N+1} - 1}.$$

The answer is:

$$\boxed{\pi_n = \frac{\rho - 1}{\rho^{N+1} - 1} \rho^n,\ n = 0, \ldots, N}$$

Separate case: $\lambda = \mu$. Then we have:

$$\pi_{n-1} - 2\pi_n + \pi_{n+1} = 0,\ n = 1, 2, \ldots, N-1. \tag{44}$$

From (39) and (42), we have:

$$\pi_0 = \pi_1,\ \pi_{N-1} = \pi_N. \tag{45}$$

We can just let $\pi_n = 1$ for all $n = 0, \ldots, N$. Then these conditions (44) and (45) are satisfied. But we need also to normalize it, so that (43) holds. Then we get:

$$\boxed{\pi_0 = \ldots = \pi_N = \frac{1}{N+1}}$$

This stationary distribution exists even if $\lambda \geq \mu$, because the queue cannot grow indefinitely.

16.3. **Finite queues with varying intensities.** Assume we have a lemonade stand. People come there with intensity $\lambda = 2$, and are served with intensity $\mu = 2$, but with the following qualifications. If there are currently no people in the queue, then the newcomer joins the queue. If there are one or two people in the queue, the newcomer joint the queue with probability $1/2$. If there are three or more people, then the newcomer does not join the queue. Let $X(t)$ be the number of people in the queue at time t. Then X is a continuous-time Markov chain with state space $\{0, 1, 2, 3\}$. When $X(t) = 1$, the intensity of jumping to 2 is $2 \cdot 0.5 = 1$. Indeed, if two people per minute on average come to the lemonade stand, but each of them chooses whether to stay or not independently with probability 0.5, then the effective rate of people joining the queue is 1 per minite. Similarly, when $X(t) = 2$, the intensity of jumping to 3 is $2 \cdot 0.5 = 1$. The intensity from 3 to 2, from 2 to 1, from 1 to 0, is 2. The generator is

$$A = \begin{bmatrix} -2 & 2 & 0 & 0 \\ 2 & -3 & 1 & 0 \\ 0 & 2 & -3 & 1 \\ 0 & 0 & 2 & -2 \end{bmatrix}$$

To find the stationary distribution $\pi = \begin{bmatrix}\pi_0 & \pi_1 & \pi_2 & \pi_3\end{bmatrix}$, solve $\pi A = \begin{bmatrix}0 & 0 & 0 & 0\end{bmatrix}$

17. Brownian Motion

17.1. **Continuous-time random walk.** Take a compound Poisson process

$$S(t) = \sum_{k=1}^{N(t)} Z_k,$$

with $N = (N(t), t \geq 0)$ being a Poisson process with intensity $\lambda = 1$, and with $Z_1, Z_2, \dots$ are i.i.d. variables with $\mathbf{P}(Z_i = 1) = \mathbf{P}(Z_i = -1) = 0.5$. This process S is called a *continuous-time random walk*, by analogy with a *discrete-time random walk* $X_n = Z_1 + \ldots + Z_n$. In fact, if you take $0 =: \tau_0 < \tau_1 < \dots$ to be the jump times of this continuous-time random walk S, then the corresponding discrete-time Markov chain is nothing else but a discrete-time random walk: $X_n = S(\tau_n)$.

$$\mathbf{E}Z_k = 1 \cdot 0.5 + (-1) \cdot 0.5 = 0, \ \mathbf{E}Z_k^2 = 1^2 \cdot 0.5 + (-1)^2 \cdot 0.5 = 1,$$

$$\text{and so} \quad \operatorname{Var} Z_k = \mathbf{E}Z_k^2 - (\mathbf{E}Z_k)^2 = 1.$$

Therefore,

$$\mathbf{E}S(t) = \lambda t \cdot \mathbf{E}Z_k = 0, \ \operatorname{Var} S(t) = t\left(\lambda(\mathbf{E}Z_k)^2 + \lambda^2 \operatorname{Var} Z_k\right) = t.$$

17.2. **Zooming out: Scaling limit is a Brownian motion.** To make jumps very small but very frequent, take a large n, and define

$$S_n(t) = \frac{1}{\sqrt{n}} \sum_{k=1}^{N(nt)} Z_k = \frac{1}{\sqrt{n}} S(nt).$$

This is also a compound Poisson process: Poisson process $(N(tn), t \geq 0)$ having intensity $\lambda n = n$, so jumps occur every n^{-1} times, on average; and each jump is up or down with equal probability, with size $n^{-1/2}$ of jump. We have:

$$\mathbf{E}S_n(t) = \frac{1}{\sqrt{n}}\mathbf{E}S(nt) = 0, \ \operatorname{Var} S_n(t) = \left(\frac{1}{\sqrt{n}}\right)^2 \operatorname{Var} S(nt) = \frac{1}{n} \cdot nt = t.$$

The random variable $S_n(t)$, ofr a large n, is the sum of many very small summands. By Central Limit Theorem, as $n \to \infty$, it converges to the normal distribution with the same mean and variance as $S_n(t)$. In other words,

$$S_n(t) = \frac{1}{\sqrt{n}} \sum_{k=1}^{N(nt)} Z_k \ \Rightarrow \ \mathcal{N}(0, t).$$

Similarly, for $t > s$, as $n \to \infty$,

$$S_n(t) - S_n(s) = \frac{1}{\sqrt{n}} \sum_{N(ns)+1}^{N(nt)} Z_k \ \Rightarrow \ \mathcal{N}(0, t-s).$$

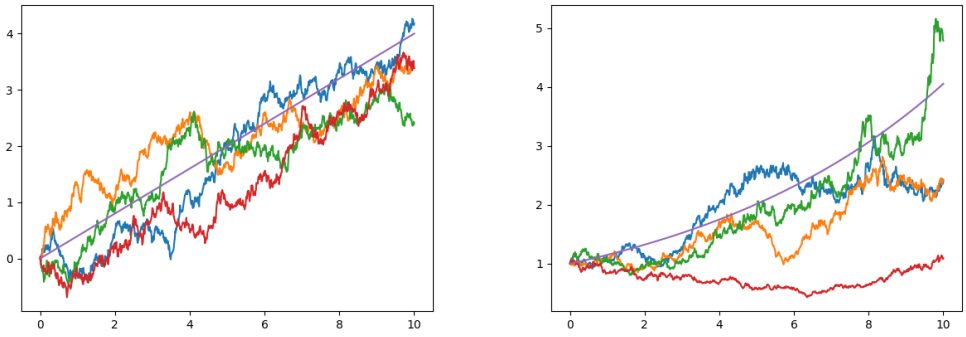
And this difference is independent of $S_n(u)$ for $u \leq s$. Therefore, $S_n(t) \Rightarrow W(t)$ as $n \to \infty$. Here, we have the following properties of W:

(a) $W(t) - W(s) \sim \mathcal{N}(0, t-s)$ for $t > s$, and is independent of $W(u)$, $u \leq s$;
(b) $W(0) = 0$, because $S_n(0) = 0$.
(c) W has continuous trajectories, because the jumps of S_n are very small: only of size n^{-1}.

This process W which satisfies (a), (b), (c), is called a *Brownian motion.* This is the most important random (stochastic) process in Probability Theory.

This movement is a sum of very frequent but very small jumps. It was first used by Einstein to model the motion of a small dust particle hit by molecules (many small hits). Brownian motion is also used to model the stock price, because its movement is influenced by a lot of new bits of information, which are random (if they were not random, they would be predictable, that is, not new).

Brownian motion is a Markov process: Future movement $W(t)$, $t > s$, does not depend on the past $W(t)$, $t < s$, if we know the present $W(s)$. This is because $W(t) - W(s)$ is independent of $W(u)$, $u \leq s$. The state space of the Brownian motion is the whole real line: $\mathbb{R}$.



(A) Brownian Motion (B) Geometric Brownian Motion

FIGURE 1. Left panel: Brownian motion with drift $g = 0.4$ and diffusion $\sigma = .5$, starting from 0, and the drift line $x = gt$. Right Panel: Geometric Brownian motion with drift $g = 0.1$ and diffusion $\sigma = 0.2$, starting from 1, and the expectation $x = e^{(g+\sigma^2/2)t}$

17.3. **Expectation, variance, and transition density.** Assume $W(s) = x$. Then $W(t) \sim \mathcal{N}(x, t-s)$, because $W(t) - W(s) \sim \mathcal{N}(0, t-s)$. Therefore, the density of $W(t)$ given $W(s) = x$ is

$$p(s,t,x,y) = \frac{1}{\sqrt{2\pi(t-s)}} \exp\left(-\frac{(y-x)^2}{2(t-s)}\right).$$

In other words, for every $a < b$,

$$\mathbf{P}(a \le W(t) \le b \mid W(s) = x) = \int_a^b p(s,t,x,y)\,\mathrm{d}y.$$

We often write this as

$$p(s,t,x,y) = \varphi(t-s, y-x),\ \varphi(t,z) := \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{z^2}{2t}\right).$$

Example 17.1. We have:

$$\mathbf{E}W(2)W(3) = \mathbf{E}W^2(2) + \mathbf{E}\left[W(2)(W(3) - W(2))\right] = \operatorname{Var} W(2) + \mathbf{E}(W(3) - W(2))\mathbf{E}W(2) = 2.$$

More generally, $\mathbf{E}W(s)W(t) = \min(s,t)$.

Example 17.2. For every number x, we have:

$$\begin{aligned}\mathbf{E}\left[W(3) \mid W(2) = x\right] &= \mathbf{E}\left[W(3) - W(2) \mid W(2) = x\right] + \mathbf{E}\left[W(2) \mid W(2) = x\right] \\ &= \mathbf{E}(W(3) - W(2)) + x = 0 + x = x.\end{aligned}$$

Example 17.3. $\mathbf{E}\left[W^2(5) \mid W(3) = 4\right] = \mathbf{E}\left(\xi + 4\right)^2 = \mathbf{E}\xi^2 + 8\mathbf{E}\xi + 16 = 2 + 0 + 16 = 18$, $\xi = W(5) - W(3) \sim \mathcal{N}(0,2)$.

Example 17.4. $\mathbf{E}\left[(W(6) - W(4))W(5)\right] = \mathbf{E}W(5)W(6) - \mathbf{E}W(4)W(5) = \min(5,6) - \min(4,5) = 1$.

Example 17.5. $\mathbf{P}(W(3) > 0) = 0.5$, because $W(3) \sim \mathcal{N}(0,3)$ is symmetric with respect to zero.

Example 17.6. Let us find $\mathbf{P}(W(3) > 1 \mid W(1) = 2)$. We have: $\xi := W(3) - W(1) \sim \mathcal{N}(0,2)$. Therefore,

$$\mathbf{P}(W(3) > 1 \mid W(1) = 2) = \mathbf{P}(\xi + 2 > 1) = \mathbf{P}(\xi > -1).$$

But $\xi = \sqrt{2}Z$ for $Z \sim \mathcal{N}(0,1)$, so

$$\mathbf{P}(\xi > -1) = \mathbf{P}(\sqrt{2}Z > -1) = \mathbf{P}(Z > -1/\sqrt{2}) = \mathbf{P}(Z > -0.707) = 0.7611.$$

Example 17.7. Let us find the density of $W^3(2)$, given $W(1) = 2$. We have: $W^3(2) = (\xi + 2)^3$ for $\xi \sim \mathcal{N}(0,1)$. Then the density of ξ is

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

and the density of $(\xi + 2)^3$ can be found as follows:

$$\mathbf{P}\left(a \le (\xi + 2)^3 \le b\right)$$

$$= \mathbf{P}\left(a^{1/3} - 2 \le \xi \le b^{1/3} - 2\right) = \int_{a^{1/3}-2}^{b^{1/3}-2} f(x)\,\mathrm{d}x$$
$$= \int_a^b f(y^{1/3} - 2)\frac{1}{3}y^{-2/3}\mathrm{d}y.$$

where we change variables $x = y^{1/3} - 2$, so $\mathrm{d}x = y^{-2/3}\mathrm{d}y/3$, and

$$a^{1/3} - 2 \le x \le b^{1/3} - 2 \;\Leftrightarrow\; a^{1/3} - 2 \le y^{1/3} - 2 \le b^{1/3} - 2 \;\Rightarrow\; a \le y \le b$$

Therefore, the density of $(\xi + 2)^3$ is given by

$$f(y^{1/3} - 2)\frac{1}{3}y^{-2/3} = \frac{1}{3\sqrt{2\pi}y^{2/3}}\exp\left(-\frac{(y^{1/3} - 2)^2}{2}\right).$$

17.4. **Brownian motion with drift and diffusion.** Take coefficients g and σ, a starting point x, and define

$$X(t) = x + gt + \sigma W(t).$$

This is a *Brownian motion starting from* x, with *drift coefficient* g, and *diffusion coefficient* σ^2. The *standard Brownian motion* has $x = 0$, $g = 0$, $\sigma = 1$. When we say "Brownian motion" without specifying the coefficients and the starting point, we mean the standard Brownian motion. This process has properties similar to the standard Brownian motion: for $t > s$,

$$X(t) - X(s) = g(t - s) + \sigma(W(t) - W(s)) \sim \mathcal{N}\left(g(t - s), \sigma^2(t - s)\right),$$

and this random variable is independent of $X(u)$, $u \le s$.

