Probability-I

Spring 2024

Soumya Bhattacharya

Literature used in the preparation of these notes:

S. Bhattacharya, *The Probabilistic Pigeonhole Principle*, The American Mathematical Monthly, Vol. 130, No. 7, p. 678, 2023

G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, eds. D. R. Heath-Brown, J. H. Silverman, Oxford University Press, 2008

S. Ross, *A first course in Probability*, Pearson Education, 2002

`https://en.wikipedia.org/`

If you have any comments about these lecture notes, please send them to
soumya.bhattacharya@iiserkol.ac.in

Life is either a daring random experiment or nothing at all.*

*apologies to Helen Keller.

Contents

Probability Spaces

Probability is a measure of how likely an event is to occur. This subject has a wide range of applications in every aspect of human life. The first use of probabilistic notions occurred in Cardano's *Book on games of chance*[*] (1663). Also, *The art of speculating*[**] (1713) by Bernoulli and *The doctrine of chances*[†] (1718) by De Moivre had immense impacts on the development of the Theory of Probability. Initially, people studying probability were concerned about the actions which could result in only finitely many outcomes. So, the probability of obtaining a particular set of outcomes was defined as

(0.1) *the proportion of the required outcomes among all the possible outcomes*

of the action. The first attempt to extend probabilistic ideas beyond finite sets was in Laplace's *Analytical Theory of Probability*[‡] (1812). Much later, the modern axioms of Probability were introduced by Kolmogorov in his book *Foundations of the Theory of Probability*[§] (1933).

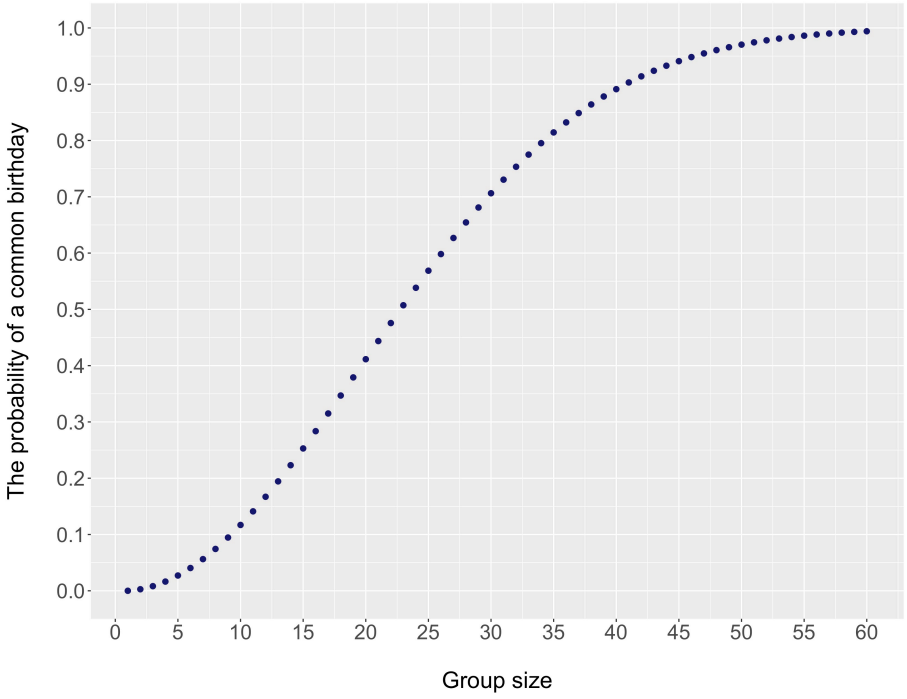
1 The probabilistic pigeonhole principle

The Pigeonhole principle asserts that if you put n pigeons into $m < n$ pigeonholes, then at least two pigeons would be in the same hole.

You may wonder what happens if there are more pigeonholes than pigeons. In any case, if the pigeons are allocated to the pigeonholes indiscriminately, regardless of the number of occupants of any pigeonhole, then it is still a possibility that two or more pigeons end up in the same hole. The *Birthday problem* is only a special

[*] in Italian
[**] in Latin
[†] in English
[‡] in French
[§] in German

case of this which asks for the probability* that in a group of n randomly chosen people, at least two shares a birthday. The apparent paradox is that the probability of two persons sharing a birthday exceeds 0.5 if the group size is at least 23. In fact, if a group has at least 57 members, the probability that at least two of the members have a common birthday is more than 0.99.



THEOREM 1 (PROBABILISTIC PIGEONHOLE PRINCIPLE). *Given $m \in \mathbb{N}$ and $p \in [0, 1)$, let $n \in \mathbb{N}$ be larger than or equal to*

$$\frac{1}{2} + \sqrt{2m \log\left(\frac{1}{1-p}\right) + \frac{1}{4}}. \tag{1.1}$$

If n pigeons are placed in m pigeonholes randomly, then the probability that at least two pigeons would be in the same hole is greater than p.

PROOF. We require to to show that for an integer n which is larger than the quantity in (1.1), if n pigeons are placed in m pigeonholes randomly, then the probability of an overlap in the placement of the pigeons is greater than p. From (0.1), it follows that the probability of an overlap in the placement is greater than p if and only if the probability of no overlap in the placement of the pigeons is less than $1 - p$. Let us compute this probability: Since every pigeon could be placed in any of the m pigeonholes, the total number of ways in which n pigeons could be placed in m holes is m^n. Whereas, the number of ways in which n pigeons could be placed in m holes with no overlaps is ${}^m\mathrm{P}_n$. Now, (0.1) implies that the required probability is the proportion of the placements of the pigeons in distinct pigeonholes among all possible placements of n pigeons in m pigeonholes. This

*For now, we take (0.1) as the definition of Probability.

proportion is ${}^m\mathrm{P}_n/m^n$. Hence, it suffices to show that if the integer n is greater than or equal to the quantity in (1.1), then ${}^m\mathrm{P}_n/m^n$ is less than $1-p$.

Let n be greater than or equal to the quantity in (1.1). Then after subtracting $1/2$ from both sides, squaring and dividing by $2m$, we obtain

$$\frac{n(n-1)}{2m} > \log\Big(\frac{1}{1-p}\Big),$$

which implies that

$$-\sum_{j=1}^{n-1}\log\big(1-\frac{j}{m}\big) = \sum_{j=1}^{n-1}\sum_{\ell=1}^{\infty}\frac{j^\ell}{\ell m^\ell} > \sum_{j=1}^{n-1}\frac{j}{m} = \frac{n(n-1)}{2m} > \log\Big(\frac{1}{1-p}\Big).$$

After multiplying both sides by -1 and exponentiating, we obtain

$$\prod_{j=1}^{n-1}\big(1-\frac{j}{m}\big) < 1-p.$$

Since the left hand side of the above inequality is ${}^m\mathrm{P}_n/m^n$, the claim follows. □

Instead of using the logarithmic inequality as above, one may also use the fact that $e^{-x} \geq 1-x$ for all $x \in \mathbb{R}$ to prove Theorem 1.

COROLLARY 1. *If $3n+1$ or more pigeons are placed in n^2 pigeonholes randomly, then the probability that at least two pigeons would be in the same hole is greater than 0.98889.*

PROOF. Putting $m = n^2$ and $p = 0.98889$ in (0.1), we obtain a quantity that is smaller than

$$\frac{1}{2} + \sqrt{9n^2 + \frac{1}{4}},$$

which is less than $3n+1$. □

2 THE LONGEST RUN IN A SEQUENCE OF TOSSES

During the last quarter of the twentieth century, it was a common practice* among probability teachers around world to demostrate a probabilistic method of lie detection: They used to give their students the homework of tossing a coin a hundred times** and noting down the sequence of outcomes. Quite predictably, it was also a common practice among the students to evade this boring homework and make an attempt to dupe the teacher by writing down a sequence of heads and tails which they thought are sufficiently *random*. As in general, human intuition of *randomness* is rather poor, in almost all cases these made-up sequences did not contain any run of heads or tails of length larger than five or six. In the next class,

*See this article: `https://www.maa.org/sites/default/files/images/upload_library/22/Polya/07468342.di020742.02p0021g.pdf`

**Or even a thousand times, as recollected by Prof. Bimal Roy from the B. Stat.(Hons.) batch of 1978 of ISI Kolkata. Their teacher Prof. Jogabrata Roy gave them this homework in their first class of Probability.

the teacher used to ask how many students obtained a run of eight* or more heads, which was often answered in the negative. Then he used to tell the students to compute the probability of their claim to see how unlikely it is! Let's demostrate how to compute such probabilities** with a small example:

EXAMPLE 1. Find the probability of obtaining no two consecutive heads when a fair coin is tossed seven times.

ANSWER. For $n \geq 1$, let F_n denote the number of outcomes with no two consecutive heads when a fair coin is tossed n times. Then we have

$$F_1 = |\{T, H\}| = 2,$$

$$F_2 = |\{TT, HT, TH\}| = 3$$

and for $n > 2$,

$$F_n = F_{n-1} + F_{n-2},$$

because, a sequence of $n \geq 2$ tosses without two consecutive heads ends either in T or TH. It follows from above that $F_7 = 34$. Since the total number of possible outcomes in 7 tosses (assuming that the coin does not land on its side) is 2^7. So, according to (0.1), the required probability is $F_7/2^7 = 17/64$.

3 THE RIEMANN HYPOTHESIS VIA PROBABILITY

As a brief motivation for the students of probability, below we describe the Riemann Hypothesis† (1859) in probabilistic terms.

DEFINITION 1 (PRIMORIAL). The k-th primorial is the product of the first k primes.

DEFINITION 2 (EULER–MASCHERONI CONSTANT). The limiting value γ of the difference between $\displaystyle\sum_{k=1}^{m} \frac{1}{k}$ and $\log m$ as m tends to infinity. Its first few digits are

$$0.57721566490153286060651209008240243104215933593992359880576 72\ldots.$$

The Riemann Hypothesis is equivalent to the claim that if an integer n is chosen at random from $\{1, 2, \ldots, N_k\}$, where N_k denotes the k-th primorial, then the probability of n being coprime to N_k is less than

$$\frac{1}{e^{\gamma} \log\log N_k}$$

for all $k \in \mathbb{N}$.

*Of course, *eight* could be replaced by a larger number, depending on the class size.

**Here we compute the probability only for one sequence of a given length. In a class, the same experiment is repeated as many times as the number of students in the class. So, the probability that nobody obtains a run of heads of length eight or more in hundred tosses reduces significantly when the class size is big.

†A 165 years old open conjecture which claims that the real part of any the nontivial zero of the Riemann Zeta function is $1/2$.

DEFINITION 3 (EXPERIMENT). An experiment is an act that can be repeated under similar conditions.

DEFINITION 4 (SAMPLE SPACE). The set Ω of all possible outcomes of an experiment is called its sample space.

DEFINITION 5 (SET OF EVENTS AND THE AXIOMS OF PROBABILITY). The set of events is a subset $\mathcal{E}$ of the power set of the sample space Ω such that

(a) $\Omega \in \mathcal{E}$.
(b) If $A \in \mathcal{E}$, then $A^c \in \mathcal{E}$.
(c) The set $\mathcal{E}$ is closed under countable unions.
(d) There exists a function $P : \mathcal{E} \to [0,1]$ such that $P(\Omega) = 1$. and for pairwise disjoint events $\{A_n\}_{n=1}^{\infty}$ with $A = \underset{n}{\cup} A_n$,

$$P(A) = \sum_{n=1}^{\infty} P(A_n).$$

In particular, for a finite sample space, we may always assume the set of events to be identical with the power set of the sample space.

DEFINITION 6 (PROBABILITY SPACE). The sample space Ω of an experiment together with the set of events $\mathcal{E}$ and the probability function $P : \mathcal{E} \to [0,1]$, is called a probability space and it is denoted by the triplet $(\Omega, \mathcal{E}, P)$.

EXAMPLE 2 (GAMES OF CHANCES).

(i) Tossing a coin: $\Omega = \{H, T\}$.
(ii) Rolling a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$.
(iii) Guessing the first letter of a stranger's name: $\Omega = \{A, B, C, \ldots, Z\}$.
(iv) Guessing someone's palm temperature in Fahrenheit: $\Omega = [55, 115]$.
(v) Guessing the number of stars in the Milky Way: $\Omega := \{1, 2, 3, \ldots\}$.

DEFINITION 7 (MUTUALLY EXCLUSIVE EVENTS). If two events $A, B \in \mathcal{E}$ is such that $A \cap B = \emptyset$, then A and B are called mutually exclusive events.

Let $(\Omega, \mathcal{E}, P)$ be a probability space. If A and $B \in \mathcal{E}$ are mutually exclusive events, then Definition 5.(d) implies that $P(A \cup B) = P(A) + P(B)$.

However, if $A \cap B \neq \emptyset$, then we may write $A = (A \cap B) \cup (A \cap B^c)$, where $B^c := \Omega \setminus B$. Since $A \cap B$ and $A \cap B^c$ are mutually exclusive, it follows from Definition 5.(d) that

$$P(A) = P(A \cap B) + P(A \cap B^c). \tag{4.1}$$

Since $A \cup B = (A \cap B^c) \cup B$ and since $A \cap B^c$ and B are mutually exclusive,

$$P(A \cup B) = P(A \cap B^c) + P(B). \tag{4.2}$$

From (4.1) and (4.2), it follows that

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{4.3}$$

By induction, the above equality can be generalized to any finite collection of of events.

DEFINITION 8 (EXHAUSTIVE EVENTS). The events in a set $S \subseteq E$ are called exhaustive if $\underset{A \in S}{\cup} A = \Omega$, i.e. the entire sample space.

EXAMPLE 3 (COUNTABLE SAMPLE SPACES). Let $\Omega = \{\omega_n\}_{n=1}^{\infty}$. If $\mathcal{E} = 2^{\Omega}$, then in particular, $\{\omega_n\} \subseteq \Omega$ for all $n \in \mathbb{N}$. Since for $n \in \mathbb{N}$, the events $\{\omega_n\}$ are pairwise mutually exclusive and exhaustive, it follows from the axioms of probability that

$$\sum_{n=1}^{\infty} P(\omega_n) = P(\Omega) = 1.$$

EXAMPLE 4. Show that if you keep on tossing a coin where the probability of obtaining a head is $p > 0$, then the probability of eventual occurrence of at least one head is 1.