Example 17.8. Consider $x = 0$, $g = -1$, $\sigma = 2$. Find $\mathbf{E}(X(3) \mid X(2) = 2)$. We have: $X(t) = -t + 2W(t)$, and so $X(3) - X(2) = -1 + 2(W(3) - W(2)) \sim \mathcal{N}(-1, 4)$. Therefore, $\mathbf{E}(X(3) \mid X(2) = 2) = \mathbf{E}(X(3) - X(2)) + 2 = -1 + 2 = 1$. Next, find $\mathbf{E}(X^2(5) \mid X(2) = 2)$: if $\xi := X(5) - X(2) \sim \mathcal{N}(-3, 12)$, then we have $X^2(5) = (2 + \xi)^2$, and the answer is

$$\mathbf{E}(2 + \xi)^2 = 4 + 4\mathbf{E}\xi + \mathbf{E}\xi^2 = 4 + 4 \cdot (-3) + (-3)^2 + 12 = 13.$$

17.5. **Levy processes.** Combine a Brownian motion with a compound Poisson process:

$$L(t) = x + gt + \sigma W(t) + \sum_{k=1}^{N(t)} Z_k. \tag{46}$$

Here, W is a standard Brownian motion, N is a Poisson process with intensity λ, and $Z_1, Z_2, \ldots$ are i.i.d. Also, W, N, Z_k are all independent. This process behaves as a Brownian motion between jumps, which occur every $\mathrm{Exp}(\lambda)$ times, and the displacement during the kth jump is Z_k. This process also has the Markov property: $L(t) - L(s)$ is independent of $L(u)$, $u \le s$. Also, it has *stationary increments*: the distribution of $L(t) - L(s)$ depends only on $t - s$ for $t > s$. For example, the distribution of $L(6) - L(4)$ and of $L(5) - L(3)$ is the same.

Example 17.9. $x = 1$, $g = 2$, $\sigma = 3$, $\lambda = 2$, $Z_k \sim \mathrm{Exp}(2)$. Then

$$\mathbf{E}L(t) = x + gt + \sigma\mathbf{E}W(t) + \mathbf{E}\sum_{k=1}^{N(t)} Z_k = 1 + 2t + \lambda t\mathbf{E}Z_k = 1 + 2t + 2t \cdot \frac{1}{2} = 1 + 3t,$$

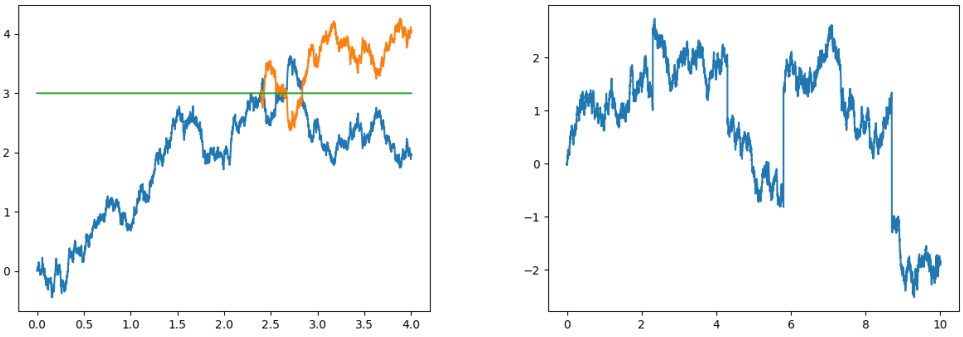
$$\mathrm{Var}\, L(t) = \sigma^2 \mathrm{Var}\, W(t) + \mathrm{Var}\sum_{k=1}^{N(t)} Z_k = \sigma^2 t + t\lambda\left((\mathbf{E}Z_k)^2 + \mathrm{Var}\, Z_k\right)$$
$$= 9t + 2t\left(0.5^2 + \frac{1}{4}\right) = 10t.$$

Let $G(u) := \mathbf{E}e^{uZ_k}$ be the moment generating function of Z_k. It is equal to

$$G(u) = \frac{2}{2 - u}.$$

Then the moment generating function of $L(t)$ is given by

$$F_t(u) := \mathbf{E}\exp(uL(t)) = \mathbf{E}\exp\left[u\left(x + gt + \sigma W(t)\right)\right]\mathbf{E}\exp\left[u\sum_{k=1}^{N(t)} Z_k\right]$$



(A) Reflection Principle (B) Lévy Process

FIGURE 2. Left: Illustration of the reflection principle for the Brownian motion. Right: A Lévy process: a standard Brownian motion plus a compound Poisson process with intensity of jumps $\lambda = 0.5$.

$$= \exp\left(ux + gut + \frac{\sigma^2 u^2}{2}t\right) \exp\left(\lambda t(G(u) - 1)\right)$$
$$= e^{ux} \exp\left(t\left[gu + \sigma^2 u^2/2 + \lambda(G(u) - 1)\right]\right) = \exp\left(u + t\left[2u + \frac{9}{2}u^2 + 2\left(\frac{2}{2-u} - 1\right)\right]\right).$$

Any process with stationary independent increments is called a Levy process. One can show that if such process has at most finitely many jumps on a finite time interval, then it can be represented as (46).

17.6. **Reflection principle and the maximum of Brownian motion.** Try to find the distribution of $M(t) = \max_{0\le s\le t} W(s)$. For example, fin

$$\mathbf{P}(M(4) \ge 3,\ 1 \le W(4) \le 2). \tag{47}$$

The trajectory of the Brownian motion which satisfies (47) crosses the line $y = 3$. Let us take the first moment of crossing and reflect the trajectory of the Brownian motion across this line, starting from this moment. Then we get another trajectory of Brownian motion, because it is symmetric (can go up or down with equal probability). But it reaches between 4 and 5 at time $t = 4$, because 4 and 5 are symmetric to the points 2 and 1 respectively, with respect to this line. Conversely, every such trajectory of a Brownian motion with $4 \le W(4) \le 5$, after being reflected after its first crossing of line $y = 3$, becomes a Brownian motion which satisfies (47). Therefore, the probability from (47) is equal to (with $W(4) = 2Z$, $Z \sim \mathcal{N}(0, 1)$):

$$\mathbf{P}(4 \le W(4) \le 5) = \frac{1}{\sqrt{2\pi}2}\int_4^5 e^{-x^2/8}\,\mathrm{d}x = \mathbf{P}(2 \le Z \le 2.5) = 0.9938 - 0.9772 = \boxed{0.0166}$$

Similarly,

$$\mathbf{P}(M(4) \ge 3,\ W(4) \le 3) = \mathbf{P}(W(4) \ge 3) = \frac{1}{\sqrt{2\pi}2}\int_3^\infty e^{-x^2/8}\,\mathrm{d}x.$$

Note that if $W(4) \ge 3$, then certainly $M(4) \ge 3$. And

$$\mathbf{P}(M(4) \ge 3) = \mathbf{P}(M(4) \ge 3,\ W(4) \le 3) + \mathbf{P}(W(4) \ge 3) = 2\mathbf{P}(W(4) \ge 3).$$

Similarly, for every $t > 0$, and $y \ge 0$,

$$\mathbf{P}(M(t) \ge y) = 2\mathbf{P}(W(t) \ge y) = \mathbf{P}(W(t) \ge y) + \mathbf{P}(W(t) \le -y) = \mathbf{P}(|W(t)| \ge y).$$

That is, $M(t)$ and $|W(t)|$ have the same distribution. It has density

$$\frac{2}{\sqrt{2\pi t}}e^{-x^2/2t}\,\mathrm{d}x,\ x \ge 0.$$

17.7. **Reflected Brownian motion.** This is the process $|W| = (|W(t)|,\, t \geq 0)$. It is a Markov process, because the behavior of $|W(t)|$ for $t \geq s$ depends only on $|W(s)|$; if you know $W(u)$, $u < s$, this does not give you any additional information. The state space of this process is $\mathbb{R}_+ := [0, \infty)$. *Skorohod representation:*

$$|W(t)| = B(t) + \ell(t),$$

where B is another Brownian motion (not W!), and ℓ is a continuous nondecreasing process with $\ell(0) = 0$, with ℓ increasing only when $|W| = 0$. As long as $|W|$ stays inside the positive half-line, it behaves as a Brownian motion. But when it hits zero, it "wants" to go down, but is not allowed to, because it has to stay positive. Then we add a push $\mathrm{d}\ell(t)$ to make it positive, and prevent it from crossing $y = 0$. The process ℓ is called a *local time* of $|W|$ at zero.

The transition density of $|W(t)|$: for any $0 \leq a < b$,

$$\begin{aligned}
&\mathbf{P}\left(a \leq |W(t)| \leq b \mid |W(s)| = x\right) \\
&= \mathbf{P}\left(a \leq W(t) \leq b \mid W(s) = x\right) + \mathbf{P}\left(-b \leq W(t) \leq -a \mid W(s) = x\right) \\
&= \frac{1}{\sqrt{2\pi(t-s)}} \int_a^b \exp\left(-\frac{(y-x)^2}{2(t-s)}\right) \mathrm{d}y + \frac{1}{\sqrt{2\pi(t-s)}} \int_{-b}^{-a} \exp\left(-\frac{(y-x)^2}{2(t-s)}\right) \mathrm{d}y \\
&= \frac{1}{\sqrt{2\pi(t-s)}} \int_a^b \left[\exp\left(-\frac{(y-x)^2}{2(t-s)}\right) + \exp\left(-\frac{(-y-x)^2}{2(t-s)}\right)\right] \mathrm{d}y
\end{aligned}$$

so given $|W(s)| = x$, $|W(t)|$ has density

$$p(s,t,x,y) = \frac{1}{\sqrt{2\pi(t-s)}} \left[\exp\left(-\frac{(y-x)^2}{2(t-s)}\right) + \exp\left(-\frac{(y+x)^2}{2(t-s)}\right)\right].$$

Example 17.10. Let us find $\mathbf{P}(|W(2)| \leq 1 \mid |W(1)| = 1)$. Let $\xi := W(2) - W(1) \sim \mathcal{N}(0,1)$. Then

$$\begin{aligned}
\mathbf{P}(|W(2)| \leq 1 \mid |W(1)| = 1) &= \mathbf{P}(-1 \leq W(2) \leq 1 \mid W(1) = 1) \\
&= \mathbf{P}(-1 \leq \xi - 1 \leq 1) = \mathbf{P}(0 \leq \xi \leq 2) = \frac{1}{\sqrt{2\pi}} \int_0^2 e^{-x^2/2}\, \mathrm{d}x = 0.48.
\end{aligned}$$

Example 17.11. Let us find $\mathbf{E}|W(3)|$. We can represent $W(3) = \sqrt{3}\xi$, $\xi \sim \mathcal{N}(0,1)$. Then

$$\begin{aligned}
\mathbf{E}|W(3)| &= \sqrt{3}\mathbf{E}|\xi| = \sqrt{3}\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |y| e^{-y^2/2}\, \mathrm{d}y \\
&= \sqrt{3}\frac{2}{\sqrt{2\pi}} \int_0^{\infty} y e^{-y^2/2}\, \mathrm{d}y = \sqrt{3}\frac{2}{\sqrt{2\pi}} \left(-e^{-y^2/2}\right)\Big|_{y=0}^{\infty} = \frac{2\sqrt{3}}{\sqrt{2\pi}}.
\end{aligned}$$

17.8. **Transformations of Brownian motion.** Note that the process $(W(9t),\, t \geq 0)$ is also a Brownian motion with diffusion coefficient 9 and drift coefficient 0. In other words, the following processes have the same distribution:

$$(W(9t),\, t \geq 0) \text{ and } (3W(t),\, t \geq 0).$$

Indeed, $W(9t)$ and $3W(t)$ have the same distribution: $\mathcal{N}(0, 9t)$. Moreover, $W(9t) - W(9s)$ and $3W(t) - 3W(s)$ have the same distribution: $\mathcal{N}(0, 9(t-s))$, for $t > s$. And $W(9t) - W(9s)$ is independent of $W(9u)$, $u \leq s$.

Consider two independent Brownian motions: $X = (X(t),\, t \geq 0)$ is a Brownian motion with drift $g_1 = -2$ and diffusion $\sigma_1^2 = 3$, starting from $x_1 = -1$, and $Y = (Y(t),\, t \geq 0)$ is a Brownian motion with drift $g_2 = 3$ and diffusion $\sigma_2^2 = 4$, starting from $x_2 = 2$. Then take the process

$$Z = (Z(t),\, t \geq 0) = 2X - 3Y = (2X(t) - 3Y(t),\, t \geq 0).$$

This is also a Brownian motion, starting from

$$Z(0) = 2X(0) - 3Y(0) = 2 \cdot (-1) - 3 \cdot 2 = -8.$$

For $t > s$, we have:

$$Z(t) - Z(s) \sim 2(X(t) - X(s)) - 3(Y(t) - Y(s)).$$

But the following increments are normally distributed and are independent:

$$X(t) - X(s) \sim \mathcal{N}(-2(t-s), 3(t-s)),\ Y(t) - Y(s) \sim \mathcal{N}(3(t-s), 4(t-s)).$$

Therefore, using results of Section 9, we get: $Z(t) - Z(s)$ is also normal with

$$\mathbf{E}(Z(t) - Z(s)) = 2\mathbf{E}(X(t) - X(s)) - 3\mathbf{E}(Y(t) - Y(s))$$

$$= -2(t-s)\cdot 2 + 3(t-s)\cdot(-3) = -13(t-s),$$

$$\begin{aligned}\operatorname{Var}(Z(t)-Z(s)) &= 2^2\operatorname{Var}(X(t)-X(s)) + (-3)^2\operatorname{Var}(Y(t)-Y(s))\\ &= 2^2\cdot 3(t-s) + (-3)^2\cdot 4(t-s) = 48(t-s).\end{aligned}$$

And $Z(t)-Z(s)$ is independent of $Z(u)$, $u\le s$, because X and Y satisfy the same property. Therefore, Z is a Brownian motion with drift coefficient -13 and diffusion coefficient 48, starting from -8.

17.9. **Derivatives in the Black-Scholes model.** This is an extension of stochastic finance theory from Section 12 for the continuous case. We model the stock price as the geometric Brownian motion:

$$S(t) = S_0\exp(gt+\sigma W(t)).$$

Here, σ is called *volatility*; this parameter shows the magnitude of fluctuations in the price. One can see the definitions of European options and derivatives in Section 12. We would like to hedge a derivative $D = f(S(T))$, where $f(x)$ is some real-valued function. In particular, we are interested in the following functions, with $a_+ := \max(a,0)$: (a) $f(x) = (x-K)_+$: *European option-call* with *strike* K; (b) $f(x) = (K-x)_+$: *European option-put* with *strike* K; (c) $f(x) = 1(x>K)$: *Binary option* with *barrier* K.

If we had random walk instead of a Brownian motion, we would switch to a risk-neutral probability, under which $\mathbf{E}S(t) = S_0$. Here, we have a (geometric) Brownian motion, which is a limit of geometric random walks with small but frequent jumps. It is possible to show that in this case, risk-neutral probability corresponds to a change in drift. That is, we need to find new g_* such that $\mathbf{E}S(t) = S_0$. We can do this as follows: Calculating the moment generating function of $W(t)\sim\mathcal{N}(0,t)$, we get:

$$\mathbf{E}\exp(g_*t+\sigma W(t)) = e^{g_*t}e^{\sigma^2 t/2} = 1 \;\Rightarrow\; g_* = -\frac{\sigma^2}{2}.$$

Therefore, under the new risk-neutral probability, the stock price is modeled by a geometric Brownian motion:

$$S(t) = S_0\exp\left(\sigma W(t) - \frac{\sigma^2}{2}t\right),\ t\ge 0. \tag{48}$$

And we need to take the expected value: $v = \mathbf{E}f(S(T))$. This is the fair price. We can find the way to *hedge*, or *replicate* these derivatives, using stochastic calculus in the next section.

Example 17.12. For the binary option with barrier $K=3$, maturity $T=2$, traded upon the stock with volatility $\sigma=3$ and initial price $S_0=1.5$, its risk-neutral dynamics is given by $S(t) = 1.5\exp(3W(t)-9t/2)$, therefore the fair price of this option is (with $W(2)=\sqrt{2}Z$ for $Z\sim\mathcal{N}(0,1)$):

$$\begin{aligned}\mathbf{E}1(S(2)>3) &= \mathbf{P}\left(1.5\exp\left(3W(2)-9\cdot 2/2\right)>3\right) = \mathbf{P}\left(\exp\left(3\sqrt{2}Z-9\right)>2\right)\\ &= \mathbf{P}\left(3\sqrt{2}Z-9>\ln 2\right) = \mathbf{P}\left(Z>\frac{3}{\sqrt{2}}+\frac{\ln 2}{3\sqrt{2}}\right) = \mathbf{P}(Z>2.28) = 0.011\end{aligned}$$

For the European option call, we get the celebrated *Black-Scholes formula* (Ecomonics Nobel Prize 1997):

Example 17.13. Try strike $K=1.5$, maturity $T=3$, volatility $\sigma=2$, and current price $S_0=1$. Then $S(t) = \exp(2W(t)-2t)$, and $S(T) = \exp(2W(3)-6)$. Then the fair price is

$$v = \mathbf{E}(S(T)-K)_+ = \mathbf{E}\left(\exp(2W(3)-6)-1.5\right)_+.$$

The option is executed if its terminal price at maturity is greater than strike, that is,

$$\exp(2W(3)-6)-1.5>0 \;\Leftrightarrow W(3) > 3+\frac{1}{2}\log 1.5 = 3.405,$$

The probability density function of $W(3)$ is given by $(6\pi)^{-1/2}\exp(-x^2/6)$. Then

$$\begin{aligned}v = \mathbf{E}\left(\exp(2W(3)-6)-1.5\right)_+ &= \frac{1}{\sqrt{6\pi}}\int_{3+0.5\log 1.5}^{\infty}\left[e^{2x-6}-1.5\right]e^{-x^2/6}\,\mathrm{d}x\\ &= \frac{1}{\sqrt{6\pi}}\int_{3+0.5\log 1.5}^{\infty}e^{2x-6}\,e^{-x^2/6}\,\mathrm{d}x - \frac{1}{\sqrt{6\pi}}\int_{3+0.5\log 1.5}^{\infty}1.5\,e^{-x^2/6}\,\mathrm{d}x.\end{aligned} \tag{49}$$

The density $(6\pi)^{-1/2}e^{-x^2/6}$ is of $\sqrt{3}Z$, $Z\sim\mathcal{N}(0,1)$. The second term in the right-hand side of (49) is equal to:

$$\frac{1}{\sqrt{6\pi}}\int_{3+0.5\log 1.5}^{\infty}1.5e^{-x^2/6}\,\mathrm{d}x = \mathbf{P}(\sqrt{3}Z>3+0.5\log 1.5)$$

$$= \mathbf{P}(\sqrt{3}Z > 3 + 0.5\log 1.5) = \mathbf{P}\left(Z > \sqrt{3} + \frac{0.5\log 1.5}{\sqrt{3}}\right).$$

The first term in the right-hand side of (49) is equal to:

$$\frac{1}{\sqrt{6\pi}}\int_{3+0.5\log 1.5}^{\infty} e^{2x-6}\, e^{-x^2/6}\,\mathrm{d}x = \frac{1}{\sqrt{6\pi}}\int_{3+0.5\log 1.5}^{\infty} e^{-x^2/6+2x-6}\,\mathrm{d}x = \frac{1}{\sqrt{6\pi}}\int_{3+0.5\log 1.5}^{\infty} e^{-(x-6)^2/6}\,\mathrm{d}x$$
$$= \mathbf{P}(\sqrt{3}Z + 6 > 3 + 0.5\log 1.5) = \mathbf{P}\left(Z > -\sqrt{3} + \frac{0.5\log 1.5}{\sqrt{3}}\right).$$

Therefore, the answer is

$$v = \mathbf{P}\left(Z > -\sqrt{3} + \frac{0.5\log 1.5}{\sqrt{3}}\right) - 1.5\cdot\mathbf{P}\left(Z > \sqrt{3} + \frac{0.5\log 1.5}{\sqrt{3}}\right)$$
$$= \mathbf{P}(Z > -1.53) - 1.5\cdot\mathbf{P}(Z > 1.93) = 0.937 - 1.5\cdot(1 - 0.9732) = \boxed{0.8968}$$

We can also find the fair price of a European derivative at any time t, if we know the price $S(t) = x$ at this time:

$$u(t,x) = \mathbf{E}(f(S(T)) \mid S(t) = x),$$

where the expectation is taken again with respect to the risk-neutral probability, that is, for the process $S(t)$ from (48). Then we simply treat time t as the initial time, and $S(t) = x$ as the initial price x. The hedging of such a derivative: that is, replicating it with a combination of cash and this stock, is slightly harder, and is deferred to the next section on stochastic calculus.

17.10. **Drawbacks of the Black-Scholes model.** *Heavy Tails.* Large fluctuations of stocks occur more frequently than prescribed by the normal distributions. A better distribution is $\mathbf{P}(\ln(S(t_2)/S(t_1)) \geq x) \approx cx^{-\alpha}$, and same for $\mathbf{P}(\ln(S(t_2)/S(t_1)) \leq -x)$; here, $c, \alpha > 0$ are some constants.

Short-term dependence. Increments of logarithms are *not* independent: say $\ln S(2.1) - \ln S(2)$ and $\ln S(2.2) - \ln S(2.1)$ are dependent, although they should be independent if we lived in the Black-Scholes model.

Volatility is not constant. If we calculate σ^2 from real data, it should be constant in the Black-Scholes model, but it is not. In fact, the graph $(S(t), \sigma(t))$, of the stock price and volatility, is not a horizontal line, but a decreasing one: *"volatility skew"* or a U-shaped curve: *"volatility smile"*. To explain this, several techniques are available, including *stochastic volatility models:* when $\sigma(t)$ is modeled itself by a stochastic process, for example using second Brownian motion $B(t)$, correlated with $W(t)$.

Bid-ask spread. The stock price is actually two different prices: *bid*, the price offered for buying, and *ask*, the price offered for selling. Selling - buying occurs when a bid is equal to an ask. The dynamics of bid offers and ask offers is modeled separately.

18. Stochastic Calculus

In this section, our aim is to differentiate Brownian motion $W = (W(t),\, t \geq 0)$. We cannot do this in the usual sense, because this is not differentiable. Indeed,

$$W(t + \mathrm{d}t) - W(t) \sim \mathcal{N}(0, \mathrm{d}t) \;\Rightarrow\; \frac{W(t+\mathrm{d}t) - W(t)}{\mathrm{d}t} \sim \mathcal{N}\left(0, \frac{1}{\mathrm{d}t^2}\cdot \mathrm{d}t\right) = \mathcal{N}\left(0, \frac{1}{\mathrm{d}t}\right). \tag{50}$$

As $\mathrm{d}t \to 0$, the variance $1/\mathrm{d}t \to \infty$. Therefore, this ratio in (50) does not converge anywhere.

18.1. **Quadratic variation.** For a function $f : [0,\infty) \to \mathbb{R}$, define its *quadratic variation* on the interval $[0,t]$:

$$\langle f\rangle_t := \lim_{N\to\infty} \sum_{k=0}^{N-1} \left(f\left(\frac{(k+1)t}{N}\right) - f\left(\frac{kt}{N}\right)\right)^2.$$

We split the interval $[0,t]$ into N equal subintervals

$$\left[0, \frac{t}{N}\right], \left[\frac{t}{N}, \frac{2t}{N}\right], \ldots, \left[\frac{(N-1)t}{N}, t\right].$$

and calculate the increment of f on each such subinterval; then square each and sum them. Then let the size of intervals go to zero. For smooth functions f, one can show that this quantity is zero. But for the Brownian motion W, we have independent identically distributed increments:

$$W\left(\frac{(k+1)t}{N}\right) - W\left(\frac{kt}{N}\right) \sim \mathcal{N}\left(0, \frac{t}{N}\right).$$

Therefore, we can represent

$$W\left(\frac{(k+1)t}{N}\right) - W\left(\frac{kt}{N}\right) = (t/N)^{1/2} Z_k,\ k = 0, \ldots, N-1,$$

where $Z_1, \ldots, Z_N$ are i.i.d. $\mathcal{N}(0,1)$. Therefore, by the Law of Large Numbers, as $N \to \infty$:

$$\left((t/N)^{1/2} Z_1\right)^2 + \ldots + \left((t/N)^{1/2} Z_N\right)^2 = \frac{t}{N}\left(Z_1^2 + \ldots + Z_N^2\right) \to t\mathbf{E}Z_1 = t.$$

We have shown that $\langle W \rangle_t = t$. We can also symbolically represent quadratic variation as

$$\langle f \rangle_t = \int_0^t (\mathrm{d}f)^2.$$

In other words, $\mathrm{d}\langle f \rangle_t = (\mathrm{d}f)^2$. For smooth functions f, we have: $\mathrm{d}f = f'(t)\,\mathrm{d}t$, and therefore

$$\mathrm{d}\langle f \rangle_t = (f'(t)\,\mathrm{d}t)^2 = f'^2(t)\,(\mathrm{d}t)^2 = 0.$$

Indeed, in such calculations we must consider $(\mathrm{d}t)^2 = 0$, and in general $(\mathrm{d}t)^a = 0$ for every $a > 1$.

18.2. **Construction of a stochastic integral.** For a deterministic function or a random process $H = (H(t),\, t \geq 0)$, define the *stochastic integral*, or *Itô's integral:*

$$\begin{aligned}\int_0^t H(s)\,\mathrm{d}W(s) &= \lim_{N\to\infty} \sum_{k=0}^{N-1} H\left(\frac{kt}{N}\right)\left(W\left(\frac{(k+1)t}{N}\right) - W\left(\frac{kt}{N}\right)\right) \\ &= H(0)\left(W\left(\frac{t}{N}\right) - W(0)\right) + H\left(\frac{t}{N}\right)\left(W\left(\frac{2t}{N}\right) - W\left(\frac{t}{N}\right)\right) \\ &\quad + \ldots + H\left(\frac{t(N-1)}{N}\right)\left(W(t) - W\left(\frac{t(N-1)}{N}\right)\right).\end{aligned} \tag{51}$$

This is similar to a Riemann sum for the integral $\int_0^t H(s)\,\mathrm{d}s$.