PROOF. Consider the random experiment, where you keep on tossing the coin till you get a head. The corresponding sample space is

$$\Omega := \{H, TH, TTH, TTTH, \ldots\} \cup \{TTTTT \ldots \infty\}$$

and the set of events is $\mathcal{E} = 2^{\Omega}$, where

$$P(H) = p,\ P(TH) = (1-p)p,\ P(TTH) = (1-p)^2 p, \ldots$$

Since probability of eventually obtaining a head is

$$\sum_{n=0}^{\infty} (1-p)^n p = \frac{p}{1-(1-p)} = 1,$$

the claim follows. □

Since the complement of the desired event (i.e. eventually obtaining a head) is $\{TTTTT \ldots \infty\}$, showing that

$$P(TTTTT \ldots \infty) = \lim_{n \to \infty} (1-p)^n = 0$$

would have also sufficed for Example 4 (see Exercise 10.i).

DEFINITION 9 (EQUALLY LIKELY). The events in a set $S \subseteq E$ are called equally likely if $P(A) = P(B)$ for all $A, B \in S$.

THEOREM 2. *All the outcomes in a countably infinite sample space can not be equally likely.*

PROOF. Let $\Omega := \{\omega_n\}_{n=1}^{\infty}$. Suppose, each of the outcomes ω_n has an equal probability p of occurrence. Since the events $\{\omega_n\}$ are mutually exclusive and exhaustive, it follows that

$$1 = P(\Omega) = \sum_{n=1}^{\infty} P(\omega_n) = \sum_{n=1}^{\infty} p,$$

which is absurd, since the right hand side does not converge if $p > 0$, whereas if $p = 0$, then the right hand side converges to zero! □

5 Earlier attempts at defining probability

Definition 10 (Random experiment). If the sample space of an experiment is known but none of the outcomes of the experiment occurs with certainty, then we call the experiment a random experiment.

In particular, all the games of chances mentioned in Example 2 are random experiments.

Example 5 (Examples of non-random experiments). All the experiments that verify certain physical/chemical/biological laws are not random experiments*.

Definition 11 (Classical definition of probability). If a random experiment results in m mutually exclusive, exhaustive and equally likely outcomes, of which exactly $n(A)$ outcomes are favourable for an event $A \in \mathcal{E}$, then the probability of the event A, denoted by $P(A)$ is given by

$$P(A) = \frac{n(A)}{m}.$$

Indeed, the last definition is meant only for finite sample spaces. However, an inherent flaw in the last definition of probability is that it is cyclic, since in order to define *equally likely* events, we have already used the notion of probability.

Definition 12 (Frequency definition of probability). Let a random experiment be repeated n times, in which the frequency of the event A was $f_n(A)$, i.e. the event A occurred exactly $f_n(A)$ times. The ratio $f_n(A)/n$ is called the relative frequency of the event A and the probability $P(A)$ is defined by the limit

$$P(A) = \lim_{n\to\infty} \frac{f_n(A)}{n}.$$

An inherent drawback of the above definition of probability is that the existence of the above limit can not be proved in any case. However, later we shall see that the law of large numbers has its roots in this idea.

6 Boole's and Bonferroni's inequalities

Theorem 3 (Boole's inequality). *Let $(\Omega, \mathcal{E}, P)$ be a probability space. Then for $A_1, A_2, \ldots A_n \in \mathcal{E}$, we have*

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) \leq P(A_1) + P(A_2) + \cdots + P(A_n).$$

Proof. For $i \in \{1, 2, \ldots, n\}$, define

$$B_i := \begin{cases} A_1 & \text{if } i = 1 \\ A_i \setminus \bigcup_{j=1}^{i-1} A_j. & \text{if } i > 1. \end{cases}$$

*However, if the range of the possible errors in such an experiment is nonempty and is known a priori, then we may also view such an experiment as a random experiment, since no particular error occurs with certainty.

By construction, the events B_i for $i \in \{1, 2, \ldots, n\}$ are disjoint. Therefore,

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(B_1 \cup B_2 \cup \ldots B_n) = \sum_{i=1}^{n} P(B_i) \leq \sum_{i=1}^{n} P(A_i),$$

where the second equality follows from Definition 5.(d) and the last inequality follows from Exercise 10.ii. □

THEOREM 4 (BONFERRONI'S INEQUALITY). *Let $(\Omega, \mathcal{E}, P)$ be a probability space. Then for $A_1, A_2, \ldots A_n \in \mathcal{E}$, we have*

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) \geq P(A_1) + P(A_2) + \cdots + P(A_n) - (n-1).$$

PROOF. We proceed to prove the claim by induction. Note that since $P(A_1 \cup A_2) \leq 1$, for $n = 2$ the claim follows trivially from (4.3). Suppose, the claim holds for an integer $m \geq 2$. Then

$$\begin{aligned} P(A_1 \cap A_2 \cap \ldots \cap A_m \cap A_{m+1}) &\geq P(A_1 \cap A_2 \cap \ldots \cap A_m) + P(A_{m+1}) - 1 \\ &\geq \sum_{i=1}^{m} P(A_i) - (m-1) + P(A_{m+1}) - 1 \\ &= \sum_{i=1}^{m+1} P(A_i) - m, \end{aligned}$$

where the first inequality holds since the claim is true for two events, whereas the second inequality is implied by the induction hypothesis. Thus, the claim follows. □

Exercises

1. Show that the probability of at least two persons sharing a birthday among a group of n persons is
$$1 - \frac{^{365}\mathrm{P}_n}{365^n} - \frac{97 \times {}^{365}\mathrm{P}_{n-1}}{146097 \times 365^{n-1}}.$$

2. Show that for any sequence $\{a_n\}$ of positive reals which diverges to infinity, if $\lfloor a_n \sqrt{n} \rfloor$ pigeons are placed in n pigeonholes, then the probability that at least two pigeons are in the same hole tends to 1 as $n \to \infty$.

3. If you toss a fair coin 20 times, find the probability of obtaining 10 heads and 10 tails.

4. Compute the probability of obtaining a total of 10 points when two unbiased dice are rolled.

5. (i) Find the probability of obtaining no four consecutive heads when a fair coin is tossed ten times.
(ii) Find the probability of obtaining neither four consecutive heads nor four consecutive tails when a fair coin is tossed ten times.

6. Let x be a point inside a convex quadrilateral $\mathcal{Q}$. Find the probability that x is neither on the boundary nor inside any of the circles drawn with the sides of the quadrilateral Q as their diameters.

7. (Bertrand's Paradox) Explain why there can't be a single answer to the question that asks for the probability of choosing a chord randomly of length less than r in a circle of radius r.

8. Let $(\Omega, \mathcal{E}, P)$ be a probability space. Show that $\mathcal{E}$ is closed under countable intersections.

9. Let $(\Omega, \mathcal{E}, P)$ be a probability space. Let $A, B \in \mathcal{E}$. Show that

$$P(B \cap A^c) = P(B) - P(B \cap A).$$

10. Let Suppose $(\Omega, \mathcal{E}, P)$ be a probability space and $A, B \in \mathcal{E}$. Using the axiomatic definition of probability, prove the following statements:
    i) $P(A^c) = 1 - P(A)$.
    ii) If $A \subseteq B$, then $P(A) \leq P(B)$.

11. Let $(\Omega, \mathcal{E}, P)$ be a probability space and $A_1, A_2, \ldots, A_n \in \mathcal{E}$. Show that $P(A_1 \cup A_2 \cup \ldots \cup A_n)$ is equal to

$$\sum_{i=1}^{n} P(A_i) - \sum_{1 \leq i < j \leq n} P(A_i \cap A_j) + \sum_{1 \leq i < j < k \leq n} P(A_i \cap A_j \cap A_k) - \ldots + (-1)^{n-1} P(A_1 \cap \ldots \cap A_n).$$

12. Show that if you keep on throwing a fair die, the probability of eventual occurence of the outcome 6 is 1.

13. Dropping two points uniformly at random on $[0, 1]$, the unit interval is divided into three segments. Find the probability that the three segments obtained in this way form a triangle.

14. Find the probability of obtaining no two consecutive heads when a fair coin is tossed n times.

15. Find the probability of obtaining a strictly increasing sequence of integers if you pick 3 integers (one at a time) (a) with replacement or (b) without replacement from $\{1, 2, 3, \ldots, 1000\}$.

16. To the choice of each $n \in \mathbb{N}$, could you assign a probability $P(n) > 0$ such that the following conditions hold?
    (1) $P(m) \neq P(n)$ for all $m, n \in \mathbb{N}$, $m \neq n$.

(2) The probability of choosing an odd positive integer is the same as the probability of choosing an even positive integer.

17. Seven students of IISER Kolkata went to participate in an event at IISER Mohali. They booked AC 3-tier tickets from Howrah to Chandigarh in Netaji Express, which has three AC 3-tier coaches. Every such coach has eight coupes (i.e. compartments), each coupe containing eight berths. If the berths were allocated randomly, find the probability of at least two among the seven students being allocated berths in the same coupe.

18. A point is selected at random inside an equilateral triangle whose side length is 3. Find the probability that the distance of the chosen point from any vertex of the traingle is greater than 1.

19. On a rainy day, 4 students went for their coaching classes with nearly identical raincoats. They put their raincoats at the same place before going to the class. Find the probability that none of the students selects his/her own raincoat after the class.

20. Suppose, the planetary system of a star contains ten planets with coplanar orbits and suppose the distance from the star to the farthest point(s) on the orbit of the outermost planet is 1.5 astronomical units. Find the probability that at an arbitrary instant of time, at least two of these planets are less than or equal to $\sqrt{2}$ astronomical units apart from each other.

21. Suppose, the planetary system of a star contains twenty-five planets with at least two planets having non-coplanar orbits and suppose the distance from the star to the farthest point(s) on the orbit of the outermost planet is 2 astronomical units. Show that the probability of at least two of these planets being less than or equal to $\sqrt{3}$ astronomical units apart from each other at an arbitrary instant of time is greater than 0.99.

22. Carrom is played with a red, nine black and nine white coins (and a striker) on a square board with a pocket in each corner. Suppose, we have a 29 inch $\times$ 29 inch carrom board (excluding the raised borders) and suppose, the diameter of all the coins are 3/5 inches. If all these coins are scattered randomly (none of them being in any pocket) on it, show that the probability of at least two of the coins being less than two inches apart is greater than $4/5$.

Conditional Probability

There's a story of a statistician who told a friend that he never took an aeroplane. Because, he computed the probability that there will be a bomb on the plane. Although this probability was low, it was still too high for his comfort.

Two weeks later, to her amazement, the same friend discovered the statistician onboard in her aeroplane. "Well, well, well.. here I see a statistician who changes his theory every two weeks!", She exclaimed. "No, no, I haven't changed my theory at all!", he replied in a rather serious voice."Then how did you gather the courage to board this flight?", she teased, without realizing that there's little point in teasing a nerd. "Oh, I just computed the probability that there would simultaneously be two bombs on a plane. This Probability was low enough for my comfort. So, now I carry my own bomb.", he grinned.

Was there any flaw in the statistician's argument? We shall come back to this question after we learn some basic terminologies in conditional probability.

7 Bayes' theorem

Given the occurrence of an event B in a random experiment, the occurrence of an event A is equivalent to the occurrence of the event $A \cap B$. However, since B has already occurred, none of the events which are disjoint from B can occur anymore. So, if the event B has already occurred, then the sample space is reduced to B. The probabilities of the sub-events of B are measured relative to

the probability of B, which in particular ensures that given B, the probability of the occurrence of the event B is 1.

DEFINITION 13 (CONDITIONAL PROBABILITY). Let $(\Omega, \mathcal{E}, P)$ be a probability space and let $A, B \in \mathcal{E}$ with $P(B) > 0$. Then the probability of A given B is defined by

$$P(A|B) := \frac{P(A \cap B)}{P(B)}.$$

LEMMA 1 (TOTAL PROBABILITY). *Let $(\Omega, \mathcal{E}, P)$ be a probability space and let $\{A_i\}_{i=0}^{\infty}$ be exhaustive and pairwise mutually exclusive events such that $P(A_0) = 0$ and $P(A_i)$ is nonzero for all $i \geq 1$. Then for all $B \in \mathcal{E}$, we have*

$$P(B) = \sum_{i=1}^{\infty} P(A_i)P(B|A_i).$$

PROOF. Since Ω is the disjoint union of the events $\{A_i\}_{i=1}^{\infty}$, we may write

$$P(B) = P(B \cap \Omega) = P\Big(B \cap \big(\bigcup_{i=0}^{\infty} A_i\big)\Big) = P\Big(\bigcup_{i=0}^{\infty} B \cap A_i\Big)$$
$$= \sum_{i=1}^{\infty} P(B \cap A_i) = \sum_{i=1}^{\infty} P(A_i)P(B|A_i),$$

where the fourth equality follows from Definition 5.(d), since $\{B \cap A_i\}_{i=0}^{\infty}$ are pairwise mutually exclusive events and since $0 \leq P(B \cap A_0) \leq P(A_0) = 0$. □

THEOREM 5 (BAYES' THEOREM). *Let $(\Omega, \mathcal{E}, P)$ be a probability space and let $\{A_i\}_{i=1}^{\infty}$ be exhaustive and pairwise mutually exclusive events such that $P(A_i)$ is nonzero for all i. Then for all $B \in \mathcal{E}$ with $P(B) > 0$ and for all $j \in \mathbb{N}$,*

$$P(A_j|B) = \frac{P(A_j)P(B|A_j)}{\sum_{i=1}^{\infty} P(A_i)P(B|A_i)}.$$

PROOF. We have

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(A_j)P(B|A_j)}{P(B)} = \frac{P(A_j)P(B|A_j)}{\sum_{i=1}^{\infty} P(A_i)P(B|A_i)},$$

where the last equality follows from Lemma 1. □

EXAMPLE 6. Consider the random experiment, where you keep on rollinng a fair die till you get an outcome of 6 points. For $n \in \mathbb{N}$, let A_n denote the event that you stop rolling the die after the n-th roll, i.e. you obtain the outcome of 6 points for the first time on the n-th roll of the die. Let B denote the event that all the outcomes preceding the last one (which is necessarily 6) are odd. Determine $P(A_m|B)$.

ANSWER. From the definition of the events A_n and B, we obtain

$$P(A_n) = \frac{1}{6}\Big(\frac{5}{6}\Big)^{n-1} \text{ and } P(B|A_n) = \Big(\frac{3}{5}\Big)^{n-1}.$$

Therefore, Bayes' theorem implies that

$$P(A_m|B) = \frac{P(A_m)P(B|A_m)}{\sum_{n=1}^{\infty} P(A_n)P(B|A_n)} = \frac{\frac{1}{6}\cdot\frac{1}{2^{m-1}}}{\frac{1}{6}\sum_{n=1}^{\infty}\frac{1}{2^{n-1}}} = \frac{1}{2^m}.$$

8 ASKING SOMEONE OUT FOR COFFEE

EXAMPLE 7. Suppose, you find someone interesting and you'd like to ask that person out for a coffee. Let's assume that there are the following three pairwise mutually exclusive, exhaustive and equally likely cases:

A. He/she finds you interesting too.
B. He/she feels indifferent towards you.
C. He/she feels repulsed by you.

In Case A, it is natural for him/her to accept your invitation with a rather high probability, say 0.9. In Case B, he/she accepts it with probability 0.5, whereas in Case C, there is still a little chance, say 0.1, that he/she accepts your invitation (in particular, if his/her affinity for coffee is more than the repulsion that he/she feels towards you).
(i) Find the probability that he/she accepts your invitation.
(ii) Given that he/she accepted your invitation, find the probability that he/she finds you interesting too.

ANSWER. (i) Since the cases A, B and C are pairwise mutually exclusive, exhaustive and equally likely, we have

$$P(A) = P(B) = P(C) = \frac{1}{3}.$$

Let Y denote the event that he/she accepts your invitation and N denote the event that he/she declines it. Then

$$\begin{aligned} P(Y) &= P(A)P(Y|A) + P(B)P(Y|B) + P(C)P(Y|C) \\ &= \frac{1}{3}(0.9 + 0.5 + 0.1) \\ &= 0.5\,. \end{aligned}$$

In other words, the events Y and N are equally likely.
(ii) We have

$$P(A|Y) = \frac{P(A)P(Y|A)}{P(Y)} = \frac{1}{3} \times \frac{0.9}{0.5} = 0.6\,.$$

In particular, given that he/she accepts your invitation, though the probability that he/she finds you interesting is greater than $1/2$, yet unfortunately, there's also much room for a mistaken conviction!

EXAMPLE 8. (Monty Hall problem*) Three paper cups are placed upside down on a table. There is a coin under one of these cups, whereas there are nothing under the other two. You don't know under which of the cups the coin lies but your friend does. She asks you to guess which cup hides the coin: You choose one of the cups randomly (but you do not lift it up). Then she lifts up another cup and shows that there's nothing under that. Now, except your initial choice, there still remains a cup which may or may not hide the coin. Given a chance to switch your choice to the remaining cup, would you switch your choice or would you stick to your initial guess? Please justify your answer!

ANSWER. Let A denote the event that the coin is under the cup which you have chosen initially and let L denote the event that your friend lifts another cup to show that there is nothing under it. Note that A^c is the event that the coin is not under the cup which you have chosen. Hence, $A^c|L$ is the event that the coin is under the remaining cup. So, we only require to compare $P(A|L)$ with $P(A^c|L)$. Since in any case, your friend always lifts up an empty cup and shows that there's nothing under that, the event L (and hence, also the event $L|A$) is a sure event. In other words, we have $P(L) = 1 = P(L|A)$. Hence,

$$P(A|L) = \frac{P(A)P(L|A)}{P(L)} = P(A)P(L|A) = P(A) = \frac{1}{3}.$$

Therefore, we conclude that

$$P(A^c|L) = 1 - P(A|L) = \frac{2}{3}.$$

Since $P(A^c|L) > P(A|L)$, you should switch your choice.

If you are not convinced by the above reasoning, then look at the following table that lists all the possible scenarios:

*See https://en.wikipedia.org/wiki/Monty_Hall_problem.

Initial guess	Coin is under	Stick to the first guess	Switch to the remaining cup
Cup 1	Cup 1	Right	Wrong
Cup 1	Cup 2	Wrong	Right
Cup 1	Cup 3	Wrong	Right
Cup 2	Cup 1	Wrong	Right
Cup 2	Cup 2	Right	Wrong
Cup 2	Cup 3	Wrong	Right
Cup 3	Cup 1	Wrong	Right
Cup 3	Cup 2	Wrong	Right
Cup 3	Cup 3	Right	Wrong
Total right guesses		3	6

So, indeed your chance of guessing correctly doubles if you switch your choice.

Example 9 (Random Monty Hall). Three paper cups are placed upside down on a table. There is a coin under one of these cups, whereas there are nothing under the other two. Neither you, nor your friend knows under which of the cups the coin lies. You are asked to guess which cup hides the coin: You choose one of the cups randomly (but you do not lift it up). Then your friend also chooses another cup randomly and lifts it up to find that there's nothing under that. Now, except your initial choice, there still remains a cup which may or may not hide the coin. Given a chance to switch your choice to the remaining cup, would you switch your choice or would you stick to your initial guess? Please justify your answer!

Answer. As before, let A denote the event that the coin is under the cup which you have chosen initially and let L denote the event that your friend lifts another cup randomly to find that there is nothing under it. Note that A^c is the event that the coin is not under the cup which you have chosen. Hence, $A^c|L$ is the event that the coin is under the remaining cup. So, we only require to compare $P(A|L)$ with $P(A^c|L)$. Given A, whichever cup your friend lifts, that will have no coin under it. So, $P(L|A) = 1$. However, $P(L|A^c) = 1/2$. Now, Bayes' theorem implies that

$$P(A|L) = \frac{P(A)P(L|A)}{P(A)P(L|A) + P(A^c)P(L|A^c)} = \frac{\frac{1}{3} \times 1}{\frac{1}{3} \times 1 + \frac{2}{3} \times \frac{1}{2}} = \frac{1}{2}.$$

Since $P(A^c|L) = P(A|L)$, you need not should switch your choice in this case.

10 Independence

Definition 14 (Independent events). Let $(\Omega, \mathcal{E}, P)$ be a probability space and let $A, B \in \mathcal{E}$. If the occurrence of any of A and B does not affect the probability of occurrence of the other, i.e. if $P(A|B) = P(A)$ and $P(B|A) = P(B)$, then we say that A and B are independent*. In other words, A and B are independent if

$$P(A \cap B) = P(A)P(B).$$

*If you wonder how independence looks like on a Venn diagram, then see this: `https://www.youtube.com/watch?v=pV3nZAsJxl0`

The events in a set $S \subseteq \mathcal{E}$ are said to be pairwise independent if the above equation is satisfied by all pairs $A, B \in S$, whereas all the events in S are said to be mutually independent if

$$P\Big(\bigcap_{A\in T} A\Big) = \prod_{A\in T} P(A)$$

for all countable sets $T \subseteq S$.

Example 10. Consider the random experiment where a fair tetrahedral die (whose faces are marked with points from 1 to 4) is rolled. Let the events A, B and C be defined as follows:

$$\begin{aligned} A &: \text{ The die comes to rest on face 1 or face 2.} \\ B &: \text{ The die comes to rest on face 2 or face 3.} \\ C &: \text{ The die comes to rest on face 3 or face 1.} \end{aligned}$$

Note that

$$P(A\cap B) = P(\{2\}) = \frac{1}{4} = P(A)P(B),\ P(B\cap C) = P(\{3\}) = \frac{1}{4} = P(B)P(C),$$

$$P(C \cap A) = P(\{4\}) = \frac{1}{4} = P(C)P(A).$$

However,

$$P(A \cap B \cap C) = P(\emptyset) = 0 \neq P(A)P(B)P(C).$$

In other words, pairwise independence doesn't ensure mutual independence.

Example 11. Let's reconsider the story of the statistician from the beginning of this chapter. He computed the probability p that there will be a bomb on the aeroplane. Then he computed the probability of the event that there are two bombs B_1 and B_2 on the same aeroplane. Apparently, under the assumption of independence, he obtained

$$P(B_1 \cap B_2) = P(B_1)P(B_2) = p^2,$$

which according to him, was low enough compared to p. So, he carries his own bomb B_1! That implies, B_1 is given. Hence, he must instead consider

$$P(B_2|B_1) = \frac{P(B_1 \cap B_2)}{P(B_1)} = \frac{p^2}{p} = p.$$

In other words, to the utter discomfort of the statistician, the probability of there being another bomb on the plane is still p.

11 Base Rate Fallacy

Example 12 (Base Rate Fallacy / False Positive Paradox). Consider an extremely rare disease which affects only 0.1% of the population. Suppose, a test which checks whether the person has the disease, has 99% sensitivity* (true positive rate) and 99% specificity (true negative rate). In other words, the test correctly identifies a positive in 99% of the cases and also correctly identifies a

*See `https://en.wikipedia.org/wiki/Sensitivity_and_specificity`.

negative in 99% of the cases*. Let's assume that the test always gives either a positive or a negative result.
(i) Find the probability that a person has the disease given that he/she tests positive.
(ii) Given that a person has tested positive once, find the probability that he/she has the disease if he/she tests positive again.
(iii) Given that a person has tested positive twice, find the probability that he/she has the disease if he/she tests positive again.

ANSWER. (i) Let D denote the event that the person undergoing the test has the disease. Let P denote the event that the test shows a positive result and let N denote the event that the test shows a negative result. Since the test has 99% sensitivity, we have $P(\mathrm{P}|\mathrm{D}) = 0.99$ and since the test has 99% specificity, we have $P(\mathrm{N}|\mathrm{D}^c) = 0.99$. Hence, $P(\mathrm{P}|\mathrm{D}^c) = 1 - P(\mathrm{N}|\mathrm{D}^c) = 1 - 0.99 = 0.01$.

$$\begin{aligned} P(\mathrm{D}|\mathrm{P}) &= \frac{P(\mathrm{D})P(\mathrm{P}|\mathrm{D})}{P(D)P(\mathrm{P}|\mathrm{D}) + P(D^c)P(\mathrm{P}|\mathrm{D}^c)} \\ &= \frac{0.001 \times 0.99}{0.001 \times 0.99 + 0.999 \times 0.01} \approx 0.09. \end{aligned}$$

Despite the high base rates, viz. $P(\mathrm{P}|\mathrm{D})$ and $P(\mathrm{N}|\mathrm{D}^c)$ apparently indicating a high degree of accuracy of the test, it is surprising that the probability $P(\mathrm{D}|\mathrm{P})$ is so low. This happens because the disease is very rare. More precisely, among 1000 persons, the disease affects only about 1000×0.001=1 person. However, the specificity of the test is 99%, i.e. the test correctly identifies a negative in 99% of the cases. So, among 1000 persons, about 1000(1-0.99)=10 persons get false positive results. Thus, even if the high sensitivity of the test guarantees that the test result is almost certainly positive for any person having the disease, there are also about 10 more persons in every 1000, who gets false positive results. So, even if a person gets a positive result, his/her chance of having the disease is only about $1/11 \approx 0.09$. The higher the prevalence of the disease, the greater would be the probability for a person who tested positive to actually have the disease.
(ii) Since the person has tested positive once, Part (i) implies that his/her probability of having the disease $\mathrm{P}(D)$ has now increased from 0.001 to 0.09. Hence, after the second test, we have

$$\begin{aligned} P(\mathrm{D}|\mathrm{P}) &= \frac{P(\mathrm{D})P(\mathrm{P}|\mathrm{D})}{P(D)P(\mathrm{P}|\mathrm{D}) + P(D^c)P(\mathrm{P}|\mathrm{D}^c)} \\ &= \frac{0.09 \times 0.99}{0.09 \times 0.99 + 0.91 \times 0.01} \approx 0.9. \end{aligned}$$

(iii) Since the person has tested positive twice, Part (ii) implies that his/her probability of having the disease $\mathrm{P}(D)$ has now increased from 0.001 to 0.9.

*Such an ideal test does not exist in practice. See `www.statpearls.com/ArticleLibrary/viewarticle/96435#ref_18158403`

Hence, after the third test, we have

$$P(\mathrm{D}|\mathrm{P}) = \frac{P(\mathrm{D})P(\mathrm{P}|\mathrm{D})}{P(D)P(\mathrm{P}|\mathrm{D}) + P(D^c)P(\mathrm{P}|\mathrm{D}^c)}$$
$$= \frac{0.9 \times 0.99}{0.9 \times 0.99 + 0.1 \times 0.01} \approx 0.999.$$

In other words, the degree of accuracy of the test increases dramatically, if we repeat the test a few times.

12 Return of the coffee

Example 13. In Example 7, check that $P(Y|A) + P(N|A^c) > 1$ to conclude from Exercise 9 that*

$$P(A|\underbrace{YY\ldots Y}_{n}) \to 1 \text{ as } n \to \infty.$$

Answer. It is given in Exercise 7 that $P(Y|A) = 0.9$. Also, we have

$$P(N|A^c) = 1 - P(Y|A^c) = 1 - \frac{P(Y)P(A^c|Y)}{P(A^c)}$$
$$= 1 - \frac{P(Y)(1 - P(A|Y))}{1 - P(A)} = 1 - \frac{0.5 \times (1 - 0.6)}{0.67}$$
$$\approx 0.7\,,$$

where we have put the values of $P(Y), P(A)$ and $P(A|Y)$ from Example 7. In particular, we have $P(Y|A) + P(N|A^c) > 1$. So, under the suitable assumtions of independence, the last exercise implies that

$$P(A|\underbrace{YY\ldots Y}_{n}) \to 1 \text{ as } n \to \infty.$$

In other words, if the person accepts your invitation as many times as you ask him or her out, then you could be almost sure that this person indeed finds you interesting.

*Of course, here we are making two contradicting assumptions about the person of your interest: (1) For him/her, the first impression is the last impression, i.e. his/her feelings of interest / indifference / repulsion towards you does not change over time.
(2) He/she lives in the moment: In particular, given that the person finds you interesting (or not), each time when you ask him/her out for a coffee, his/her spontaneous response is independent of all the past/future responses which you received/will receive from him/her.

On the contrarty, if you do not want to reveal your interest (in coffee or in the person), just occasionally decline invitations for coffee!

Exercises

1. A hostel room is shared by two students, each of whom is equally likely to be either hard working or careless. Given that one of the roommates is hard working, what is the probability that the other one is careless?