Example 18.1. Let $H(s) \equiv c$. Then the sum in (51) becomes $c(W(t) - W(0)) = cW(t)$. That is,

$$\int_0^t c\mathrm{d}W(s) = cW(t).$$

18.3. **Stochastic integral of a deterministic function.** More generally, take a deterministic (non-random) function $f : [0,t] \to \mathbb{R}$, and let $H := f$. Then the sum in (51) is a linear combination of independent normal random variables

$$Z_k := W\left(\frac{(k+1)t}{N}\right) - W\left(\frac{kt}{N}\right) \sim \mathcal{N}\left(0, \frac{t}{N}\right),\quad k = 0, \ldots, N-1.$$

Each of which has $\mathbf{E}Z_k = 0$ and $\operatorname{Var} Z_k = t/N$. We can rewrite this sum as

$$S_N := f(0)\,Z_0 + f\left(\frac{t}{N}\right) Z_1 + \ldots + f\left(\frac{(N-1)t}{N}\right) Z_{N-1}.$$

It also has normal distribution (as a linear combination of independent normal random variables), with

$$\mathbf{E}S_N = f(0)\,\mathbf{E}Z_0 + \ldots + f\left(\frac{(N-1)t}{N}\right)\mathbf{E}Z_{N-1} = 0,$$

$$\operatorname{Var} S_N = f^2(0)\,\operatorname{Var} Z_0 + \ldots + f^2\left(\frac{(N-1)t}{N}\right)\operatorname{Var} Z_{N-1} = \frac{t}{N}\sum_{k=0}^{N-1} f^2\left(\frac{kt}{N}\right).$$

Note that $\operatorname{Var} S_N$ is a Riemann sum for the integral $I := \int_0^t f^2(s)\,\mathrm{d}s$. Therefore, $\operatorname{Var} S_N \to I$. And the stochastic integral is distributed as the normal random variable:

$$\int_0^t f(s)\,\mathrm{d}W(s) \sim \mathcal{N}(0, I).$$

Example 18.2. Consider the integral $\int_0^2 (3-t)\,\mathrm{d}W(t)$. It is distributed as $\mathcal{N}(0, \sigma^2)$ with

$$\sigma^2 = \int_0^2 (3-s)^2\,\mathrm{d}s = \int_0^2 \left(9 - 6s + s^2\right)\,\mathrm{d}s = \left(9s - 3s^2 + \frac{s^3}{3}\right)\Big|_{s=0}^{s=2} = \frac{26}{3}.$$

Example 18.3. Consider the integral $\int_0^1 t\,\mathrm{d}X(t)$, where X is a Brownian motion with drift $g=-1$ and diffusion $\sigma^2=4$. Then we have: $\mathrm{d}X(t)=-\mathrm{d}t+2\mathrm{d}W(t)$, and therefore

$$\int_0^1 t\,\mathrm{d}X(t)=-\int_0^1 t\,\mathrm{d}t+2\int_0^1 t\,\mathrm{d}W(t)\sim\mathcal{N}\left(-\int_0^1 t\,\mathrm{d}t,4\int_0^1 t^2\,\mathrm{d}t\right)=\mathcal{N}\left(-\frac{1}{2},\frac{4}{3}\right).$$

If the integrator $H=(H(t),\,t\geq 0)$ is random, then similarly we can show that

$$\mathbf{E}\int_0^T X(t)\,\mathrm{d}W(t)=0 \text{ and } \mathbf{E}\left[\int_0^T X(t)\,\mathrm{d}W(t)\right]^2=\mathbf{E}\int_0^t X^2(t)\,\mathrm{d}t.$$

However, in this case the distribution of this stochastic integral is not necessarily normal.

18.4. **Relation between stochastic integral and quadratic variation.** The stochastic integral

$$X(t)=\int_0^t H(s)\,\mathrm{d}W(s),\ t\geq 0 \tag{52}$$

can be itself viewed as a random process. We can write it in the form

$$\mathrm{d}X(t)=H(t)\,\mathrm{d}W(t).$$

Its quadratic variation is equal to

$$\langle X\rangle_t=\int_0^t(\mathrm{d}X(s))^2=\int_0^t H^2(s)\,(\mathrm{d}W(s))^2=\int_0^t H^2(s)\mathrm{d}\langle W\rangle_s=\int_0^t H^2(s)\,\mathrm{d}s.$$

Alternatively, we can write this as

$$\mathrm{d}\langle X\rangle_t=H^2(t)\,\mathrm{d}t.$$

When X is a sum of a stochastic integral and a usual integral:

$$X(t)=X(0)+\int_0^t G(s)\,\mathrm{d}s+\int_0^t H(s)\,\mathrm{d}W(s), \tag{53}$$

we can rewrite this as $\mathrm{d}X(t)=G(t)\,\mathrm{d}t+H(t)\,\mathrm{d}W(t)$. Therefore,

$$\mathrm{d}\langle X\rangle_t=(\mathrm{d}X(t))^2=G^2(t)(\mathrm{d}t)^2+2G(t)H(t)\mathrm{d}t\mathrm{d}W(t)+H^2(t)\,\mathrm{d}t.$$

The term $(\mathrm{d}t)^2$ should be neglected in such calculation, as explained before. The term $\mathrm{d}t\mathrm{d}W(t)$ is of order $(\mathrm{d}t)^{3/2}$, because $\mathrm{d}W(t)=W(t+\mathrm{d}t)-W(t)\sim\mathcal{N}(0,\mathrm{d}t)$ is of order $(\mathrm{d}t)^{1/2}$. So it should also be neglected. At the end, we get: $\mathrm{d}\langle X\rangle_t=H^2(t)\,\mathrm{d}t$. The processes (52) and (53) have the same quadratic variation. Adding a usual (non-stochastic) integral does not influence the quadratic variation. Take expectation in (53):

$$\mathbf{E}(X(t)-X(0))=\int_0^t\mathbf{E}G(s)\,\mathrm{d}s.$$

In particular, if

$$\mathbf{E}X(t)=\mathbf{E}X(0)\quad\text{for allt}\geq 0,\quad\text{then}\quad\mathbf{E}G(s)=0\quad\text{for all}\quad s\geq 0. \tag{54}$$

18.5. **Itô's formula.** This is the main formula in stochastic calculus. Take a stochastic process $X=(X(t),\,t\geq 0)$ with quadratic variation $\langle X\rangle_t$. Consider a function $f:\mathbb{R}\to\mathbb{R}$. Apply f to $X(t)$ and get $Y(t)=f(X(t))$. Then

$$\mathrm{d}Y(t)=f'(X(t))\,\mathrm{d}X(t)+\frac{1}{2}f''(X(t))\,\mathrm{d}\langle X\rangle_t. \tag{55}$$

Let us explain this: use Taylor decomposition

$$f(y)-f(x)\approx f'(x)(y-x)+\frac{1}{2}f''(x)(y-x)^2\text{ for }y\approx x.$$

$$\begin{aligned}\mathrm{d}Y(t)=&Y(t+\mathrm{d}t)-Y(t)=f(X(t+\mathrm{d}t))-f(X(t))\\&=f'(X(t))(X(t+\mathrm{d}t)-X(t))+\frac{1}{2}f''(X(t))\,(X(t+\mathrm{d}t)-X(t))^2\\&=f'(X(t))\,\mathrm{d}X(t)+\frac{1}{2}f''(X(t))(\mathrm{d}X(t))^2=f'(X(t))\,\mathrm{d}X(t)+\frac{1}{2}f''(X(t))\mathrm{d}\langle X\rangle_t.\end{aligned}$$

Example 18.4. Let $f(x) = x^2$ and $X(t) = W(t)$. Then $f'(x) = 2x$ and $f''(x) = 2$. Also, $\langle W\rangle_t = t$. Therefore, applying (55), we get:

$$\mathrm{d}Y(t) = \mathrm{d}W^2(t) = 2W(t)\,\mathrm{d}W(t) + \mathrm{d}t.$$

We can write this as (because $W(0) = 0$):

$$W^2(t) = 2\int_0^t W(s)\,\mathrm{d}W(s) + t \;\Rightarrow\; \int_0^t W(s)\,\mathrm{d}W(s) = \frac{W^2(t) - t}{2}. \tag{56}$$

For smooth functions f with $f(0) = 0$, we have:

$$\int_0^t f(s)\,\mathrm{d}f(s) = \int_0^t f(s)f'(s)\,\mathrm{d}s = \frac{f^2(t)}{2}.$$

This additional term $-t/2$ is due to difference between ordinary and stochastic calculus. From this expression (56), we can immediately calculate the quadratic variation of $W^2(t)$:

$$\langle W^2\rangle_t = \int_0^t 4W^2(s)\,\mathrm{d}s.$$

For a function $f(t, x)$, Itô's formula takes form

$$\mathrm{d}f(t, X(t)) = \frac{\partial f}{\partial t}(t, X(t))\,\mathrm{d}t + \frac{\partial f}{\partial x}(t, X(t))\,\mathrm{d}X(t) + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(t, X(t))\mathrm{d}\langle X\rangle_t. \tag{57}$$

Example 18.5. Find quadratic variation of the process $tW(t)$. Apply Itô's formula to $f(t, x) = tx$ and $X(t) = W(t)$, with $\langle W\rangle_t = t$. We have:

$$\frac{\partial f}{\partial t} = x,\ \frac{\partial f}{\partial x} = t,\ \frac{\partial^2 f}{\partial x^2} = 0.$$

$$\mathrm{d}(tW(t)) = W(t)\,\mathrm{d}t + t\,\mathrm{d}W(t) \;\Rightarrow\; \langle (tW(t))\rangle_t = \int_0^t s^2\,\mathrm{d}s = \frac{t^3}{3}.$$

Let us introduce a *geometric Brownian motion* with drift g and diffusion σ^2, starting from x:

$$X(t) = xe^{gt+\sigma W(t)}.$$

Example 18.6. Apply Itô's formula to geometric Brownian motion $X(t)$, starting from $x = 3$, with parameters $g = 2$ and $\sigma = 3$, and to the function $f(x) = x^2 t$. We have:

$$X(t) = 3e^{2t+3W(t)} = e^{Y(t)},\ Y(t) := \ln 3 + 2t + 3W(t).$$

We need to solve this problem in three steps:

1. Find $\mathrm{d}X(t)$ and $\langle X\rangle_t$. To this end, apply Itô's formula to $g(x) := e^x$ and $Y(t)$. Then $g'(x) = g''(x) = e^x$, and $g'(Y(t)) = g''(Y(t)) = X(t)$. Also, $\mathrm{d}Y(t) = 2\mathrm{d}t + 3\mathrm{d}W(t)$, so $\langle Y\rangle_t = 9t$. Therefore,

$$\begin{aligned}\mathrm{d}X(t) &= \mathrm{d}g(Y(t)) = X(t)\,\mathrm{d}Y(t) + \frac{1}{2}X(t)\mathrm{d}\langle Y\rangle_t \\ &= X(t)\left(2\mathrm{d}t + 3\mathrm{d}W(t) + \frac{9}{2}\mathrm{d}t\right) = X(t)\left(\frac{13}{2}\mathrm{d}t + 3\mathrm{d}W(t)\right); \\ \langle X\rangle_t &= \int_0^t (9X^2(s))\,\mathrm{d}s \;\Rightarrow\; \mathrm{d}\langle X\rangle_t = 9X^2(t)\,\mathrm{d}t.\end{aligned}$$

2. Calculate

$$\frac{\partial f}{\partial t} = x^2,\ \frac{\partial f}{\partial x} = 2tx,\ \frac{\partial^2 f}{\partial x^2} = 2t.$$

3. Apply (57). We have:

$$\begin{aligned}\mathrm{d}f(t, X(t)) =& X^2(t)\,\mathrm{d}t + 2tX(t)\mathrm{d}X(t) + t\,\mathrm{d}\langle X\rangle_t \\ &= 2X^2(t)\,\mathrm{d}t + 13tX^2(t)\,\mathrm{d}t + 6tX^2(t)\,\mathrm{d}W(t) + 9tX^2(t)\,\mathrm{d}t \\ &= (2 + 22t)X^2(t)\,\mathrm{d}t + 6tX^2(t)\,\mathrm{d}W(t).\end{aligned}$$

18.6. **Hedging a European derivative.** Consider a European derivative $f(S(T))$, with

$$S(t) = S_0 e^{gt+\sigma W(t)}$$

is the Black-Scholes model of a stock price, T is the maturity, and $f(x)$ is a real-valued function. Assume at time t we constuct a portfolio: We split our wealth $V(t)$ at time t between $H(t)$ shares of the stock and $V(t) - H(t)S(t)$ in cash. Then the change in this wealth during time $[t, t + \mathrm{d}t]$ is equal to

$$H(t)(S(t + \mathrm{d}t) - S(t)) = H(t)\,\mathrm{d}S(t).$$

Thus $\mathrm{d}V(t) = H(t)\,\mathrm{d}S(t)$. We need $V(T) = f(S(T))$: at time T our wealth should exactly match the derivative. Then $V(0) = v$ would be the *fair price* at time $t = 0$, and $V(t)$ is the *fair price* at time t. The function $H(t)$ gives our *hedging strategy*, or *replicating portfolio*. Let us find $V(t)$ and $H(t)$. As discussed in the previous section, the fair price at time t given $S(t) = x$ is given by

$$u(t, x) = \mathbf{E}(f(S(T)) \mid S(t) = x),\ S(t) = S_0 \exp(\sigma W(t) - \sigma^2 t/2); \quad u(t, S(t)) = V(t). \tag{58}$$

Note that $\mathbf{E}u(t, S(t)) = \mathbf{E}f(S(T))$, because the expectation of a conditional expectation is always equal to the *unconditional expectation*. Therefore, $\mathbf{E}u(t, S(t))$ does not depend on t: It is constant. Now, decompose $u(t, S(t))$ according to Itô's formula:

$$\mathrm{d}u(t, S(t)) = \frac{\partial u}{\partial t}(t, S(t))\,\mathrm{d}t + \frac{\partial u}{\partial x}(t, S(t))\,\mathrm{d}S(t) + \frac{1}{2}\frac{\partial^2 u}{\partial x^2}(t, S(t))\,\mathrm{d}\langle S\rangle_t. \tag{59}$$