2. Let $(\Omega, \mathcal{E}, P)$ be a probability space and let $A_1, A_2, \ldots, A_n \in \mathcal{E}$ with $P(A_1 \cap \cdots \cap A_n) \neq 0$. Show that $P(A_1 \cap \cdots \cap A_n)$ is equal to
$$P(A_1)\, P(A_2 \,|\, A_1)\, P(A_3, |A_2 \cap A_1) \cdots P(A_n \,|\, A_{n-1} \cap \cdots \cap A_1).$$

3. Let $(\Omega, \mathcal{E}, P)$ be a probability space and let $A_1, A_2, \ldots, \in \mathcal{E}$ be pairwise mutually exclusive. Let $A = \cup_{n=1}^{\infty} A_n$ and let $B \in \mathcal{E}$ with $P(B) \neq 0$. Show that
$$P(A \,|\, B) = \sum_{n=1}^{\infty} P(A_n \,|\, B).$$

4. Find the probability that an analog watch shows a specified time if you look at the watch at a random instant*.

5. Find the probability of obtaining n heads in m tosses of a fair coin, given that the m-th toss resulted in a head and no two of the n heads occurred in two consecutive tosses.

6. Let's consider a random experiment in which you keep on tossing a fair coin till you obtain either 4 heads or 4 tails in total. Find the probability that you stop tossing after 6 tosses, given that the first two tosses resulted in heads.
7. There are n boxes numbered $1, 2, \ldots, n$, among which the rth box contains $r - 1$ white cubes and $n - r$ red cubes. Suppose, we choose a box at random and we remove two cubes from it, one after another, without replacement.
   (a) Find the probability of the second cube being red.
   (b) Find the probability of the second cube being red, given that the first cube is red.

8. We are familiar with the famous Monty Hall problem. Now suppose, instead of 3 doors, there are n doors, only one among which has a prize behind it.
   (a) Find the probability of winning upon switching given that Monty opens k doors. Will switching benefit you?

*You may assume that the second-hand of the clock moves discretely.

(b) Find the probability of winning upon switching given that Monty opens maximum number of doors. Will switching benefit you?
(c) Find the probability of winning upon switching given that Monty opens no doors. Will switching benefit you?

9. In a similar situation as in Example 12, show that if

$$\textit{sensitivity} + \textit{specificity} > 100\%,$$

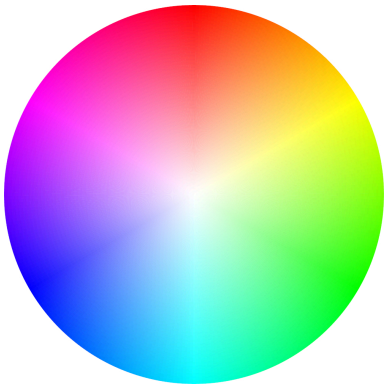
then

$$P(D|\underbrace{PP\ldots P}_{n}) \to 1 \text{ as } n \to \infty.$$

10. Suppose, you satisfy all the assumptions in Example 13. Figure out how often at least you need to decline someone's invitation for coffee if you do not want to reveal your interest in that person.

Random Variables

If you roll a die, you obtain an outcome between 1 to 6 points. When you ask a person for his/her phone number, you obtain* a ten digit number. However in general, the outcomes of a random experiment need not be numbers. For example, consider random experiments like guessing the first letter of a stranger's name or guessing someone's favourite shade of colour.



Often is is useful to quantify the elements of the sample space. A random variable is just a tool for this quanification. For example, in the penultimate example, one may quantify the twenty-six roman alphabets by twenty-six real numbers, whereas in the last example, RGB coordinates may be used to quantify the shades of colours.

In complete generality, a *ranodm variable* is a *measurable function*** on the sample space. But here we restrict ourselves only to the real valued measurable functions:

DEFINITION 15 (RANDOM VARIABLE). A random variable is a real valued function on the sample space such that the preimage of every interval is an event.

DEFINITION 16 (PROBABILITY DISTRIBUTION OF A RANDOM VARIABLE). Let $(\Omega, \mathcal{E}, P)$ be a probability space and let $X : \Omega \to \mathbb{R}$ be a random variable. Then for all $a \in \mathbb{R}$, the preimage of the interval $(-\infty, a]$ under X is in $\mathcal{E}$. The random variable X translates the probability space $(\Omega, \mathcal{E}, P)$ to the probability space $(\mathbb{R}, \mathcal{E}_X, P_X)$,

*assuming that you get a precise numerical response.

**In particular, if the range of the function is $\mathbb{R}^3$ (resp. $\mathbb{R}^2$, $\mathbb{R}$), then the preimage of every set having finite/ infinite volume (resp. area, length) is an event, i.e. a subset S of the sample space having *probability measure* $P(S)$.

where

$$\mathcal{E}_X := \{A \subseteq \mathbb{R} \mid X^{-1}(A) \in \mathcal{E}\} \tag{12.1}$$

and the function $P_X : \mathcal{E}_X \to [0,1]$ is defined by

$$P_X(A) := P(X^{-1}(A)) \text{ for all } A \in \mathcal{E}_X. \tag{12.2}$$

The function P_X is called the *probability distribution function* of X.

Henceforth, we use the notations $\mathrm{P}(X \in A)$ and $P_X(A)$ interchangeably. In particular, we write $\mathrm{P}(X \leq a)$ for $P_X((-\infty, a])$.

LEMMA 2. *Let $(\Omega, \mathcal{E}, P)$ be a probability space, let $X : \Omega \to \mathbb{R}$ be a random variable and let the probability space $(\mathbb{R}, \mathcal{E}_X, P_X)$ be defined as above. Then for all $x \in \mathbb{R}$, we have $\{x\} \in \mathcal{E}_X$.*

PROOF. Since $\mathcal{E}$ is closed under complementation and countable unions, it follows from (12.1) that $\mathcal{E}_X$ is also closed under complementation and countable unions. Since both $(-\infty, x)$ and $(x, \infty) \in \mathcal{E}_X$, it follows that

$$\{x\} = [x, x] = ((-\infty, x) \cup (x, \infty))^c \in \mathcal{E}_X.$$

□

COROLLARY 2. *For a random variable X and for all $x \in \mathbb{R}$, $\mathrm{P}(X = x)$ is well-defined.*

DEFINITION 17 (PROBABILITY MASS FUNCTION). Let X be a random variable. The function from $\mathbb{R}$ to $[0,1]$ which maps

$$x \mapsto \mathrm{P}(X = x)$$

is called the *probability mass function* (PMF) of X.

13 CUMULATIVE DISTRIBUTION FUNCTION

DEFINITION 18 (CUMULATIVE DISTRIBUTION FUNCTION). Let X be a random variable. We define the cumulative density function (CDF) of X by

$$F_X(a) := \mathrm{P}(X \leq a)$$

for all $a \in \mathbb{R}$.

LEMMA 3. *The CDF of a random variable is a nondecreasing function.*

PROOF. For $a \leq b$, we have $X^{-1}((-\infty, a]) \subseteq X^{-1}((-\infty, b])$. Hence, from Exercise 10 in Chapter 1, we obtain

$$F_X(a) = \mathrm{P}(X \leq a) \leq \mathrm{P}(X \leq b) = F_X(b).$$

In other words, $F_X : \mathbb{R} \to [0,1]$ is a nondecreasing function. □

EXAMPLE 14. Let X denote the number of heads obtained in a toss of a fair coin. Write down the CDF of X.

ANSWER. Here we have $P(X < 0) = P(0 < X < 1) = P(X > 1) = 0$ and $P(X = 0) = P(X = 1) = 1/2$. Hence the CDF of X is given by

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1/2 & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1. \end{cases}$$

THEOREM 6. *The CDF of every random variable is right continuous.*

PROOF. Let X be a random variable with CDF F. We require to show that for every sequence $\{x_n\}$ in (x, ∞) with $x_n \to x$, we have

$$\lim_{n\to\infty} F(x_n) = F(x) \tag{13.1}$$

Note that it suffices to prove the above equality for every *decreasing* sequence $\{x_n\}$ with $x_n \to x$: Suppose, $\{x_n\}$ be such a sequence which satisfies (13.1). In other words, for every $\varepsilon > 0$, there is an $N \in \mathbb{N}$ such that for all $n \geq N$, we have $F(x_n) \in B_\varepsilon(F(x))$. Since F is a monotonic function, it follows that for all $a \in (x, x_n]$, we have $F(a) \in B_\varepsilon(F(x))$. Let $\{y_n\}$ be an arbitrary sequence in (x, ∞) with $y_n \to x$. Then there exists an $M \in \mathbb{N}$ such that for all $m \geq M$, we have $y_m \in B_{|x_n - x|}(x) \cap (x, \infty) = (x, x_n)$. Hence, for all $m \geq M$, we have $F(y_m) \in B_\varepsilon(F(x))$. That implies, $\lim_{n\to\infty} F(y_n) = F(x)$. So, if (13.1) is satisfied by a decreasing sequence which converges to x, then it is satisfied by every sequence which converges to x from above.

Therefore, we may assume that $\{x_n\}$ is a decreasing sequence which converges to x. Let $I := (x, \infty)$ and let $I_n := (x_n, \infty)$ for all $n \in \mathbb{N}$. Then we have an ascending chain of intervals

$$I_1 \subseteq I_2 \subseteq \cdots \subseteq I_n \subset I_{n+1} \subseteq \cdots$$

and

$$\bigcup_{n\in\mathbb{N}} I_n = I.$$

Define the intervals I'_n by $I'_1 = I_1$ and $I'_n := I_n \setminus I_{n-1}$ for all integers $n > 1$. Then the intervals $\{I'_n\}_{n\in\mathbb{N}}$ are disjoint with

$$\bigcup_{n\in\mathbb{N}} I'_n = I.$$

So, it follows from Definition 5 that

$$P_X(I) = \sum_{n=1}^{\infty} P_X(I'_n) = \lim_{m\to\infty} \sum_{n=1}^{m} P_X(I'_n) = \lim_{m\to\infty} P_X(I_m).$$

In other words,

$$\mathrm{P}(X > x) = \lim_{n\to\infty} \mathrm{P}(X > x_n).$$

i.e.

$$1 - F(x) = \lim_{n\to\infty} (1 - F(x_n)),$$

which implies the claim. □

COROLLARY 3. *The CDF F_X of a random variable X is continuous if and only if F_X is left continuous.*

14 CONTINUOUS RANDOM VARIABLES

DEFINITION 19 (CONTINUOUS RANDOM VARIABLES). A random variable with a continuous CDF* is called a continuous random variable.

THEOREM 7. *Let X be a random variable. For $x \in \mathbb{R}$, we have*

$$\mathrm{P}(X = x) > 0$$

if and only if the cumulative distribution function F_X is discontinuous at x.

PROOF. From Exercise 4, we know that F_X can have only a jump discontinuities. Since F_X is right continuous (see Theorem 6), a jump discontinuity of F_X at x occurs if and only if $F_X(x-) < F_X(x)$, i.e.

$$P(X < x) < P(X \leq x).$$

Since $P(X \leq x) = P(X < x) + P(X = x)$, it follows that F_X has a discontinuity at x if and only if

$$\mathrm{P}(X = x) > 0.$$

□

COROLLARY 4. *Let X be a continuous random variable. Then*

$$\mathrm{P}(X = x) = 0$$

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

THEOREM 8. *The CDF of a random variable has at most countably many discontinuities.*

PROOF. Follows immediately from Lemma 3 and Exercise 5. □

THEOREM 9. *For every random variable X, the sum*

$$\sum_{x\,:\,\mathrm{P}(X=x)>0} \mathrm{P}(X = x)$$

is well-defined and converges in $[0, 1]$.

PROOF. Let $\mathcal{D}_X$ denote the set of discontinuities of the CDF of the random variable X. Theorem 7 implies that

$$\mathcal{D}_X = \{x : \mathrm{P}(X = x) > 0\}$$

and Theorem 8 implies that $\mathcal{D}_X$ is countable. Hence, the sum of $P(X = x)$ over all $x \in \mathcal{D}_X$ is well-defined and we have

$$0 \leq \sum_{x\,:\,\mathrm{P}(X=x)>0} \mathrm{P}(X = x) = \mathrm{P}(X \in \mathcal{D}_X) \leq \mathrm{P}(X \in \mathbb{R}) = 1,$$

where the sum is equal to 0 if and only if $\mathcal{D}_X = \emptyset$, i.e. X is a continuous random variable. Also, if $\mathcal{D}_X$ is finite, the claim is trivial. Otherwise, the partial sums

*By Exercise 3, such a CDF is also uniformly continuous.

of the above series of positive real numbers form a bounded increrasing sequence in $[0,1]$. So, the above series converges in $[0,1]$ by the Monotone Convergence Theorem. □

Definition 20 (Classification of random variables). Let X be a random variable. If

$$\sum_{x\,:\,\mathrm{P}(X=x)>0} \mathrm{P}(X=x) = \begin{cases} 0, & \text{then } X \text{ is a } \textit{continuous} \text{ random variable} \\ \mu & \text{for some } \mu \in (0,1),\ X \text{ is a } \textit{mixed} \text{ random variable} \\ 1, & \text{then } X \text{ is called a } \textit{discrete} \text{ random variable.} \end{cases}$$

Definition 21 (Probability density function). For a continuous random variable X, if there exists a function $f_X : \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that

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

for all $x \in \mathbb{R}$, then f_X is called a *probability density function* (PDF) of X. Here F_X denotes the CDF of X. In particular, we have

$$\mathrm{P}(a < X \leq b) = F_X(b) - F_X(a).$$

and

$$\int_{-\infty}^{\infty} f_X(t)dt = 1.$$

In particular, if F_X is differentiable, then we may take f_X to be the derivative of F_X. Note that a PDF of a continuous random variable is not necessarily continuous. However, if f_X is continuous, then the Fundamental Theorem of Calculus implies that

$$f_X(t) = F_X'(t).$$

Example 15. Define the CDF of a random variable X by

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \dfrac{x}{x+1} & \text{otherwise.} \end{cases}$$

Since $F : \mathbb{R} \to [0,1)$ is continuous, X is a continuous random variable. Moreover, since F is differentaible, we may take F' as a PDF of X. Thus, we may define $f_X : \mathbb{R} \to [0,1]$ by

$$f_X(t) = \begin{cases} 0 & \text{if } t < 0 \\ \dfrac{1}{(t+1)^2} & \text{otherwise.} \end{cases}$$

15 Mixed random variables

Theorem 10. *If the CDF F_X of a random variable X*

(a) *has a discontinuity in $\mathbb{R}$ and*

(b) *is continuous and increasing*[*] *in an interval of nonzero length*[**]*,*

then X is a mixed random variable.

PROOF. Let I be an interval of nonzero length such that F_X is continuous on I. Let $a, b \in I$ such that $a < b$. Then F_X is continuous and increasing on $(a, b]$. Hence, Theorem 7 implies that

$$\mathrm{P}(X = x) = 0 \text{ for all } x \in (a, b].$$

However,

$$\mathrm{P}(X \in (a, b]) = F_X(b) - F_X(a) > 0. \tag{15.1}$$

Let $\mathcal{D}_X$ be the set of discontinuities of F_X. Since F_X is continuous on $(a, b]$, it follows that $\mathcal{D}_X \cap (a, b] = \emptyset$. Hence, from Definition 5, we obtain

$$\mathrm{P}(X \in \mathcal{D}_X) + \mathrm{P}(X \in (a, b]) = \mathrm{P}(X \in \mathcal{D}_X \cup (a, b]) \leq \mathrm{P}(R) = 1.$$

That implies,

$$\mathrm{P}(X \in \mathcal{D}_X) \leq 1 - \mathrm{P}(X \in (a, b]) < 1, \tag{15.2}$$

because $\mathrm{P}(X \in (a, b]) > 0$ (see (15.1)). Since F_X has a discontinuity in $\mathbb{R}$, it follows that $\mathcal{D}_X \neq \emptyset$ and Theorem 7 implies that $\mathcal{D}_X = \{x : \mathrm{P}(X = x) > 0\}$. Hence, from (15.2), it follows that

$$0 < \sum_{x\,:\,\mathrm{P}(X=x)>0} \mathrm{P}(X = x) < 1.$$

Therefore, X is a mixed random variable. □

EXAMPLE 16. Let the range of a random variable X be $\{1, 4\} \cup [2, 3]$, where $[2, 3]$ denotes the closed interval with 2 and 3 at its extremities. Define the CDF of X by

$$F_X(x) = \begin{cases} 0 & \text{if } x < 1 \\ 1/4 & \text{if } 1 \leq x < 2 \\ x/4 & \text{if } 2 \leq x \leq 3 \\ 3/4 & \text{if } 3 \leq x < 4 \\ 1 & \text{if } x \geq 4. \end{cases}$$

Then F_X is continuous and increasing in the interval [2,3] and it has discontinuities at $1, 2$ and 4. Hence, Theorem 10 implies that X is a mixed random variable. We could have also checked this directly: Since there are only 3 values of $x \in \mathbb{R}$ such that $\mathrm{P}(X = x)$ is nonzero, viz.

$$\mathrm{P}(X = 1) = \mathrm{P}(X = 2) = \mathrm{P}(X = 4) = \frac{1}{4},$$

[*]by *increasing*, we mean *strictly increasing*.

[**]Note that, though the CDF of a random variable has countably many discontinuities, but in general, there may not be any interval of nonzero length on which the CDF is continuous. For example, consider a random variable which has nonzero probabilities only at the rational numbers.

it follows that

$$\sum_{x\,:\,\mathrm{P}(X=x)>0} \mathrm{P}(X=x) = \frac{3}{4} \in (0,1).$$

Henceforth, we would restrict ourselves only to discussions about continuous and discrete random variables.

16 Discrete random variables

Theorem 11. *If a random variable X assumes only countably many values, then X is a discrete random variable.*

Proof. Let $\mathcal{C}$ denote the set of values which X assumes, let

$$\mathcal{C}_+ = \{x \in \mathcal{C} \,:\, \mathrm{P}(X=x) > 0\} \text{ and } \mathcal{C}_0 = \{x \in \mathcal{C} \,:\, \mathrm{P}(X=x) = 0\}.$$

Then $\mathcal{C}_+ \cup \mathcal{C}_0 = \mathcal{C}$ and $\mathcal{C}_+ \cap \mathcal{C}_0 = \emptyset$. Hence, from Definition 5, it follows that

$$\begin{aligned}\sum_{x\in\mathcal{C}_+} \mathrm{P}(X=x) &= \sum_{x\in\mathcal{C}_+} \mathrm{P}(X=x) + \sum_{x\in\mathcal{C}_0} \mathrm{P}(X=x)\\ &= \sum_{x\in\mathcal{C}_+\cup\mathcal{C}_0} \mathrm{P}(X=x) = \sum_{x\in\mathcal{C}} \mathrm{P}(X=x)\\ &= \mathrm{P}(X\in\mathcal{C}).\end{aligned}$$

Since X does not assume any value in $\mathbb{R} \setminus \mathcal{C}$, we have

$$\mathrm{P}(X \in \mathbb{R} \setminus \mathcal{C}) = 0.$$

In particular, it follows from the above equation and Exercise 10 that

$$\mathrm{P}(X=x) = 0 \text{ for all } x \in \mathbb{R} \setminus \mathcal{C}.$$

Hence, the set of all the real numbers x such that $\mathrm{P}(X=x) > 0$ is $\mathcal{C}_+$, i.e.

$$\{x \,:\, \mathrm{P}(X=x) > 0\} = \mathcal{C}_+.$$

Thus, we obtain

$$\begin{aligned}\sum_{x\,:\,\mathrm{P}(X=x)>0} \mathrm{P}(X=x) &= \sum_{x\in\mathcal{C}_+} \mathrm{P}(X=x)\\ &= \mathrm{P}(X\in\mathcal{C})\\ &= \mathrm{P}(X\in\mathcal{C}) + \mathrm{P}(X\in\mathbb{R}\setminus\mathcal{C})\\ &= \mathrm{P}(X\in\mathbb{R})\\ &= 1.\end{aligned}$$

So, X is indeed a discrete random variable according to Definition 20. □

Since every set of isolated points in $\mathbb{R}$ is countable, we conclude:

Corollary 5. *If a random variable X takes values only in a discrete set*, then X is a discrete random variable.*

*A set of isolated points.

DEFINITION 22 (BERNOULLI/BINARY RANDOM VARIABLE). If a random variable X assumes only a pair of distinct values, we call X a Bernoulli/Binary random variable.

In particular, if a random experiment has only two possible outcomes, e.g.

(a) tossing a coin: $\Omega = \{H, T\}$
(b) writing an exam: $\Omega = \{\text{pass}, \text{fail}\}$
(c) undertaking a mission: $\Omega = \{\text{success}, \text{failure}\}$
(d) asking someone out for coffee: $\Omega = \{\text{positive reply}, \text{negative reply}\}$,

then any random variable $X : \Omega \to \mathbb{R}$ which is not a constant function, is a Bernoulli random variable. However, having exactly two element is not at all a necessary condition on the sample space Ω for $X : \Omega \to \mathbb{R}$ to be a Bernoulli random variable. For example, the range of human body temperature is uncountable, whereas a person is said to have a fever if his/her temperature is higher than $98.6°$ Fahrenheit. So, for a randomly chosen person, we may define a Bernoulli random variable X on the range of human body temperature, say $[55, 115]$ in Fahrenheit such that $X = 1$ if the temperature of the person is higher than $98.6°$ F and $X = 0$ otherwise.

NOTATION 1. Let X be a Bernoulli random variable taking values in $\{0, 1\}$ such that $\mathrm{P}(X = 1) = p$. Then we write

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

The probability p is called the *parameter* of Bernoulli distribution.

DEFINITION 23 (DISCRETE UNIFORM RANDOM VARIABLE). If a random variable X assumes only n equally probable distinct values, then we call X a Discrete Uniform random variable and we write

$$X \sim \text{Uniform}(n).$$

17 INDEPENDENCE OF RANDOM VARIABLES

DEFINITION 24. Let X and Y be two random variables defined on the same sample space and let $A \in \mathcal{E}_X$ and $B \in \mathcal{E}_Y$. If $\mathrm{P}(Y \in B) > 0$, we define the conditional probability

$$\mathrm{P}(X \in A \mid Y \in B) := \frac{\mathrm{P}(X \in A, Y \in B)}{\mathrm{P}(Y \in B)},$$

where

$$\mathrm{P}(X \in A, Y \in B) := \mathrm{P}(\{X \in A\} \cap \{Y \in B\}).$$

Two random variables are *independent* if prior knowledge about the value assumed by any one of them does not alter the probability of occurrence of the other in any given subset of $\mathbb{R}$, i.e. the random variables X and Y are independent if and only if

$$\mathrm{P}(X \in A \mid Y \in B) = \mathrm{P}(X \in A) \text{ and } \mathrm{P}(Y \in B \mid X \in A) = \mathrm{P}(Y \in B),$$

provided $\mathrm{P}(Y \in B) > 0$ and $\mathrm{P}(X \in A) > 0$, respectively. In other words, two random variables X and Y are independent if and only if

$$\mathrm{P}(X \in A, Y \in B) = \mathrm{P}(X \in A)\,\mathrm{P}(Y \in B). \tag{17.1}$$

for all $A \in \mathcal{E}_X$ and $B \in \mathcal{E}_Y$. Let $\{X_i\}_{i\in S}$ be a set of random variables. The random variables X_i are said to be *pairwise independent* if the above equation is satisfied by all pairs $\{X_i, X_j\}$ for distinct $i, j \in S$, whereas all the random variables X_i are said to be *mutually independent* if

$$\mathrm{P}\Big(\bigcap_{i\in T}\{X_i \in A_i\}\Big) = \prod_{i\in T}\mathrm{P}(X_i \in A_i). \tag{17.2}$$

for all $A_i \in \mathcal{E}_{X_i}$ and for all countable sets $T \subseteq S$.

Definition 25 (iid random variables). Let $\{X_i\}_{i\in S}$ be a set of random variables. The random variables X_i are said to be *independent and identically distributed* (iid) if they satisfy both of the following conditions:

(a) The random variables X_i are mutually independent.
(b) $F_{X_i}(x) = F_{X_j}(x)$ for all $i, j \in S$ and for all $x \in \mathbb{R}$.

18 Random vectors and conditional distribution

A random vector is only a generalized random variable, whose range is $\mathbb{R}^n$ for some $n \in \mathbb{N}$.

Definition 26 (Random vector and joint CDF). Let $X_1, X_2, \ldots, X_n$ be random variables defined on the same sample space. Then the ordered n-tuple

$$\widetilde{X} := (X_1, X_2, \ldots, X_n)$$

is called a random vector*. The random vector $\widetilde{X}$ is said to be *discrete* or *continuous* if all its component random variables X_i are discrete or continuous, respectively. The joint CDF of $X_1, X_2, \ldots, X_n$ is defined by

$$\begin{aligned} F_{\widetilde{X}}(x_1, x_2, \ldots, x_n) &:= \mathrm{P}(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n) \\ &= \mathrm{P}(\{X_1 \le x_1\} \cap \{X_2 \le x_2\} \cap \cdots \cap \{X_n \le x_n\}). \end{aligned}$$

Definition 27 (Joint PMF). Let $X_1, X_2, \ldots, X_n$ be discrete random variables defined on the same sample space. We define the joint PMF of the random variables $X_1, \ldots, X_n$ by

$$\mathrm{P}(X_1 = x_1, \ldots, X_n = x_n) := \mathrm{P}(\{X_1 = x_1\} \cap \{X_2 = x_2\} \cap \cdots \cap \{X_n = x_n\}).$$

Example 17. Let's consider the random experiment in which a fair coin is tossed thrice. Let the random variables X and Y denote the number of heads and tails obtained, respectively. The joint PMF $\mathrm{P}(X = i, Y = j)$ is presented in the following table:

*Similarly, a matrix whose elements are random variables defined on the same sample space is called a random matrix.

X=i \ Y=j	0	1	2	3	$\mathrm{P}(X=i)$
0	0	0	0	1/8	1/8
1	0	0	3/8	0	3/8
2	0	3/8	0	0	3/8
3	1/8	0	0	0	1/8
$\mathrm{P}(Y=j)$	1/8	3/8	3/8	1/8	

The PMFs of X and Y are given by the row sums and the column sums in the above table, respectively. Since they are written on the margins of the table, they are called *marginal PMF*s. In general, the PMF of a component random variable X_{i_0} of a discrete random vector $\widetilde{X}$ is obtained similarly by summing the joint PMF over the entire range of all the components of $\widetilde{X}$ except X_{i_0}.

Definition 28 (Marginal PMF). Let $X_1, X_2, \ldots, X_n$ be discrete random variables defined on the same sample space. Then for all $i \in \{1, 2, \ldots, n\}$ the map from $\mathbb{R}$ to $[0, 1]$ that sends $x \mapsto \mathrm{P}(X_i = x)$ is called the i-th marginal PMF of the joint distribution of $X_1, X_2, \ldots, X_n$.

Lemma 4. *Let $X_1, X_2, \ldots, X_n$ be discrete random variables defined on the same sample space. Then*

$$\mathrm{P}(X_1 = x_1) = \sum_{x_2,\ldots,x_n \,:\, \mathrm{P}(X_1=x_1,\ldots,X_n=x_n)>0} \mathrm{P}(X_1 = x_1, \ldots, X_n = x_n).$$

Proof. Follows immediately from the Lemma of total probability (see Lemma 1 in Chapter 2). □

Note that for all $i \in \{1, 2, \ldots, n\}$, after interchanging X_i and X_1, it follows that a similar result as above holds for all $i \in \{1, 2, \ldots, n\}$.

Definition 29 (Conditional PMF and conditional CDF). Let X and Y be discrete random variables defined on the same sample space. We define* the conditional PMF of X given Y as the quotient of the joint PMF of X and Y with the marginal PMF of Y. More precisely, the conditional PMF of X given $Y = y$ is given by

$$\mathrm{P}(X = x \mid Y = y) = \frac{\mathrm{P}(X = x, Y = y)}{\mathrm{P}(Y = y)}$$

for all y with $\mathrm{P}(Y = y) > 0$. Hence, for all such y, we define the conditional CDF of X given $Y = y$ by

$$\begin{aligned}\mathrm{P}(X \le x \mid Y = y) &= \sum_{\substack{r \le x \\ \mathrm{P}(X=r \mid Y=y)>0}} \mathrm{P}(X = r \mid Y = y) \\ &= \frac{1}{\mathrm{P}(Y = y)} \sum_{\substack{r \le x \\ \mathrm{P}(X=r \mid Y=y)>0}} \mathrm{P}(X = r, Y = y).\end{aligned}$$

*See Definition 24.