Apply Itô's formula with $f(t, x) = e^{\sigma x - \sigma^2 t/2}$ to $x = W(t)$. Then we have:

$$\frac{\partial f}{\partial t} = -\frac{\sigma^2}{2} f,\ \frac{\partial f}{\partial x} = \sigma f,\ \frac{\partial^2 f}{\partial x^2} = \sigma^2 f.$$

$$\begin{aligned}\mathrm{d}S(t) = \mathrm{d}f(t, W(t)) &= \left[\frac{\partial f}{\partial t}(t, S(t)) + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(t, S(t))\right]\mathrm{d}t + \frac{\partial f}{\partial x}(t, S(t))\,\mathrm{d}W(t)\\ &= \sigma f(t, S(t))\,\mathrm{d}W(t) = \sigma S(t)\,\mathrm{d}W(t).\end{aligned}$$

Therefore, $\mathrm{d}\langle S\rangle_t = \sigma^2 S^2(t)\,\mathrm{d}t$. Thus we can rewrite (59) as

$$\mathrm{d}u(t, S(t)) = \left[\frac{\partial u}{\partial t}(t, S(t)) + \frac{1}{2}\sigma^2 S^2(t)\frac{\partial^2 u}{\partial x^2}(t, S(t))\right]\mathrm{d}t + \frac{\partial u}{\partial x}(t, S(t)) \cdot \sigma S(t)\,\mathrm{d}W(t). \tag{60}$$

But we need that $u(t, S(t)) = V(t)$ satisfies $\mathrm{d}V(t) = H(t)\mathrm{d}S(t)$. Then the coefficient at $\mathrm{d}t$ is equal to 0:

$$\frac{\partial u}{\partial t} + \frac{1}{2}\sigma^2 x^2 \frac{\partial^2 u}{\partial x^2} = 0, \tag{61}$$

This is called the *Black-Scholes equation*. And $\frac{\partial u}{\partial x}(t, S(t)) = H(t)$. This is called *delta hedging:* The derivatives of $u(t, x)$ with respect to volatility σ, current price x, current time t, are called *delta*, *vega*, and *theta*, respectively. They measure sensitivity of the option price with respect to the parameters of the model. Collectively, they are called *"greeks"*. The *delta derivative* gives us the quantity of shares we need to have at time t in our portfolio.

Example 18.7. $f(x) = x^2$, $\sigma = T = 1$. Then $S(t) = e^{W(t)-t/2}$, and the fair price is given by

$$V(t) = u(t, S(t)) = \mathbf{E}\left[\left(e^{W(1)-1/2}\right)^2 \mid S(t)\right] = \mathbf{E}\left[e^{2W(1)-1} \mid W(t)\right] = e^{2W(t)-1}\mathbf{E}\,e^{2(W(1)-W(t))}.$$

In the last derivation, we used that $W(1) = W(t) + (W(1) - W(t))$, and $W(1) - W(t)$ is independent of $W(t)$. Continuing our calculations, we have: Since $W(1) - W(t) \sim \mathcal{N}(0, 1 - t)$, calculate the moment generating function:

$$V(t) = e^{2W(t)-1}e^{2^2(1-t)/2} = e^{2W(t)-1}e^{2(1-t)} = e^{2W(t)+1-2t}.$$

Now express this in terms of $S(t)$: $W(t) = t/2 + \ln S(t)$, and thus

$$V(t) = \exp\left(2(t/2 + \ln S(t)) + 1 - 2t\right) = \exp\left(2\ln S(t) + 1 - t\right) = S^2(t)e^{1-t}.$$

Therefore, the fair price function $u(t, x) = x^2 e^{1-t}$, and the delta hedging gives us:

$$\frac{\partial u}{\partial x} = 2xe^{1-t},\ H(t) = \frac{\partial u}{\partial x}(t, S(t)) = 2S(t)e^{1-t}.$$

As an example, if $S(0.5) = 1.4$, then at time $t = 0.5$ we need to have $2 \cdot 1.4 \cdot e^{1-0.5} = 4.616$ shares of this stock.

19. Stochastic Differential Equations

19.1. **Definition.** Take two functions $g, \sigma : \mathbb{R} \to \mathbb{R}$. An equation

$$\mathrm{d}X(t) = g(X(t))\,\mathrm{d}t + \sigma(X(t))\,\mathrm{d}W(t) \tag{62}$$

is called a *stochastic differential equation* (SDE) with *drift* g and *diffusion* σ. If we impose the *initial condition* $X(0) = x_0$, then this equation (under some conditions) has a unique solution. One can rewrite (62) together with this initial condition in the form

$$X(t) = x_0 + \int_0^t g(X(s))\,\mathrm{d}s + \int_0^t \sigma(X(s))\,\mathrm{d}W(s). \tag{63}$$

Example 19.1. If $g(x) \equiv g$ and $\sigma(x) \equiv \sigma$ are constant functions, then the equation (63) takes the form

$$X(t) = x_0 + \int_0^t g\,\mathrm{d}s + \int_0^t \sigma\,\mathrm{d}W(s) = x_0 + gt + \sigma W(t).$$

This is a Brownian motion with drift g and diffusion σ.

19.2. **Geometric Brownian motion.** Let $g(x) = g_0 x$ and $\sigma(x) = \sigma_0 x$ for some constants g_0 and σ_0. Then (62) takes the form

$$\mathrm{d}X(t) = X(t)\left[g_0\,\mathrm{d}t + \sigma_0\,\mathrm{d}W(t)\right]. \tag{64}$$

Let us show this is a geometric Brownian motion. From (64), we get:

$$\mathrm{d}\langle X\rangle_t = \sigma_0^2 X^2(t)\,\mathrm{d}t.$$

Apply Itô's formula to (64) with $f(x) = \log x$. Then

$$f'(x) = \frac{1}{x},\ f''(x) = -\frac{1}{x^2},$$

$$\begin{aligned}\mathrm{d}\log X(t) = \mathrm{d}f(X(t)) &= f'(X(t))\,\mathrm{d}X(t) + \frac{1}{2}f''(X(t))\,\mathrm{d}\langle X\rangle_t \\ &= \frac{1}{X(t)}\cdot X(t)\left[g_0\,\mathrm{d}t + \sigma_0\,\mathrm{d}W(t)\right] - \frac{1}{2X^2(t)}\sigma_0^2 X^2(t)\,\mathrm{d}t = \mu\,\mathrm{d}t + \sigma_0\mathrm{d}W(t),\end{aligned}$$

where $\mu = g_0 - \sigma_0^2/2$ is the drift. Therefore, $\log X(t) = \log X(0) + \mu t + \sigma_0 W(t) \ \Rightarrow\ X(t) = X(0)e^{\mu t + \sigma_0 W(t)}$.

Example 19.2. Let $\mathrm{d}X(t) = 4X(t)\,\mathrm{d}t + 0.5X(t)\,\mathrm{d}W(t)$, and $X(0) = 3$. Then $g_0 = 4$, $\sigma_0 = 0.5$, and $\mu = \frac{31}{8}$. Therefore,

$$X(t) = 3e^{\frac{31}{8}t + 0.5W(t)}.$$

Recall the moment generating functionfor normal random variables:

$$Z \sim \mathcal{N}(m, \rho^2) \ \Rightarrow\ \mathbf{E}e^{uZ} = \exp\left(mu + \frac{1}{2}\rho^2 u^2\right). \tag{65}$$

$$Z = \frac{31}{8}t + 0.5W(t) \ \Rightarrow\ m = \mathbf{E}Z = \frac{31}{8}t,\ \rho^2 = \operatorname{Var} Z = \operatorname{Var}(0.5W(t)) = 0.5^2 \operatorname{Var} W(t) = \frac{t}{4}. \tag{66}$$

Plugging this Z from (66) and $u = 1$ into (65), we get after calculation: $\mathbf{E}X(t) = 3\mathbf{E}e^Z = 3e^{4t}$. Similarly, plugging $u = 2$ into (66), we can find the second moment:

$$\mathbf{E}X^2(t) = 3^2\mathbf{E}e^{2Z} = 9\exp\left(\frac{33}{4}t\right).$$

19.3. **Ornstein-Uhlenbeck process.** Assume $g(x) = c(m - x)$ and $\sigma(x) = \sigma$ for $c, \sigma > 0$. Then the equation (62) takes the form

$$\mathrm{d}X(t) = c(m - X(t))\,\mathrm{d}t + \sigma\,\mathrm{d}W(t). \tag{67}$$

This has *mean-reverting property:* if $X(t) > m$, then X tends to move, on average, down to m; if $X(t) < m$, then X tends to move, on average, up to m. Therefore, this process tends to oscillate around m. Compare this with Brownian motion, which just diffuses (as a random walk) to infinity, without any limits. Or with geometric Brownian motion, which does the same, but on the logarithmic scale.

The equation (67) is an example of a *linear* SDE, which has drift $g(x)$ and diffusion $\sigma(x)$ to be linear in x.

19.4. **Linear equations.** An SDE is called *linear* if the drift and diffusion coefficients are linear functions. For such SDE, we can find their mean and variance. As an example, consider the process

$$\mathrm{d}X(t) = (X(t) - 2)\,\mathrm{d}t + 2X(t)\,\mathrm{d}W(t),\ X(0) = -3. \tag{68}$$

Find $m(t) := \mathbf{E}X(t)$. We have: $m(0) = -3$, and

$$X(t) = -3 + \int_0^t (X(s) - 2)\,\mathrm{d}s + 2\int_0^t X(s)\,\mathrm{d}W(s).$$

Taking the expectation, note that the stochastic integral has expectation zero, and

$$m(t) = -3 + \int_0^t (m(s) - 2)\,\mathrm{d}s \ \Rightarrow\ m'(t) = m(t) - 2,\ m(0) = -3.$$

Let us now solve this linear SDE using integrating factor. Because $m' = m \ \Rightarrow\ m(t) = Ce^t$, the integrating factor is $m(t) = e^{-t}$, and so

$$\begin{aligned} m' - m = -2 \ &\Rightarrow\ e^{-t}m' - e^{-t}m = -2e^{-t} \ \Rightarrow\ (e^{-t}m)' = -2e^{-t} \\ &\Rightarrow\ e^{-t}m = \int (-2e^{-t})\,\mathrm{d}t \ \Rightarrow\ e^{-t}m = 2e^{-t} + C \ \Rightarrow\ \boxed{m = 2 + Ce^t} \end{aligned}$$

Because $m(0) = -3$, we find $2 + C = -3 \ \Rightarrow\ C = -5$. Thus,

$$m(t) = 2 - 5e^t.$$

Now, apply Itô's formula to (68) with $f(x) = x^2$, $f'(x) = 2x$, $f''(x) = 2$:

$$\mathrm{d}X^2(t) = f'(X(t))\,\mathrm{d}X(t) + \frac{1}{2}f''(X(t))\,\mathrm{d}\langle X\rangle_t = 2X(t)\,\mathrm{d}X(t) + \mathrm{d}\langle X\rangle_t.$$

But $d\langle X\rangle_t = 4X^2(t)\,\mathrm{d}t$. Therefore,

$$\begin{aligned} \mathrm{d}X^2(t) =& 2X(t)\left[(X(t) - 2)\,\mathrm{d}t + 2X(t)\,\mathrm{d}W(t)\right] + 4X^2(t)\,\mathrm{d}t \\ =& \left[6X^2(t) - 4X(t)\right]\,\mathrm{d}t + 4X^2(t)\,\mathrm{d}W(t). \end{aligned}$$

We can rewrite this as

$$X^2(t) = X^2(0) + \int_0^t \left[6X^2(s) - 4X(s)\right]\,\mathrm{d}s + \int_0^t 4X^2(s)\,\mathrm{d}W(s).$$

Take expectation and use that $X(0) = -3$:

$$\mathbf{E}X^2(t) = (-3)^2 + \int_0^t \mathbf{E}\left[6X^2(s) - 4X(s)\right]\,\mathrm{d}s.$$

Denote $a(t) := \mathbf{E}X^2(t)$. Then

$$a(t) = 9 + \int_0^t \left[6a(s) - 4m(s)\right]\,\mathrm{d}s.$$

We can rewrite this as

$$a'(t) = 6a(t) - 4m(t) = 6a(t) - 4(2 - 5e^t),\ a(0) = 9.$$

Now it remains to solve this linear ODE. Using the same integrating factor method, we solve this equation.

19.5. **General remarks.** Every SDE is a continuous-time Markov process with state space $\mathbb{R}$, the whole real line, unlike Markov chains, which have state space $\{1, 2, 3\}$ or another finite or countable set. This Markov process has transition density $p(t, x, y)$: the density of $X(t)$ at point y given that it started from $X(0) = x$. We already know this transition density for Brownian motion. For the general SDE, this transition density is a solution of two PDEs, called *forward* and *backward Kolmogorov equations.*

Sometimes this SDE has a stationary distribution. For example, Brownian motion and geometric Brownian motion do not have a stationary distribution, but Ornstein-Uhlenbeck process does have one, and it is normal:

$$\mathcal{N}(m, \rho^2),\ \text{with } \rho^2 := \frac{\sigma^2}{2c}.$$

The rate of convergence to such stationary distribution is known for the Ornstein-Uhlenbeck process, but not for the general SDE.

20. Continuous-Time Martingales

The content of this section largely duplicates Section 13, adapting it for continuous time. We did not avoid some repetitions, doing this for the convenience of a reader.