In particular, if X and Y are independent, then their conditional PMF and CDF are the same as the unconditional ones.

DEFINITION 30 (JOINT PDF). Let $X_1, X_2, \ldots, X_n$ be continuous random variables defined on the same sample space and let $F_{\widetilde{X}}$ denote their joint CDF. If there exists a function $f_{\widetilde{X}} : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ such that

$$F_{\widetilde{X}}(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_n} f_{\widetilde{X}}(t_1, t_2, \ldots, t_n) dt_1 dt_2 \cdots dt_n,$$

then $f_{\widetilde{X}}$ is called a joint PDF of $X_1, X_2, \ldots, X_n$. We have

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\widetilde{X}}(t_1, t_2, \ldots, t_n) dt_1 dt_2 \cdots dt_n = 1.$$

In particular, if $F_{\widetilde{X}}$ is differentable to order n, then

$$f_{\widetilde{X}}(t_1, t_2, \ldots, t_n) = \frac{\partial^n F_{\widetilde{X}}(t_1, t_2, \ldots, t_n)}{\partial t_1 \partial t_2 \cdots \partial t_n}.$$

LEMMA 5. *Let X and Y be two independent continuous random variables with continuous PDF f_X and f_Y. Then the joint PDF $f_{X,Y}$ is given by*

$$f_{X,Y}(x, y) = f_X(x) f_Y(y)$$

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

PROOF. Since f_X and f_Y are continuous, for $x, y \in \mathbb{R}$ and for a sufficiently small $\varepsilon > 0$, we have

$$\begin{aligned}
&\mathrm{P}(x - \varepsilon/2 \leq X \leq x + \varepsilon/2, y - \varepsilon/2 \leq Y \leq y + \varepsilon/2) \\
&= \mathrm{P}(x - \varepsilon/2 \leq X \leq x + \varepsilon/2)\, \mathrm{P}(y - \varepsilon/2 \leq Y \leq y + \varepsilon/2) \\
&= \int_{x-\varepsilon/2}^{x+\varepsilon/2} f_X(t_1) dt_1 \int_{y-\varepsilon/2}^{y+\varepsilon/2} f_X(t_2) dt_2 \\
&\approx f_X(x) f_Y(y) \varepsilon^2,
\end{aligned}$$

where the first equality follows from (17.1) since X and Y are independent. Now, dividing both sides by ε^2 and letting $\varepsilon \to 0$, we obtain

$$f_{X,Y}(x, y) = f_X(x) f_Y(y)$$

for all $x, y \in \mathbb{R}$. □

Similarly, from (17.2), we obtain the following:

COROLLARY 6. *If $X_1, X_2, \ldots, X_n$ are mutually independent random variables with continuous PDF $f_{X_1}, \ldots, f_{X_n}$, then for each subset $\{X_{i_1}, X_{i_2}, \ldots, X_{i_m}\} \subseteq \{X_1, X_2, \ldots, X_n\}$, the joint PDF $f_{X_{i_1}, X_{i_2}, \ldots, X_{i_m}}$ is given by*

$$f_{X_{i_1}, X_{i_2}, \ldots, X_{i_m}}(x_{i_1}, x_{i_2}, \ldots, x_{i_m}) = f_{X_{i_1}}(x_{i_1}) f_{X_{i_2}}(x_{i_2}) \cdots f_{X_{i_m}}(x_{i_m})$$

for all $x_{i_1}, x_{i_2}, \ldots, x_{i_m} \in \mathbb{R}$.

DEFINITION 31 (MARGINAL PDF). Let $X_1, X_2, \ldots, X_n$ be continuous random variables defined on the same sample space with PDFs $f_{X_1}, f_{X_2}, \ldots, f_{X_n}$. Then for all $i \in \{1, 2, \ldots, n\}$ the map from $\mathbb{R}$ to $[0, 1]$ that sends $x \mapsto f_{X_i}(x)$ is called the i-th marginal PDF of the joint distribution of $X_1, X_2, \ldots, X_n$.

Similarly as in the case of a discrete random vector, the PDF of a component random variable X_{i_0} of a continuous random vector $\widetilde{X}$ is obtained by integrating the joint PDF over the entire range of all the components of $\widetilde{X}$ except X_{i_0}.

DEFINITION 32 (CONDITIONAL PDF AND CONDITIONAL CDF). Let X and Y be continuous random variables defined on the same sample space. Similarly as in the case of the discrete random variables, we define the conditional PDF of X given Y as the quotient of the joint PDF of X and Y with the marginal PDF of Y. More precisely, if $f_{X,Y}$ denotes the joint PDF of X and Y, then the conditional PDF of X given $Y = y$ is given by

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

for all y with $f_Y(y) > 0$.

Hence, for all such y, we define the conditional CDF of X given $Y = y$ by

$$\begin{aligned} F_{X|Y}(x \mid y) &= \mathrm{P}(X \leq x \mid Y = y) \\ &= \int_{-\infty}^{x} f_{X|Y}(t \mid y)dt \\ &= \frac{1}{f_Y(y)} \int_{-\infty}^{x} f_{X,Y}(t, y)dt. \end{aligned}$$

In particular, if X and Y are independent, then their conditional PDF and CDF are the same as the unconditional ones (See Lemma 5).

19 EXPECTATION, MOMENTS AND VARIANCE

We define expectation and variance for only discrete random variables or continuous random variables having a PDF.

DEFINITION 33 (EXPECTATION OF A FUNCTION OF A DISCRETE RANDOM VARIABLE). Let X be a discrete random variable and let $g : \mathbb{R} \to \mathbb{R}$ be a function. If the series

$$\sum_{x \,:\, \mathrm{P}(X=x)>0} g(x)\, \mathrm{P}(X = x)$$

converges absolutely, then we define the *expectation* of $g(X)$ by

$$E(g) := \sum_{x \,:\, \mathrm{P}(X=x)>0} g(x)\mathrm{P}(X = x).$$

DEFINITION 34 (EXPECTATION OF A FUNCTION OF A CONTINUOUS RANDOM VARIABLES). Let X be a continuous random variable with PDF f_X and let $g : \mathbb{R} \to \mathbb{R}$ be a function. If the integral

$$\int_{-\infty}^{\infty} g(x) f_X(x) dx$$

converges absolutely, then we define the *expectation* of g by

$$E(g) := \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

DEFINITION 35 (n-TH MOMENT AND EXPECTATION). Let X be a random variable. We call $E(X^n)$ the *n-th moment* of X. In particular, the first moment $E(X)$ is called the *expectation* of X and it is often denoted by μ_X.

EXAMPLE 18. Let $X \sim \text{Bernoulli}(p)$. Then

$$E(X) = 1 \cdot \mathrm{P}(X = 1) + 0 \cdot \mathrm{P}(X = 0) = p.$$

DEFINITION 36 (VARIANCE). Let X be a random variable with expectation $\mu = E(X)$. We define the variance of X by

$$\begin{aligned} \mathrm{Var}(X) &:= E((X - \mu)^2) \\ &= E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - \mu^2 \\ &= E(X^2) - E(X)^2, \end{aligned}$$

where the third equality follows from Exercise 19, assuming the existence of the second moment of X.

Note that variance of a random variable is a nonnegative quantity by definition.

DEFINITION 37 (STANDARD DEVIATION). Let X be a random variable with variance $\mathrm{Var}(X)$. Then

$$\sigma_X := \sqrt{\mathrm{Var}(X)}$$

is called the *standard deviation* of X.

DEFINITION 38 (EXPECTATION OF A FUNCTION OF n DISCRETE RANDOM VARIABLES). Let $X_1, X_2, \ldots, X_n$ be a set of discrete random variables defined on the same sample space and let $g : \mathbb{R}^n \to \mathbb{R}$ be a function. If the series

$$\sum_{x_1,\ldots,x_n \,:\, \mathrm{P}(X_1=x_1,\ldots,X_n=x_n)>0} g(x_1, \ldots, x_n)\, \mathrm{P}(X_1 = x_1, \ldots, X_n = x_n)$$

converges absolutely, then we define the *expectation* of $g(X_1, \ldots, X_n)$ by

$$E(g) := \sum_{x_1,\ldots,x_n \,:\, \mathrm{P}(X_1=x_1,\ldots,X_n=x_n)>0} g(x_1, \ldots, x_n)\, \mathrm{P}(X_1 = x_1, \ldots, X_n = x_n).$$

DEFINITION 39 (EXPECTATION OF A FUNCTION OF n CONTINUOUS RANDOM VARIABLES). Let $X_1, X_2, \ldots, X_n$ be a set of n continuous random variables defined on the same sample space. Let $f_{\widetilde{X}}$ denote the joint PDF of $X_1, X_2, \ldots, X_n$ and let $g : \mathbb{R}^n \to \mathbb{R}$ be a function. If the integral

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(x_1, x_2 \ldots, x_n)\, f_{\widetilde{X}}(x_1, x_2 \ldots, x_n)dx_1dx_2\cdots dx_n$$

converges absolutely, then we define the *expectation* of $g(X_1, \ldots, X_n)$ by

$$E(g) := \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(x_1, x_2 \ldots, x_n)\, f_{\widetilde{X}}(x_1, x_2 \ldots, x_n)dx_1dx_2\cdots dx_n.$$

THEOREM 12. *Let X and Y be two random variables with finite expectations, defined on the same sample space. Then for all $a, b \in \mathbb{R}$, we have*

$$E(aX + bY) = aE(X) + bE(Y).$$

PROOF.

Case 1. (Discrete case)
Since both $E(X)$ and $E(Y)$ are finite, we have

$$\begin{aligned} aE(X) + bE(Y) &= a \sum_{x\,:\,\mathrm{P}(X=x)>0} x\,\mathrm{P}(X = x) + b \sum_{y\,:\,\mathrm{P}(Y=y)>0} y\,\mathrm{P}(Y = y) \\ &= \sum_{x\,:\,\mathrm{P}(X=x)>0} ax\,\mathrm{P}(X = x) + \sum_{y\,:\,\mathrm{P}(Y=y)>0} by\,\mathrm{P}(Y = y), \end{aligned}$$

where both of the above sums converge absolutely. Now, Lemma 4 implies that

$$\mathrm{P}(X = x) = \sum_{y\,:\,\mathrm{P}(X=x,Y=y)>0} \mathrm{P}(X = x, Y = y)$$

and

$$\mathrm{P}(Y = y) = \sum_{x\,:\,\mathrm{P}(X=x,Y=y)>0} \mathrm{P}(X = x, Y = y).$$

Putting these back in the expression for $aE(X) + bE(Y)$ above, we obtain

$$\begin{aligned} aE(X) + bE(Y) = &\sum_{x\,:\,\mathrm{P}(X=x)>0} ax \sum_{y\,:\,\mathrm{P}(X=x,Y=y)>0} \mathrm{P}(X = x, Y = y) \\ &+ \sum_{y\,:\,\mathrm{P}(Y=y)>0} by \sum_{x\,:\,\mathrm{P}(X=x,Y=y)>0} \mathrm{P}(X = x, Y = y), \end{aligned}$$

where all the above sums converge absolutely. Hence, by Dirichlet's theorem on absolutely convergent series, it follows that

$$\begin{aligned} aE(X) + bE(Y) &= \sum_{x,y\,:\,\mathrm{P}(X=x,Y=y)>0} (ax + by)\,\mathrm{P}(X = x, Y = y) \\ &= E(aX + bY). \end{aligned}$$

Case 2. (Continuous case)
Since both $E(X)$ and $E(Y)$ are finite, we have

$$\begin{aligned} aE(X) + bE(Y) &= a\int_{-\infty}^{\infty} x\, f_X(x)dx + b\int_{-\infty}^{\infty} y\, f_Y(y)dy \\ &= \int_{-\infty}^{\infty} ax \int_{-\infty}^{\infty} f_{X,Y}(x,y)dydx + \int_{-\infty}^{\infty} by \int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy. \end{aligned}$$

Since both the above integrals converge absolutely, by Fubini's theorem* we obtain

$$aE(X) + bE(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(ax+by)f_{X,Y}(x,y)dxdy \quad = E(aX+bY).$$

□

COROLLARY 7. *Let $X_1, X_2, \ldots, X_n$ be a set of random variables with finite expectations, defined on the same sample space. Then for all $a_1, a_2, \ldots, a_n \in \mathbb{R}$, we have*

$$E(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1E(X_1) + \cdots + a_nE(X_n).$$

THEOREM 13. *Let X and Y be two random variables with finite expectations, defined on the same sample space. If X and Y are independent, then*

$$E(XY) = E(X)E(Y).$$

PROOF.
Case 1. (Discrete case)
Since X and Y are independent, for all $x, y \in \mathbb{R}$, we have

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

Hence, it follows that

$$\begin{aligned} E(XY) &= \sum_{x,y\,:\,\mathrm{P}(X=x,Y=y)>0} xy\,\mathrm{P}(X=x,Y=y) \\ &= \sum_{x\,:\,\mathrm{P}(X=x)>0 \text{ and } y\,:\,\mathrm{P}(Y=y)>0} xy\,\mathrm{P}(X=x)\,\mathrm{P}(Y=y). \end{aligned}$$

*Or by considering the absolute convergence of the Riemann sum whose limiting value is equal to the above integral and then rearranging the order of summation (by Dirichlet's theorem on absolutely convergent series)

Since the above sum converges absolutely, by Dirichlet's theorem on absolutely convergent series, we may rewrite the above sum as

$$\begin{aligned} E(XY) &= \sum_{x\,:\,\mathrm{P}(X=x)>0} x\,\mathrm{P}(X=x) \sum_{y\,:\,\mathrm{P}(Y=y)>0} y\,\mathrm{P}(Y=y). \\ &= \sum_{x\,:\,\mathrm{P}(X=x)>0} x\,\mathrm{P}(X=x)\,E(Y) \\ &= E(X)E(Y). \end{aligned}$$

Case 2. (Continuous case)
Since X and Y are independent, for all $x, y \in \mathbb{R}$, we have $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. Hence, it follows that

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{X,Y}(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_X(x)f_Y(y)dxdy \end{aligned}$$

Since the above integrals converge absolutely, by Fubini's theorem* we may rewrite the above integral as

$$E(XY) = \int_{-\infty}^{\infty} x f_X(x)dx \int_{-\infty}^{\infty} y f_Y(y)dy = E(X)E(Y).$$

□

COROLLARY 8. *Let $X_1, X_2, \ldots, X_n$ be a set of random variables with finite expectations, defined on the same sample space. If $X_1, X_2, \ldots, X_n$ are mutually independent, then*

$$E(X_1X_2\ldots X_n) = \prod_{i=1}^{n} E(X_i).$$

20 COVARIANCE AND CORRELATION

Covariance is a crude measure of linear dependence between two random variables. Recall from Lemma 13 that if X and Y are independent random variables, then we have $E(XY) = E(X)E(Y)$. This is however, not true for arbitrary random variables X and Y. In general, we require a correction factor in the above equation, viz. the covariance of X and Y.

DEFINITION 40 (COVARIANCE). We define the covariance of two random variables X and Y by

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

*Or by considering the absolute convergence of the Riemann sum whose limiting value is equal to the above integral and then rearranging the order of summation (by Dirichlet's theorem on absolutely convergent series)

LEMMA 6. *For a random variable X and for $Y = aX + b$ for some $a, b \in \mathbb{R}$, we have*

$$\mathrm{Cov}(X, Y) = a\mathrm{Var}(X).$$

PROOF. Since $Y = aX + b$, we have

$$\begin{aligned}\mathrm{Cov}(X, Y) &= E(XY) - E(X)E(Y) \\ &= E(aX^2 + bX) - E(X)E(aX + b) \\ &= aE(X^2) + bE(X) - aE(X)^2 - bE(X) \\ &= a\mathrm{Var}(X),\end{aligned}$$

where the third equality follows from Theorem 12. □

EXAMPLE 19. Consider a random variable X such that both its expectation and third moment is zero* Then for $Y = X^2$, we have $\mathrm{Cov}(X, Y) = 0$.

PROOF. Since $Y = X^2$, we have

$$\begin{aligned}\mathrm{Cov}(X, Y) &= E(XY) - E(X)E(Y) \\ &= E(X^3) - E(X)E(X^2) \\ &= 0,\end{aligned}$$

since the existence of the third moment implies that $E(X^2)$ exists** and since both $E(X)$ and $E(X^3)$ are zero. □

DEFINITION 41 (CORRELATION COEFFICIENT). The correlation coefficient of two random variables X and Y is a normalized measure of their linear dependence, defined by

$$\rho(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

if both σ_X and σ_Y, i.e. the standard deviations of X and Y are nonzero.

THEOREM 14. *For any two random variables X and Y, we have $|\rho(X, Y)| \leq 1$.*

PROOF. Since variance is the expectation of a nonnegative quantity (see Definition 36), variance of any function is nonnegative. In particular, we have

$$\begin{aligned}0 &\leq \mathrm{Var}\left(\frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y}\right) \\ &= \frac{1}{\sigma_X^2}\mathrm{Var}(X) + \frac{1}{\sigma_Y^2}\mathrm{Var}(Y) \pm \frac{2}{\sigma_X \sigma_Y}\mathrm{Cov}(X, Y) \\ &= 2(1 \pm \rho(X, Y)).\end{aligned}$$

where the second equality follows from Exercise 20. Hence, we conclude that

$$-1 \leq \rho(X, Y) \leq 1.$$

□

*For example, when X is the standard normal random variable.

**See Exercise 17.

THEOREM 15. *For a random variable X and for $Y = aX + b$ for some $a, b \in \mathbb{R}$, we have*

$$\rho(X, Y) = \operatorname{sgn}(a),$$

where sgn *denotes the sign of a.*

PROOF. Since $Y = aX + b$, we have

$$\begin{aligned}
\operatorname{Var}(Y) &= E(Y^2) - E(Y)^2 = E(aX + b)^2 - (aE(X) + b)^2 \\
&= E(a^2X^2 + 2abX + b^2) - a^2E(X)^2 - 2abE(X) - b^2 \\
&= a^2E(X^2) + 2abE(X) + b^2 - a^2E(X)^2 - 2abE(X) - b^2 \\
&= a^2(E(X^2) - E(X)^2) \\
&= a^2\operatorname{Var}(X).
\end{aligned}$$

Hence, it follows that

$$\sigma_Y = \sqrt{\operatorname{Var}(Y)} = \sqrt{a^2\operatorname{Var}(X)} = |a|\sigma_X.$$

From Lemma 6, we know that $\operatorname{Cov}(X, Y) = a\operatorname{Var}(X)$. Therefore,

$$\rho(X, Y) = \frac{\operatorname{Cov}(X, Y)}{\sigma_X\sigma_Y} = \frac{a\operatorname{Var}(X)}{|a|\sigma_X^2} = \frac{a}{|a|} = \operatorname{sgn}(a).$$

□

You'll often encounter the following matrix in Statistics, Machine Learning, Financial Mathematics / Economics courses.*

DEFINITION 42 (COVARIANCE MATRIX). Let $X_1, X_2, \ldots, X_n$ be jointly distributed random variables. The $n \times n$ matrix whose (i, j)t-h entry is

$$\operatorname{Cov}(X_i, X_j)$$

is called the covariance matrix of $X_1, X_2, \ldots, X_n$.

21 CONDITIONAL EXPECTATION AND CONDITIONAL VARIANCE

We have already seen conditional distribution of random variables in Section 18. In particular, if X and Y are jointly distributed discrete random variables, then the conditional PMF of X given Y is

$$\mathrm{P}(X = x \mid Y = y) = \frac{\mathrm{P}(X = x, Y = y)}{\mathrm{P}(Y = y)}$$

for all y such that $\mathrm{P}(Y = y) > 0$, whereas if X and Y are jointly distributed discrete random variables, then the conditional PDF of X given Y is

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

for all y such that $f_Y(y) > 0$.

*In particular, this matrix is essential in principal component analysis and dimensionality reduction of enormous data sets as well as in diversification of investment portfolios, which you'll learn in these later courses.

DEFINITION 43 (CONDITIONAL EXPECTATION OF A FUNCTION OF A DISCRETE RANDOM VARIABLE). Let X and Y be two jointly distributed discrete random variables and let $g : \mathbb{R} \to \mathbb{R}$ be a function. Let $y \in \mathbb{R}$ be such that $\mathrm{P}(Y = y) > 0$. If the series

$$\sum_{x\,:\,\mathrm{P}(X=x \mid Y=y)>0} g(x)\,\mathrm{P}(X = x \mid Y = y)$$

converges absolutely, then we define the *expectation* of $g(X)$ given $Y = y$ by

$$E(g(X) \mid Y = y) := \sum_{x\,:\,\mathrm{P}(X=x \mid Y=y)>0} g(x)\mathrm{P}(X = x \mid Y = y).$$

DEFINITION 44 (CONDITIONAL EXPECTATION OF A FUNCTION OF A CONTINUOUS RANDOM VARIABLE). Let X and Y be two jointly distributed continuous random variables and let $g : \mathbb{R} \to \mathbb{R}$ be a function. Let $y \in \mathbb{R}$ be such that $f_Y(y) > 0$. If the integral

$$\int_{-\infty}^{\infty} g(x) f_{X \mid Y}(x \mid y)dx$$

converges absolutely, then Then we define $E(g(X) \mid Y = y)$ by

$$E(g(X) \mid Y = y) = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x \mid y)dx.$$

Recall from Definition 36 that for a random variable X, we have

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

Now, we define conditional variance as follows:

DEFINITION 45 (CONDITIONAL VARIANCE). Let X and Y be two jointly distributed discrete (resp. continuous) random variables and let $y \in \mathbb{R}$ be such that $\mathrm{P}(Y = y) > 0$ (resp. $f_Y(y) > 0$). Then we define $\mathrm{Var}(X \mid Y = y)$ by

$$\mathrm{Var}(X \mid Y = y) = E(X^2 \mid Y = y) - E(X \mid Y = y)^2.$$

In particular, for two jointly distributed random variables X and Y and for a function $g : \mathbb{R} \to \mathbb{R}$, $E(g(X) \mid Y)$ is a function of the random variable Y whose value at $Y = y$ is $E(X \mid Y = y)$. So, we define $E(E(g(X) \mid Y))$ as the expectation of the function $E(g(X) \mid Y)$ of Y (see Definition 33 and Definition 34).

THEOREM 16. *For two jointly distributed random variables X and Y, we have*

$$E(X) = E(E(X \mid Y)).$$

PROOF.
Case 1. (Discrete case) From the absolute convergence of the series defining $E(E(X \mid Y))$ and from Dirichlet's theorem on absolutely convergent series, we

conclude that

$$\begin{aligned}
E(E(X \mid Y)) &= \sum_{y \,:\, \mathrm{P}(Y=y)>0} E(X \mid Y=y)\,\mathrm{P}(Y=y) \\
&= \sum_{y \,:\, \mathrm{P}(Y=y)>0} \; \sum_{x\,:\,\mathrm{P}(X=x \,\mid\, Y=y)>0} x\,\mathrm{P}(X=x \mid Y=y)\,\mathrm{P}(Y=y) \\
&= \sum_{x,y \,:\, \mathrm{P}(X=x,Y=y)>0} x\,\mathrm{P}(X=x,Y=y) \\
&= \sum_{x \,:\, \mathrm{P}(X=x)>0} \; \sum_{y\,:\,\mathrm{P}(Y=y \,\mid\, X=x)>0} x\,\mathrm{P}(Y=y \mid X=x)\,\mathrm{P}(X=x) \\
&= \sum_{x \,:\, \mathrm{P}(X=x)>0} x \sum_{y\,:\,\mathrm{P}(X=x,Y=y)>0} \mathrm{P}(X=x,Y=y) \\
&= \sum_{x \,:\, \mathrm{P}(X=x)>0} x\,\mathrm{P}(X=x) \\
&= E(X).
\end{aligned}$$

Case 2. (Continuous case) From the absolute convergence of the integral defining $E(E(X \mid Y))$ and from Fubini's theorem*, we conclude that

$$\begin{aligned}
E(E(X \mid Y)) &= \int_{-\infty}^{\infty} E(X \mid Y=y)\, f_Y(y) dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, f_{X|Y}(x \mid y)\, f_Y(y) dx dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, f_{X,Y}(x,y) dx dy \\
&= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy dx \\
&= \int_{-\infty}^{\infty} x f_X(x) dx \\
&= E(X).
\end{aligned}$$

□

THEOREM 17. *For two jointly distributed random variables X and Y, we have*

$$\mathrm{Var}(X) = E(\mathrm{Var}(X \mid Y)) + \mathrm{Var}(E(X \mid Y)).$$

PROOF. We have

$$\mathrm{Var}(E(X \mid Y)) = E(E(X \mid Y)^2) - E(E(X \mid Y))^2 = E(E(X \mid Y)^2) - E(X)^2,$$

*or by considering the absolute convergence of the Riemann sum whose limiting value is equal to the above integral and then rearranging the order of summation (by Dirichlet's theorem on absolutely convergent series)

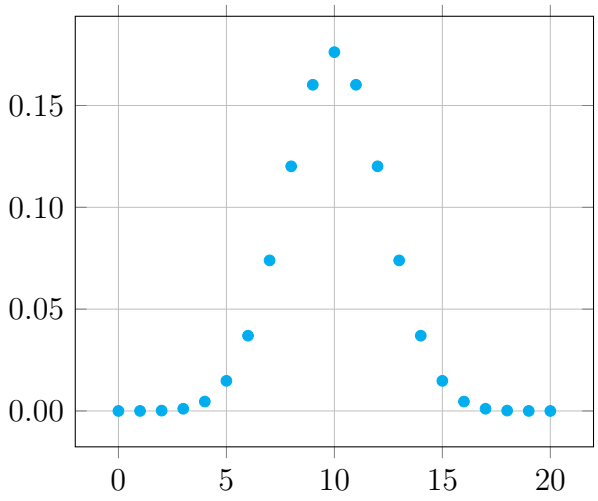
where the second equality follows from Theorem 17. Also, from Definition 45, we have

$$E(\mathrm{Var}(X \mid Y)) = E(E(X^2 \mid Y)) - E(E(X \mid Y)^2) = E(X^2) - E(E(X \mid Y)^2),$$

where the second equality follows again from Theorem 17. Now, the claim follows by adding the above two equations, since $\mathrm{Var}(X) = E(X^2) - E(X)^2$. □

22 Binomial distribution

Consider a random experiment where you toss a fair coin 20 times and let the random variable X denote the total number of heads obtained. Such a random variable X is a typical example of a binomial random variable. The following picture shows its the probability distribution.



Definition 46 (Binomial random variable). A binomial random variable X is a sum of n iid Bernoulli random variables, i.e.

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

where $X_i \sim \text{Bernoulli}(p)$ for all $i \in \{1, 2, \ldots, n\}$. The numbers n and p are called the *parameters* of Binomial distribution and we write

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

Clearly, we have $\mathrm{P}(X = r) = 0$ for all $r \notin \{0, 1, 2, \ldots, n\}$, whereas for $r \in \{0, 1, 2, \ldots, n\}$, the PMF of the X is given by

$$\begin{aligned} P(X = r) &= \sum_{\substack{S \subseteq \{1,2,\ldots,n\} \\ |S| = r}} \left(\prod_{i \in S} \mathrm{P}(X_i = 1) \right) \left(\prod_{j \notin S} \mathrm{P}(X_j = 0) \right) \\ &= \binom{n}{r} p^r (1-p)^{n-r}. \end{aligned}$$

From Example 18 and Corollary 7, it follows that if $X \sim \text{Binomial}(n, p)$, then

$$E(X) = np. \tag{22.1}$$

EXAMPLE 20. Suppose, on average you receive three WhatsApp messages per hour. Calculate the probability of obtaining at least one such message within the next fifteen minutes, assuming that no two messages arrive at the same minute.