20.1. **Definition and examples.** A process $M = (M(t),\, t \geq 0)$ is called a *martingale* if for every $0 \leq s \leq t$,

$$\mathbf{E}\left[M(t) \mid M(u),\, 0 \leq u \leq s\right] = M(s) \;\Leftrightarrow\; \mathbf{E}\left[M(t) - M(s) \mid M(u),\, 0 \leq u \leq s\right] = 0. \tag{69}$$

If in (69) we have $\geq$ or $\leq$ instead of equality, then M is called a *submartingale* or *supermartingale,* respectively. The equation (69) means that the best prediction of the future value at time t is the current value at time s. In particular, for a martingale M, we have $\mathbf{E}M(t) = \mathbf{E}M(0)$.

Example 20.1. A Brownian motion $W = (W(t),\, t \geq 0)$ is a martingale: Since $W(t) - W(s)$ is independent of $W(u)$, $0 \leq u \leq s$, and $W(t) - W(s) \sim \mathcal{N}(0, t-s)$, we have:

$$\begin{aligned}\mathbf{E}\left[W(t) \mid W(u),\, 0 \leq u \leq s\right] &= \mathbf{E}\left[W(t) - W(s) \mid W(u),\, 0 \leq u \leq s\right] + \mathbf{E}\left[W(s) \mid W(u),\, 0 \leq u \leq s\right] \\ &= \mathbf{E}\left[W(t) - W(s)\right] + W(s) = W(s).\end{aligned}$$

Example 20.2. Compensated Poisson process $M(t) = N(t) - \lambda t$, where $N(t)$ is a Poisson process with intensity λ. Then $N(t) - N(s) \sim \mathrm{Poi}(\lambda(t-s))$ is independent of $N(u)$, $0 \leq u \leq s$, and we can write $M(t) - M(s) = N(t) - N(s) - \lambda(t-s)$. Thus,

$$\begin{aligned}\mathbf{E}\left[M(t) - M(s) \mid M(u),\, 0 \leq u \leq s\right] &= \mathbf{E}\left[N(t) - N(s) \mid N(u),\, 0 \leq u \leq s\right] - \lambda(t-s) \\ &= \mathbf{E}\left[N(t) - N(s)\right] - \lambda(t-s) = 0.\end{aligned}$$

Note that in both of these examples, we used the property that $M(t) - M(s)$ is independent of $M(u)$, $0 \leq u \leq s$. Thus for every Lévy process $L(t)$, if $\mathbf{E}\left[L(t) - L(s)\right] = 0$, then this is a martingale. If $\mathbf{E}\left[L(t) - L(s)\right] \geq 0$, then L is a submartingale.

Example 20.3. Consider a Lévy process $L(t) = 1 + 2t + 2W(t) + \sum_{k=1}^{N(t)} Z_k$, with N a Poisson process with intensity $\lambda = 3$. Then it is a martingale if and only if $\mathbf{E}\left[L(t) - L(s)\right] = 2(t-s) + \lambda \cdot \mathbf{E}Z_1 = 0$.

We can consider a geometric Brownian motion $X(t) = X(0)\exp\left(gt + \sigma W(t)\right)$. When is it a martingale?

$$X(t) = X(s)\exp\left(g(t-s) + \sigma(W(t) - W(s))\right), \quad s < t.$$

And $W(t) - W(s)$ is independent of $X(u)$, $0 \leq u \leq s$. Therefore, we can write

$$\mathbf{E}\left[X(t) \mid X(u),\, 0 \leq u \leq s\right] = X(s)\mathbf{E}\exp\left(g(t-s) + \sigma(W(t) - W(s))\right) = X(s)\exp\left((t-s)(g + \sigma^2/2)\right).$$

Indeed, $W(t) - W(s) \sim \mathcal{N}(0, t-s)$ and therefore $\mathbf{E}\exp(\sigma(W(t) - W(s))) = \exp(\sigma^2(t-s)/2)$. Thus, this geometric Brownian motion is a martingale if and only if $g + \sigma^2/2 = 0$. More generally, take a Lévy process $L = (L(t),\, t \geq 0)$. Then $(e^{L(t)},\, t \geq 0)$ is a martingale if and only if

$$\mathbf{E}e^{L(t) - L(s)} = 1 \quad \text{for every } 0 \leq s < t.$$

Example 20.4. Consider $L(t) = W(t) + ct + \sum_{k=1}^{N(t)} Z_k$ with $Z_k \sim \mathcal{N}(0,1)$ and N having intensity 2. Then we can find the moment generating function, as in Section 17: Because $N(t) - N(s) \sim \mathrm{Poi}(2(t-s))$, and Z_k has moment generating function $e^{u^2/2}$

$$\begin{aligned}\mathbf{E}\left[e^{L(t) - L(s)}\right] &= e^{c(t-s)}\mathbf{E}\left[e^{W(t) - W(s)}\right]\mathbf{E}\left[\exp\left(\sum_{k=N(s)+1}^{N(t)} Z_k\right)\right] \\ &= e^{c(t-s)}\exp\left(\frac{1}{2}(t-s)\right)\exp\left(2(t-s)e^{1/2}\right) = \exp\left[\left(c + \frac{1}{2} + 2e^{1/2}\right)(t-s)\right].\end{aligned}$$

Therefore, we need $c = -0.5 - 2e^{1/2}$ for e^L to be a martingale.

For every random variable ξ and a process $X = (X(t),\, t \geq 0)$, the following conditional expectation is a martingale:

$$M(t) = \mathbf{E}\left[\xi \mid X(u),\, 0 \leq u \leq t\right].$$

Indeed, $M(t)$ is the best prediction of ξ at time t; and the best prediction of *this* at earlier time s is equal to the best prediction of ξ at time s. In particular, take a stock price $S = (S(t),\, t \geq 0)$. This last remark applies to the fair price process $V(t)$ of a European derivative $f(S(T))$ with maturity T. Indeed, this fair price, as shown in Section 17, is

$$V(t) = \mathbf{E}_*\left[f(S(T)) \mid S(u),\, 0 \leq u \leq t\right],$$

where $\mathbf{E}_*$ is the expectation taken with respect to the risk-neutral probability, under which S is a martingale.

Finally, let us mention that any stochastic integral with respect to the Brownian motion is a martingale:

$$M(t) = \int_0^t X(s)\,\mathrm{d}W(s),\ t \geq 0,$$

if only $X(t)$ is itself dependent on $W(u)$, $0 \leq u \leq t$. Indeed, fix $0 \leq s < t$ and split the interval $[s,t]$ into many small subintervals: $s = u_0 < u_1 < \ldots < u_N = t$. We can express the difference

$$M(t) - M(s) = \int_s^t X(u)\,\mathrm{d}W(u) \approx \sum_{k=0}^{N-1} X(u_k)\left(W(u_{k+1}) - W(u_k)\right). \tag{70}$$

If X is deterministic (non-random), then each summand in the right-hand side of (70) is an increment of the Brownian motion, which is independent of $W(u)$, $0 \leq u \leq s$ and has expectation zero. Therefore, its conditional expectation with respect to $M(u)$, $0 \leq u \leq s$, is zero. For stochastic (random) X, it is also possible to show, although not so easy.

20.2. **Optional stopping.** For every continuous-time martingale

$$M = (M(t),\ t \geq 0), \quad (M(\varepsilon n),\ n = 0, 1, 2, \ldots)$$

is a discrete-time martingale. And we can approximate a continuous-time martingale by such discrete-time martingales. Therefore, we can extend an optional stopping theorem from Section 13 for continuous-time martingales.

We say that τ is a *stopping time* with respect to a process $X = (X(t),\ t \geq 0)$, if for every $t \geq 0$, the event $\{\tau \leq t\}$ depends only on $X(u)$, $0 \leq u \leq t$. One can think about this as time to sell the stock, based only on its observed prices so far.

Example 20.5. Stopping times τ:

$$\inf\{t \geq 0 \mid X(t) \leq 0\}, \quad \inf\{t \geq 0 \mid X(t) \in [0,1]\},$$
$$\min(\inf\{t \geq 0 \mid X(t) = 0\}, 3), \quad \inf\{t \geq 0 \mid X(t-1) \geq 0\}$$

Example 20.6. These are *not* stopping times:

$$\tau = \inf\{t \geq 0 \mid X(t+1) \geq 0\}, \quad \tau = \sup\{t \geq 0 \mid X(t) = 0\}.$$

If $M(t)$ is a martingale, and τ is bounded, or $M(t)$, $0 \leq t \leq \tau$, is bounded, then

$$\mathbf{E}M(\tau) = \mathbf{E}M(t) \quad \text{for all} \quad t \geq 0.$$

The boundedness condition is essential. We can weaken it, but not get rid of it entirely. Indeed, consider a Brownian motion W starting with zero. Eventually it will hit 1, and so we let $\tau = \inf\{t \geq 0 \mid W(t) = 1\}$. Then $W(\tau) = 1$ but $W(0) = 0$. This does not contradict the optional stopping theorem, however, since τ is unbounded (we can wait for a very long time until we hit 1; in fact $\mathbf{E}\tau = \infty$), and $(W(t),\ 0 \leq t \leq \tau)$ is also unbounded (the Brownian motion can go arbitrarily far into the minus until it hits the level 1).

This corresponds to the *doubling strategy,* already discussed in Section 13: Play the coin until you end up with cumulative win. But this is only feasible if you are prepared to wait for unilimited time (unfeasible) and borrow arbitrarily much from someone (unfeasible). If you decide to wait only until time $T = 100$, then you get $\tau' := \min(\tau, 100)$, which is bounded. Then we have $\mathbf{E}W(\tau') = \mathbf{E}W(0) = 0$. Or if you decide to borrow up to 100, then you get $\tau'' := \inf\{t \geq 0 \mid W(t) \geq 1 \text{ or } W(t) \leq -100\}$. Then $W(t)$ is bounded for $t \leq \tau''$, and $\mathbf{E}W(\tau'') = \mathbf{E}W(0) = 0$.

20.3. **Jensen's inequality.** Recall from Section 13 that $g : \mathbb{R} \to \mathbb{R}$ is a *convex function* if

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y).$$

For example, $g(x) = x$ and $g(x) = x^2$ are convex functions, while $g(x) = \sin x$ is not. Equivalently, if you connect any two points on the graph of g by a segment, then it lies above the graph. For a twice differentiable function g, it is convex if and only if its second derivative is nonnegative: $g''(x) \geq 0$ for all x. Jensen's inequality says that if g is a convex function, then

$$\mathbf{E}g(Z) \geq g(\mathbf{E}Z). \tag{71}$$

One example of this is a well-known fact that $\mathbf{E}Z^2 \geq (\mathbf{E}Z)^2$. This is true, and $\mathbf{E}Z^2 - (\mathbf{E}Z)^2 = \operatorname{Var} Z \geq 0$. This immediately follows from (18): jsut apply $g(x) = x^2$.

Similarly, we can show (71) for conditional expectation instead of the unconditional:

$$\mathbf{E}\left[g(Z) \mid Y_1, \ldots, Y_n\right] \geq g\left(\mathbf{E}\left[Z \mid Y_1, \ldots, Y_n\right]\right). \tag{72}$$

20.4. **Transformation of martingales.** Take a martingale $M = (M(t), t \geq 0)$. Apply a convex function g to (69). By Jensen's inequality (72), we have:

$$\mathbf{E}\left[g(M(t)) \mid M(u),\, 0 \leq u \leq s\right] \geq g\left(\mathbf{E}\left[M(t) \mid M(u),\, 0 \leq u \leq s\right]\right) = g(M(t)).$$

Therefore, $g(X) = (g(M(t)),\, t \geq 0)$ is a submartingale.

Let us apply this to option pricing. We already discussed European options and other European derivatives in Section 17. Recall that a *call option* is the right to buy a stock at a certain *strike price* K. A European call option has *maturity time* T, when you can *exercise* this option: demand to buy this stock at price K. If the market price $S(T)$ of this stock at time T is less than K, then you can just buy the stock at the market price and forget about your option. Then your option does not have value. However, if the market price $S(T) \geq K$, then you should exercise this option, and its value is $S(T) - K$. In general, the value of this option is $(S(T) - K)_+ = g(S(T))$, where $g(x) = (x - K)_+$, $a_+ := \max(a, 0)$.

An American call option is different from a European one in the following way: the former can be exercised at any time until maturity N, while the latter must be exercised at maturity. Therefore, let τ be the time you decide to exercise your American call option to get the best expected value $\mathbf{E}g\left(S(\tau)\right)$. When is the best exercise time τ? This is a stopping time, since your decision to exercise at time n or not is based only on your observations until time n, but not on future observations. But the function g is convex (draw a graph and check). Frequently, the stock price X is modeled by a martingale. Then $g(S) = (g(S(t)),\, t \geq 0)$, is a submartingale. By the optional stopping theorem,

$$\mathbf{E}g\left(S(\tau)\right) \leq \mathbf{E}g\left(S(T)\right).$$

Therefore, the best time to exercise your American call option is at maturity $t = T$. In fact, American and European call options are of the same value in this case. Additional freedom to choose exercise time does not give you anything.

20.5. **Doob's inequalities.** These are generalizations of Markov's and Chebyshev's inequalities. Take a nonnegative submartingale $X = (X(t),\, t \geq 0)$ and a number $\lambda > 0$. Then

$$\mathbf{P}\left(\max_{0 \leq t \leq T} X(t) \geq \lambda\right) \leq \frac{\mathbf{E}X(T)}{\lambda}.$$

This is proved by approximating the continuous-time submartingale $(X(t),\, 0 \leq t \leq T)$ by discrete-time submartingales $(X(\varepsilon n),\ n = 0, 1, \ldots)$ for small $\varepsilon > 0$. In particular, applying a convex function $f(x) = x^2$ to a martingale $M = (M(t),\, t \geq 0)$, we get a submartingale $M^2(t)$. Therefore, we get *Kolmogorov's inequality:*

$$\mathbf{P}\left(\max_{0 \leq t \leq T} |M(t)| \geq \lambda\right) = \mathbf{P}\left(\max_{0 \leq t \leq T} M^2(t) \geq \lambda^2\right) \leq \frac{\mathbf{E}M^2(T)}{\lambda^2}.$$

Example 20.7. For a geometric Brownian motion $M(t) = e^{W(t) - t/2}$, which is a martingale, apply Kolmorogov's inequality:

$$\begin{aligned}\mathbf{P}\left(\max_{0 \leq t \leq 4} M(t) \geq 20\right) &\leq \frac{1}{20^2}\mathbf{E}\left[e^{W(4)-2}\right]^2 = \frac{1}{400}\mathbf{E}\left[e^{2W(4)-4}\right] \\ &= \frac{1}{400}\exp\left(4 \cdot 4/2 - 4\right) = \boxed{0.136}\end{aligned}$$

Example 20.8. In the previous example, we can also apply the function $f(x) = x^3$, which is convex on $x > 0$: $f''(x) = 6x > 0$. Thus

$$\begin{aligned}\mathbf{P}\left(\max_{0 \leq t \leq 4} M(t) \geq 20\right) &\leq \frac{1}{20^3}\mathbf{E}\left[e^{W(4)-2}\right]^3 = \frac{1}{8000}\mathbf{E}\left[e^{3W(4)-6}\right] \\ &= \frac{1}{8000}\exp\left(9 \cdot 4/2 - 6\right) > 1,\end{aligned}$$

so it is useless...

21. Appendix. Simulations

In this section, we outline simple ways to simulate random variables and random procecces. At the beginning of each subsection, we indicate to which section of the main theoretical text it corresponds.

21.1. **Bernoulli and binomial random variables.** We can simulate a Bernoulli random variable Z, which takes values 1 and 0 with probabilities p and $1-p$, as follows: Simulate $U \sim \text{Uni}[0,1]$. If $U \le p$, then $Z := 1$; else $Z := 0$. A binomial random variable $X \sim \text{Bin}(N,p)$ can be simulated as the sum of N i.i.d. Bernoulli random variables: $X = Z_1 + \ldots + Z_N$.

21.2. **Discrete random variables.** Consider a random variable X which takes values $0, 1, 2, \ldots$ with probabilities $p_0, p_1, p_2, \ldots$ Split $[0,1]$ into

$$[0, p_0],\ [p_0, p_0+p_1],\ [p_0+p_1, p_0+p_1+p_2],\ \ldots$$

Simulate $U \sim \text{Uni}[0,1]$ and consider these intervals one by one: If $0 \le U \le p_0$, then let $X := 0$; if $p_0 \le U \le p_0+p_1$, then let $X := 1$; etc.

21.3. **Joint discrete distributions.** Similarly we can simulate jointly distributed discrete random variables. As an example, simulate (X,Y) with joint distribution

X	Y	Prob.
0	0	0.5
0	1	0.3
1	0	0.2

Simulate $U \sim \text{Uni}[0,1]$. Next,

$$\begin{cases} 0 \le U \le 0.5 \ \Rightarrow X := 0,\ Y := 0; \\ 0.5 \le U \le 0.8 \ \Rightarrow \ X := 0,\ Y := 1; \\ 0.8 \le U \le 1 \ \Rightarrow \ X := 1,\ Y := 0. \end{cases}$$

21.4. **Continuous random variables: the inverse function method.** Consider a random variable X with cumulative distribution function $F(x) := \mathbf{P}(X \le x)$, with density $p(x) = F'(x)$. Let us simulate it. Construct the inverse function:

$$F^{-1}(u) := x, \quad \text{which solves} \quad F(x) = u.$$

Note that F is increasing, so $F^{-1}(u) \le x$ if and only if $u \le F(x)$. Take a uniform $[0,1]$ random variable U. Then we can take $X = F^{-1}(U)$, because it has the correct cumulative distribution function:

$$\mathbf{P}\left(F^{-1}(U) \le x\right) = \mathbf{P}\left(U \le F(x)\right) = F(x),$$

because for this uniform random variable U we have: $\mathbf{P}(U \le a) = a, \quad 0 \le a \le 1$.

Example 21.1. Try to simulate $X \sim \text{Exp}(\lambda)$. It has density $p(x) = \lambda e^{-\lambda x}$ and cumulative distribution function

$$F(x) = \int_0^x p(y)\mathrm{d}y = \int_0^x \lambda e^{-\lambda y}\mathrm{d}y = \left(-e^{-\lambda y}\right)\Big|_{y=0}^{y=x} = 1 - e^{-\lambda x}, \quad x \ge 0.$$

(And $F(x) = 0$ for $x \le 0$.) Now, let us find the inverse function. To this end, let us solve $F(x) = u$:

$$1 - e^{-\lambda x} = u \ \Rightarrow \ e^{-\lambda x} = 1 - u \ \Rightarrow \ x = -\lambda^{-1}\ln(1-u).$$

Now, let $\boxed{X = -\lambda^{-1}\ln(1-U)}$

21.5. **Rejection method.** Sometimes it is too difficult to simulate X using the inverse function method. For example, let $X \sim \mathcal{N}(0,1)$, then

$$F(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^x e^{-y^2/2}\mathrm{d}y,$$

which is difficult to calculate and even more difficult to invert. You can do this, using tables, but there is a better way. Let $f(x)$ be the density of X. In this case, it is

$$f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}.$$

It is bounded from above by $c = (2\pi)^{-1/2}$. Moreover, $-10 \le X \le 10$ with very high probability. Then we can simulate independent $X \sim \text{Uni}[-10,10]$ and $Y \sim \text{Uni}[0,c]$. The point (X,Y) is then uniformly distributed in the rectangle $[-10,10] \times [0,c]$. We *accept* X if this point lies below the density curve: that is, if $f(X) \ge Y$. Otherwise, we *reject* this point and repeat this again, until we get an acceptable point (X,Y). This method works equally well for multivariate distributions.

Example 21.2. Simulate (X,Y) with density $p(x,y) = x + y$ for $x, y \in [0,1]$. Then $p \le 2$. Simulate independent $X \sim \text{Uni}[0,1]$, $Y \sim \text{Uni}[0,1]$, $Z \sim \text{Uni}[0,2]$, and accept (X,Y) if $p(X,Y) = X + Y \ge Z$.

21.6. **Value at risk.** Assume we have i.i.d. claims $X_1, \dots, X_n$, and we would like to find value-at-risk corresponding to the level of confidence 95% by simulation. Simulate these claims, rank them from top to bottom, and find the $.05 \cdot n$-th ranked value. This is the value-at-risk for the current simulation. Repeat this a lot of times and average the resulting values-at-risk.

21.7. **Monte Carlo method.** Assume you have an integral $I = \int_0^1 f(x)\mathrm{d}x$, which you cannot calculate directly. One way to approximate this integral is to split $[0,1]$ into small subintervals and use Simpson's rule or other similar rules. Another way is to generate N i.i.d. uniform random variables $X_1, \dots, X_N$ on $[0,1]$ and let

$$I_N := \frac{1}{N}\sum_{k=1}^{N} f(X_k).$$

By Law of Large Numbers, because $\mathbf{E}f(X_k) = \int_0^1 f(x)\mathrm{d}x$, and $f(X_1), \dots, f(X_N)$ are i.i.d., we have: $I_N \to I$. More generally, if we have

$$I := \int_{-\infty}^{\infty} f(x)p(x)\mathrm{d}x,$$

where p is some probability density, let us generate N i.i.d. random variables $X_1, \dots, X_N$ with density p, and let

$$I_N := \frac{1}{N}\sum_{k=1}^{N} f(X_k) \to I, \quad N \to \infty,$$

because $\mathbf{E}f(X_k) = \int f(x)p(x)\mathrm{d}x = I$. This convergence is with speed $N^{-1/2}$, as follows from the Central Limit Theorem.

Example 21.3. Let $I = \int_{-\infty}^{\infty} e^{-x^4}\mathrm{d}x$. Generate $X_1, \dots, X_N \sim \mathcal{N}(0,1)$, and represent I as

$$I = \sqrt{2\pi}\int_{-\infty}^{\infty} f(x)p(x)\mathrm{d}x, \quad f(x) = e^{-x^4+x^2/2}, \quad p(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}.$$

Then we can approximate $I \approx I_N := \frac{1}{N}\sum\limits_{k=1}^{N} f(X_k)$.

This method works equally well for multiple integrals. This is a particular advantage of Monte Carlo, because the usual methods become slow when the dimension grows (*"curse of dimensionality"*).

Example 21.4. Consider the following integral on the 10-dimensional block $[0,1]^{10}$ in $\mathbb{R}^{10}$:

$$I = \int_0^1\int_0^1 \dots \int_0^1 e^{x_1^2+x_2^2+\dots+x_{10}^2}\mathrm{d}x_1\mathrm{d}x_2\dots\mathrm{d}x_{10}.$$

To split it into subblocks $[0, 0.1]^{10}$ and similar, you need 10^{10} blocks. But it is just as easy to use Monte Carlo for this case as for dimension one. Generate i.i.d. 10-dimensional vectors $U_i = (U_{i1}, \dots, U_{i10})$, with each component i.i.d. uniform on $[0,1]$. Then let

$$I_N := \frac{1}{N}\sum_{i=1}^{N} \exp\left(U_{i1}^2 + U_{i2}^2 + \dots + U_{i10}^2\right) \to I.$$

21.8. **Discrete-time Markov chains.** Consider the example with McDonalds-Subway. Let 0 and 1 correspond to McDonalds and Subway. Assume that the initial distribution is $x(0) = \begin{bmatrix}0.7 & 0.3\end{bmatrix}$. Let $U_0 \sim \mathrm{Uni}[0,1]$. If $U_0 \le 0.7$, we let $X_0 := 0$; otherwise $X_0 := 1$. Recall the transition matrix:

$$P = \begin{bmatrix}0.8 & 0.2\\ 0.1 & 0.9\end{bmatrix}$$

Simulate $U_1 \sim \mathrm{Uni}[0,1]$ independently of U_0. If $X_0 = 0$, let $X_1 := 0$ if $U_1 \le 0.8$, $X_1 := 1$ otherwise. If $X_0 = 1$, let $X_1 := 0$ if $U_1 \le 0.1$, and $X_1 := 1$ otherwise. If this Markov chain has more than two (say four) states, then we should consider four cases for U_1. In the same way, we simulate $X_2, X_3, \dots$

Assume the Markov chain is irreducible. Then the occupation time shares converge to the stationary distribution π. One can empirically calculate the stationary distribtuion: Fix a large number N of steps. For each state i, find the number M_i of times $n = 1, \dots, N$ when $X_n = i$. Then $p_i := M_i/N$ is an approximation for the stationary distribution π.

21.9. **Google's PageRank.** Take an oriented graph (edges are arrows), where states are Web pages and arrows are hyperlinks. Let us rank pages. A page is ranked highly if there are a lot of links pointing to it. However, the importance of each link is different. A link from an important Web page is also important. But if this important Web page provides lots of links, each of them is not so important. (A recommendation letter from Bill Gates carries more weight than the one from elementary school teacher. However, if Bill Gates is generous and writes thousands of letters every year, maybe his letter does not carry much weight after all.)

Let $r(i)$ be the rank of the ith Web page. It is called PageRank in honor of Larry Page (who founded Google with Sergey Brin as Ph.D. students at Stanford in 1998). Let $|i|$ be the number of links from page i. Then

$$r(i) = \sum_{j \to i} \frac{r(j)}{|j|}.$$

This is a mathematical formulation of the idea above. If we normalize $r(i)$: divide it by $R = \sum r(i)$, the sum over all existing Web pages, then we have a probability distribution $r(i)$ on the set of all Web pages. However, it is very hard to solve this system of equation explicitly. Brin and Page invented a Markov chain, which they called a *random crawler*: if at a certain moment it is at Web page i, it chooses randomly (uniformly) among the links from this page and moves along this link. This is, in effect, the random walk on a (oriented) graph. The transition matrix A of this Markov chain is given by

$$A_{ij} = \begin{cases} \frac{1}{|i|}, \ i \to j; \\ 0, \ \text{otherwise.} \end{cases}$$

But we can rewrite the equation for r above as

$$r(i) = \sum_{j \to i} r(j) A_{ji} = \sum_j r(j) A_{ji} \quad \Rightarrow \quad r = rA.$$

Therefore, r is a stationary distribution for A. If we run this random crawler long enough, it converges to r. (In fact, we need to make some adjustments to make sure this process indeed converges.)

21.10. **Random walk.** Consider a random walk with $p = 0.75$, starting from $S_0 = 1$. We simulate i.i.d. random variables $Z_1, Z_2, \ldots$ with distribution

$$\mathbf{P}(Z_k = 1) = .75, \ \mathbf{P}(Z_k = -1) = .25.$$

We do this by taking i.i.d. $U_1, U_2, \ldots \sim \text{Uni}[0,1]$ and letting $Z_k = 1$ if $U_k \le .75$, $Z_k = -1$ if $U > .75$. Then let

$$S_0 := 1, \ S_1 = S_0 + Z_1, \ S_2 = S_1 + Z_2, \ldots$$

21.11. **Poisson process.** Simulate i.i.d $T_1, T_2, \ldots \sim \text{Exp}(\lambda)$ and let

$$\tau_1 := T_1, \ \tau_2 := \tau_1 + T_2, \ \tau_3 := \tau_2 + T_3, \ldots$$

Then we find $N(t)$ by comparing t and τ_k until $t < \tau_k$. If k is the first such that $t < \tau_k$, then $N(t) = k - 1$. For example, if $\tau_3 \le t$, but $t < \tau_4$, then $N(t) = 3$. For a compound Poisson process

$$X(t) = \sum_{k=1}^{N(t)} Z_k,$$

we aslo need to simulate i.i.d. steps $Z_1, Z_2, \ldots$ and let $X(t) = Z_1 + \ldots + Z_{k-1}$ (and $X(t) = 0$ in case $k = 0$).

21.12. **Continuous-time Markov chains.** The difference from the discrete-time case is that we cannot simulate the value $X(t)$ for every time t. Instead, we do exactly as for the Poisson process: We simulate the jump times, and the values of this process at these jump times. Essentially, we switch from continuous to discrete time. As an example, consider a continuous-time Markov chain $X = (X(t), \, t \ge 0)$ with generator

$$A = \begin{bmatrix} -3 & 1 & 2 \\ 0 & -1 & 1 \\ 1 & 1 & -2 \end{bmatrix}$$

The corresponding discrete-time Markov chain $Y = (Y_n)_{n \ge 0}$ has transition matrix

$$P = \begin{bmatrix} 0 & 1/3 & 2/3 \\ 0 & 0 & 1 \\ 0.5 & 0.5 & 0 \end{bmatrix}$$

Simulate $Y_0 = X(0)$ as shown above. Assume $X(0) = Y_0 = 1$. Then the Markov chain will spend time $T_1 \sim \text{Exp}(3)$ in this state. After exiting this state, it will go to 2 with probability $1/3$ and to 3 with probability $2/3$. One can simulate Y_1 as shown in the subsection devoted to discrete-time Markov chains: Let $U_1 \sim \text{Uni}[0,1]$; if $U_1 \le 2/3$ then $Y_1 := 3$; otherwise $Y_1 := 2$. Assume $Y_1 = 3$. Then the Markov chain will spend time $T_2 \sim \text{Exp}(2)$ in this state. After exiting this state, it will go to 1 and 2 with equal probabilities .5. Let $\tau_1 := T_1$ and $\tau_2 := \tau_1 + T_2$. In this way we can simulate jump times τ_k and jump positions Y_k. For every fixed $t \ge 0$, find the k such that $\tau_k \le t < \tau_{k+1}$ and assign $X(t) := Y_k$.

Similarly to discrete-time Markov chains, we can estimate the stationary distribution for irreducible continuous-time Markov chains by calculating the share of time spent in each state over a long time interval. To this end, we fix a large number N of jumps. Take each state i, and find all $k = 1, \dots, N$ such that $Y_k = i$. Sum T_{k+1} for all such k, and divide the sum over N. The result will be p_i, and $\begin{bmatrix} p_1 & p_2 & \dots & p_N \end{bmatrix}$ is an approximation to the stationary distribution.

21.13. **Brownian motion.** It can be approximated by a discrete-time random walk: Take t_0 to be the step size. Then

$$W(kt_0) \approx \sqrt{t_0}\,(Z_1 + \dots + Z_k)\,,$$

where $Z_1, Z_2, \dots$ are i.i.d. taking values ± 1 with equal probability. This follows from the Central Limit Theorem: $\mathbf{E}Z_k = 0$ and $\operatorname{Var} Z_k = 1$, therefore

$$\frac{Z_1 + \dots + Z_k}{\sqrt{k}} \approx \mathcal{N}(0,1).$$

And $Z_1 + \dots + Z_k \approx \sqrt{k}\mathcal{N}(0,1) = \mathcal{N}(0,k)$. But $W(kt_0) \sim \mathcal{N}(0, t_0 k) = \mathcal{N}(0,k)\sqrt{t_0}$. Thus, the simulation method is as follows: take a small t_0, say $t_0 = .01$, and let $W((k+1)t_0) := W(kt_0) + \sqrt{t_0} Z_{k+1}$, with initial value $W(0) := 0$.

21.14. **Stochastic integrals.** We approximate them by taking a small step t_0, as in the previous subsection:

$$\int_0^2 W(s)\,\mathrm{d}W(s) \approx \sum_{k=0}^{2t_0-1} W(kt_0)\,(W((k+1)t_0) - W(kt_0))\,. \tag{73}$$

Having simulated the Brownian motion as in the previous subsection, we can then calculate the sum in (73).

21.15. **Stochastic differential equations.** Similarly to the Brownian motion, we take a small time step t_0:

$$\mathrm{d}X(t) = -3X(t)\,\mathrm{d}t + X(t)\,\mathrm{d}W(t),\ X(0) = 1. \tag{74}$$

We approximate the equation (74) as follows:

$$X((k+1)t_0) - X(kt_0) = -3X(kt_0)t_0 + X(kt_0)Z_{k+1},\ k = 0, 1, 2, \dots$$

where $Z_{k+1} := W((k+1)t_0) - W(kt_0)$ can be approximated by $\sqrt{t_0}Z_{k+1}$, with $Z_{k+1} = \pm 1$ with equal probabilities. Start from $X(0) = 1$, simulate Z_1 and calculate $X(t_0)$, then simulate Z_2 and calculate $X(2t_0)$, etc.

22. Simulation Problems

Problem 22.1. Simulate the random variables (X, Y) 10000 times, if they take values $(1,0)$, $(0,1)$, $(-1,0)$, $(0,-1)$. with equal probability 0.25. Calculate the empirical covariance $\operatorname{Cov}(X,Y)$, and compare with the theoretical value.

Problem 22.2. Simulate the die flip 10000 times, and find the empirical mean and variance.

Problem 22.3. Simulate a geometric random variable 10000 times with probability of success $p = 20\%$, by simulating a sequence of Bernoulli trials and then taking the number of the first successful trial. Compute empirical mean and variance, and compare with theoretical values.

Problem 22.4. Simulate 1000 times a Poisson random variable Poi(1.5) by splitting $[0,1]$ according to this distribution. Find its empirical mean, and compare with theoretical value. Simulate also $\text{Bin}(100, 0.15)$ by 100 Bernoulli trials at a time, repeat 1000 times. Compare empirical distributions with each other: find the difference

$$|\mathbf{P}(\text{Poi}(1.5) = k) - \mathbf{P}(\text{Bin}(100, 0.15) = k)|, \quad , k = 0, 1, 2, 3, 4.$$

For each distribution, find its empirical mean, and compare with theoretical value.

Problem 22.5. Simulate 10000 times the random variable X with density

$$p(x) = \frac{2}{\pi(1+x^2)},\ x \ge 0,$$

using the inverse function method. Calculate an empirical probability $\mathbf{P}(X \ge 2)$, compare with theoretical value.

Problem 22.6. Simulate 1000 times the Gamma variable $\Gamma(3, 2.6)$, using that it is the sum of three exponential random variables, which can be generated by the inverse function method. Compare the empirical and theoretical mean and variance.

Problem 22.7. Simulate 1000 times the Gamma variable $\Gamma(3, 2.6)$ using rejection method. Cut its tail at some large enough point. Compare the empirical and theoretical mean and variance.

Problem 22.8. Simulate the following jointly distributed random variables 10000 times using the rejection method:

$$p(x, y) = y, \ x \le y \le x + 1, \ 0 \le x \le 1.$$

Calculate empirical covariance, and compare with the theoretical result.

Problem 22.9. Using Monte Carlo method with 10000 tries, simulate the double integral

$$\int_0^1 \int_0^1 (x + y) \, \mathrm{d}x \, \mathrm{d}y.$$

Problem 22.10. Simulate 100 i.i.d. claims of size $Z \sim \mathrm{Exp}(1.5)$. Repeat this 1000 times. Find the value-at-risk for confidence level 95%. Compare with theoretical value.

Problem 22.11. Simulate 10000 steps of a Markov chain with transition matrix

$$P = \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.4 & 0.6 & 0 \\ 0.1 & 0.9 & 0 \end{bmatrix}$$

starting from 1. Find the stationary distribution by calculating the share of time spent in each state for this simulation. Compare with theoretical result.

Problem 22.12. Simulate 10000 steps of a Markov chain with transition matrix

$$P = \begin{bmatrix} 0.5 & 0.3 & 0.1 & 0.1 \\ 0.3 & 0.6 & 0.1 & 0 \\ 0.2 & 0.2 & 0 & 0.6 \\ 0 & 0.2 & 0.8 & 0 \end{bmatrix}$$

starting from 1. Find empirical probability of hitting 2 before 3. Compare with theoretical result.

Problem 22.13. Simulate 1000 times the random walk with $p = 0.3$ and $q = 0.7$, starting from 2. Find the empirical probability that it hits -1 before 4. Compare with theoretical value.

Problem 22.14. Simulate 10 steps of a random walk with $p = 0.6$ and $q = 0.4$, starting from -1. Repeat this simulation 10000 times. Find the empirical probability $\mathbf{P}(S_{10} = -3)$. Compare with theoretical value.

Problem 22.15. Simulate a stock price 1000 times given by geometric symmetric random walk, starting from $P_0 = 1$. Find the empirical fair value of the European option call at maturity $T = 30$ with strike $K = 2$.

Problem 22.16. Simulate 1000 times the first 50 jumps of a Poisson process N with intensity $\lambda = 2$. Calculate the empirical expectation $\mathbf{E}\,[N(1)N(2)]$, and compare it with the true value.

Problem 22.17. Simulate 1000 times the first 50 jumps of a compound Poisson process X with increments $\mathcal{N}(2.5, 4)$, and intensity $\lambda = 0.5$. Use this to find empirical value $\mathrm{Var}\, X(4)$ and compare this with the true value.

Problem 22.18. An insurance company receives claims during the year according to the Poisson process $N(t)$ with intensity $\lambda = 2$. Each claim is distributed as $\mathrm{Exp}(2.3)$. We measure time in days (365 days in a year). From 100 simulations, find the value-at-risk for confidence level 90%, and compare with theoretical value.

Problem 22.19. Until time $T = 100$, simulate a continuous-time Markov chain with generator

$$A = \begin{bmatrix} -3 & 1 & 2 \\ 2 & -5 & 3 \\ 5 & 0 & -5 \end{bmatrix}$$

starting from state 1. Find the stationary distribution by calculating the share of time spent in each state. Compare with the theoretical value.

Problem 22.20. Simulate 1000 times, starting from state 2, a continuous-time Markov chain with generator

$$A = \begin{bmatrix} -3 & 3 \\ 2 & -2 \end{bmatrix}$$

Find the empirical distribution of this chain at time $t = 2$. Compare with the theoretical value.

Problem 22.21. Simulate the M/M/1 queue with arrival intensity $\lambda = 2$ and service intensity $\mu = 4$, until time horizon $T = 100$. Find the empirical stationary distribution by calculating the share of time spent in each state. Compare this with the theoretical value.

Problem 22.22. Simulate a Brownian motion with step size .01 for time horizon $T = 10$. Repeat this 1000 times. Calculate the empirical probability $\mathbf{P}(W(10) > 2)$, and compare with true value.

Problem 22.23. Simulate a geometric Brownian motion X with drift $g = -.3$ and variance $\sigma^2 = 2$, starting from $X(0) = 0.3$, until for time horizon $T = 1$, with step size .01. Repeat this 1000 times. Calculate the empirical value of $\mathbf{E}X^2(1)$, and compare with the true value.

Problem 22.24. Find an empirical fair price of a *barrier option* which pays 1 if the price of a stock at some time until maturity $T = 2.5$ exceeds $K = 2$, if the initial price is $S_0 = 1.4$, and the volatility is $\sigma = 0.4$. Use step size 0.01 and 1000 simulations.

Problem 22.25. Simulate the following Itô integral 1000 times with step size 0.01:

$$\int_0^2 e^{W(s)}\,\mathrm{d}W(s).$$

Calculate the empirical variance and compare with the true result.

Problem 22.26. Simulate the following Ornstein-Uhlenbweck process:

$$\mathrm{d}X(t) = -2X(t)\,\mathrm{d}t + 3\,\mathrm{d}W(t),\ X(0) = -2,$$

with step 0.01, for time horizon $T = 1$. Repeat this 1000 times. Find empirical mean $\mathbf{E}X(1)$, and compare with the true value.

Cumulative Probabilities of the Standard Normal Distribution

The table gives the probabilities $\alpha = \Phi(z)$ to the left of given z–values for the standard normal distribution.

For example, the probability that a standard normal random variable Z is less than 1.53 is found at the intersection of the 1.5 rows and the 0.03 column, thus $\Phi(1.53) = P(Z \leq 1.53) = 0.9370$. Due to symmetry it holds $\Phi(-z) = 1 - \Phi(z)$ for all z.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Quantiles of the Standard Normal Distribution

For selected probabilities α, the table shows the values of the quantiles z_α such that $\Phi(z_\alpha) = P(Z \leq z_\alpha) = \alpha$, where Z is a standard normal random variable.

The quantiles satisfy the relation $z_{1-\alpha} = -z_\alpha$.

α	0.9	0.95	0.975	0.99	0.995	0.999
z_α	1.282	1.645	1.960	2.326	2.576	3.090