ANSWER. Let $W \sim \text{Binomial}(60, p)$ denote the number of WhatsApp messages which you receive in an hour. Then (22.1) implies that $E(W) = 60p$ and since you receive three WhatsApp messages per hour, it follows that $E(W) = 3$. Therefore, we conclude that $p = 1/20$. Now, let X be a random variable denoting the number of WhatsApp messages which you receive in a 15 minute interval. Then

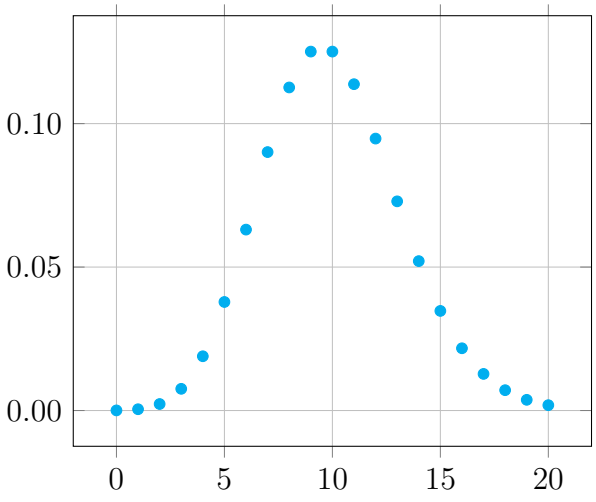
$$X \sim \text{Binomial}(15, 1/20).$$

Hence, the probability of obtaining at least one message within the next fifteen minutes is

$$\begin{aligned} \mathrm{P}(X > 0) &= 1 - \mathrm{P}(X = 0) = 1 - \binom{15}{0}\left(\frac{1}{20}\right)^0\left(1 - \frac{1}{20}\right)^{15} \\ &= 1 - \left(\frac{19}{20}\right)^{15} \approx 0.54\,. \end{aligned}$$

23 POISSON DISTRIBUTION

A random variables counting the number of occurrences of an event in a specified interval of time is a typical example of a Poisson random variable. For instance, the number of times you blink or nod during an hour-long lesson in Probability, the number of times your phone rings in a day, the number of Google searches on a specific topic per week or the number or the number of clicks on a particular website per hour - are all examples of Poisson random variables. Also, with Poisson distributions, we often approximate Binomial distributions. For example, if X denotes the number of heads obtained in 20 tosses of a fair coin, then the Poisson approximation of the distribution of the Binomial random variable X looks like the following.



Definition 47 (Poisson distribution). The Poisson distribution is the limiting distribution of the random variable $X_n \sim \text{Binomial}(n, p_n)$, where $n \to \infty$ but $E(X) = np_n = \lambda$ remains constant. The number $\lambda > 0$ is called the *parameter* of Poisson distribution and if X is a random variable that follows Poisson distribution with parameter λ, we write

$$X \sim \text{Poisson}(\lambda).$$

Clearly, we have $\mathrm{P}(X = r) = 0$ for all $r \notin \{0\} \cup \mathbb{N}$, whereas for $r \in \{0\} \cup \mathbb{N}$, the PMF of X is given by

$$\begin{aligned}
P(X = r) &= \lim_{n\to\infty} \mathrm{P}(X_n = r) \\
&= \lim_{n\to\infty} \binom{n}{r} p_n^r (1 - p_n)^{n-r} \\
&= \lim_{n\to\infty} \frac{n(n-1)\cdots(n-r+1)}{r!} \left(\frac{\lambda}{n}\right)^r \left(1 - \frac{\lambda}{n}\right)^{n-r} \\
&= \frac{\lambda^r}{r!} \lim_{n\to\infty} \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{r-1}{n}\right) \frac{(1 - \frac{\lambda}{n})^n}{(1 - \frac{\lambda}{n})^r} \\
&= e^{-\lambda} \frac{\lambda^r}{r!}.
\end{aligned}$$

Thus, for large n, the distribution of a random variable $X \sim \text{Binomial}(n, p)$ could be approximated by that of $Y \sim \text{Poisson}(\lambda)$, where $\lambda = np$.

Review of Example 20. Let X be a random variable denoting the number of WhatsApp messages you receive in a 15 minute interval. In Example 20, we saw that

$$X \sim \text{Binomial}(15, 1/20).$$

Though $n = 15$ is not much large, let's still approximate the distribution of X with

$$Y \sim \text{Poisson}(\lambda),$$

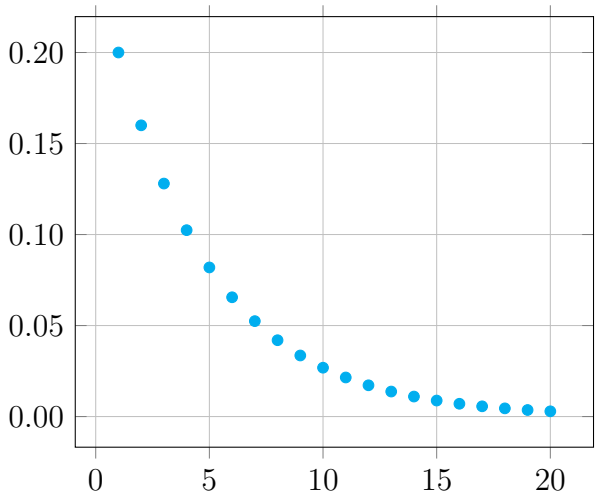
where $\lambda = 15/20 = 3/4$. Hence, the approximate probability of obtaining at least one message within the next 15 minutes is

$$\mathrm{P}(Y > 0) = 1 - \mathrm{P}(Y = 0) = 1 - e^{-\lambda}\frac{\lambda^0}{0!} = 1 - e^{-3/4} \approx 0.53\,.$$

24 Geometric distribution

Consider a random experiment where you keep on tossing a coin till you get a head*. Let X denote the number of tosses required till you get the first head. Then X is an example of a Geometric random variable. For $P(H) = 0.2$, the distribution of X looks like the following.

*Recall from Example 4 in Chapter 1 that the probability of eventually obtaining a head tends to 1 if you keep on tossing the coin.



DEFINITION 48 (GEOMETRIC RANDOM VARIABLE). Let $\{X_i\}_{i=1}^{\infty}$ be iid Bernoulli random variables with parameter p. We define the random variable X as $X = n$ if $X_n = 1$ and if $X_i = 0$ for all $i < n$. The number p is also called the *parameter* of the Geometric distribution and we write

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

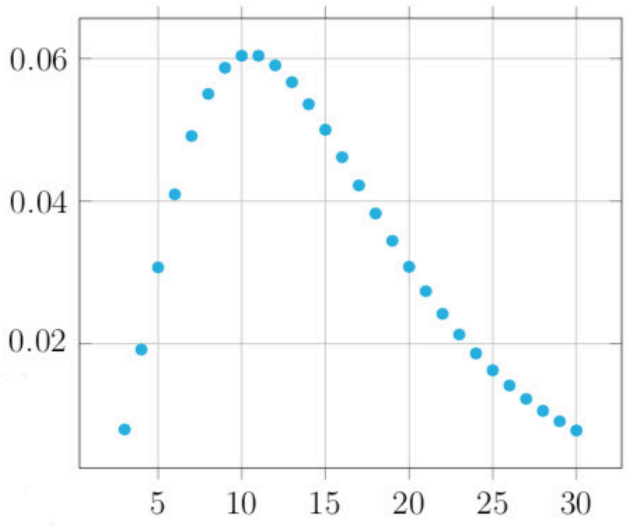
Clearly, for $r \notin \mathbb{N}$, we have $\mathrm{P}(X = r) = 0$, whereas for $r \in \mathbb{N}$ the PMF of X is given by

$$\begin{aligned}\mathrm{P}(X = r) &= \mathrm{P}(X_1 = 0, X_2 = 0, \ldots, X_{r-1} = 0, X_r = 1) \\ &= \mathrm{P}(X_1 = 0)\,\mathrm{P}(X_2 = 0) \cdots \mathrm{P}(X_{r-1} = 0)\,\mathrm{P}(X_r = 1) \\ &= (1-p)^{r-1}p,\end{aligned}$$

where the second equality follows from the mutual independence of the random variables X_i for $i \in \{1, 2, \ldots, r\}$.

25 PASCAL DISTRIBUTION

Consider a random experiment where you keep on tossing a coin till you get n heads. Let X denote the number of tosses required till you get the first n heads. Then X is an example of a Pascal/Negative binomial random variable. For $P(H) = 0.2$ and $n = 3$, the distribution of X looks like the following.



DEFINITION 49 (PASCAL/NEGATIVE BINOMIAL RANDOM VARIABLE). A Pascal/Negative binomial random variable X is a sum of n iid Geometric random variables, i.e.

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

where $X_i \sim \text{Geometric}(p)$ for all $i \in \{1, 2, \ldots, n\}$. The numbers n and p are called the *parameters* of Pascal/Negative binomial distribution and we write

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

Clearly, for $r \notin \{n, n+1, n+2, \ldots\}$, we have $\mathrm{P}(X = r) = 0$, whereas for $r \in \{n, n+1, n+2, \ldots\}$, the PMF of X is given by

$$\begin{aligned}
\mathrm{P}(X = r) &= \sum_{r_1 + r_2 + \cdots + r_n = r} \mathrm{P}(X_1 = r_1, X_2 = r_2, \ldots, X_n = r_n) \\
&= \sum_{r_1 + r_2 + \cdots + r_n = r} \mathrm{P}(X_1 = r_1)\,\mathrm{P}(X_2 = r_2) \cdots \mathrm{P}(X_n = r_n) \\
&= \sum_{r_1 + r_2 + \cdots + r_n = r} (1-p)^{r_1 - 1} p (1-p)^{r_2 - 1} p \cdots (1-p)^{r_n - 1} p \\
&= \sum_{r_1 + r_2 + \cdots + r_n = r} p^n (1-p)^{r-n} \\
&= \binom{r-1}{n-1} p^n (1-p)^{r-n},
\end{aligned}$$

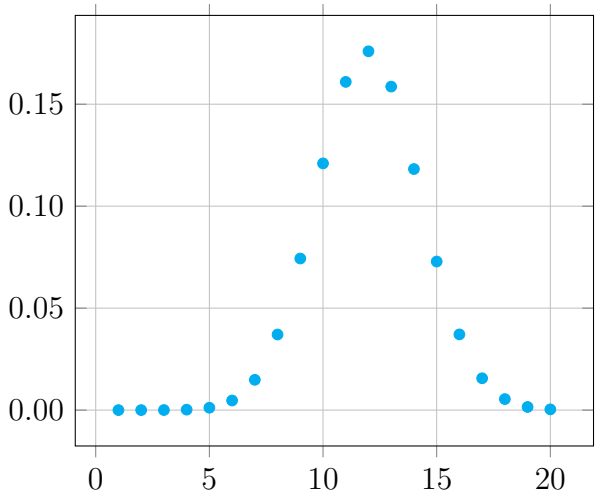
where the second equality follows from the mutual independence of the random variables $X_1, X_2, \ldots, X_n$ and the last equality follows from Exercise 29.

26 HYPERGEOMETRIC DISTRIBUTION

Consider a random experiment in which you choose* a sample of n balls from a bag containing N balls among which m are red and $N - m$ are white. Let the

*without replacement

random variable X denote the number of red balls selected in the sample. Then X is an example of a Hypergeometric random variable. For $N = 100, m = 40$ and $n = 30$, the distribution of X looks like the following.



DEFINITION 50 (HYPERGEOMETRIC RANDOM VARIABLE). For positive integers N, m and n with $m, n \leq N$, the random variable X whose PMF is given by

$$\mathrm{P}(X = r) = \begin{cases} \frac{\binom{m}{r}\binom{N-m}{n-r}}{\binom{N}{n}} & \text{if } r \in \{0, 1, 2, \ldots, n\} \\ 0 & \text{otherwise,} \end{cases}$$

is called a Hypergeometric random variable and we write

$$X \sim \mathrm{Hypergeometric}(N, m, n).$$

27 UNIFORM DISTRIBUTION

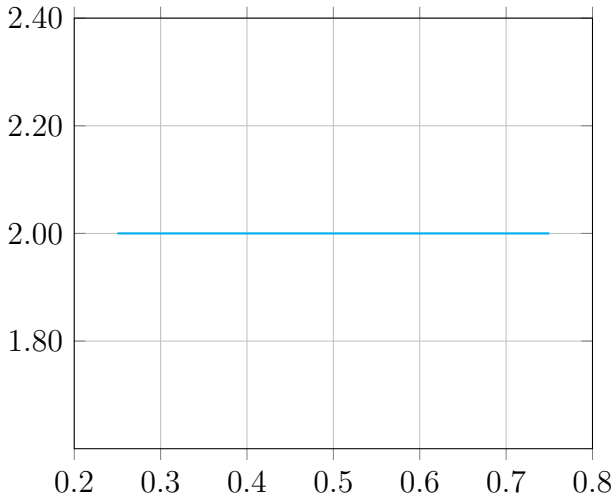
Let $(a, b) \subset \mathbb{R}$ be an open interval. Consider a random experiment where you mark a point in (a, b). Let X denote the point you marked. Then X is an example of a Continuous Uniform random variable which takes values in the interval (a, b). We write

$$X \sim \mathrm{U}(a, b).$$

DEFINITION 51 (UNIFORM DISTRIBUTION). A continuous random variable X is said to be uniformly distributed if the range of X is an interval of finite length on which the PDF of X is constant. Sine the integral of the PDF over the range of X is 1, it follows that the PDF of X is given by the reciprocal of the length of the interval, i.e. if $X \sim \mathrm{U}(a, b)$, then the PDF of X is

$$f_X(t) = \begin{cases} \frac{1}{b-a} & \text{if } a < t < b \\ 0 & \text{otherwise.} \end{cases}$$

For example, the following graph shows the PDF of $X \sim \mathrm{U}(1/4, 3/4)$.



Note that unlike PMF, the values taken by a PDF are not probabilities of events. So, as we see in the above example, a PDF can take values which are larger than 1. However, recall that a PDF is defined to be nonnegative. The reason behind this convention is that we mostly work with the continuous random variables whose CDFs are differentaible. Hence, we often take the PDF to be the derivative of the CDF. Since the CDF of a random variable is a nondecreasing function, its derivative is nonnegative. Hence, we adopted the convention that a PDF is a nonnegative function.

28 Exponential distribution

Suppose, on average you receive λ messages per hour on your phone. Let the random variable X denote the time from now till you receive the next message. Then X is an example of an Exponential random variable. The number $\lambda > 0$ is called the *parameter* of Exponential distribution.

Since on average you receive λ messages per hour, we may assume that on average you receive λt messages per t hours. Let the random variable Y_t denote the number of messages which you receive in t hours. Then

$$Y_t \sim \mathrm{Poisson}(\lambda t).$$

Note that waiting for more than t hours for the next message is the same as there being an interval of t hours in which you received no messages. Therefore, we have

$$\mathrm{P}(X > t) = \mathrm{P}(Y_t = 0) = e^{-\lambda t}. \tag{28.1}$$

That implies

$$F_X(t) = P(X \leq t) = 1 - \mathrm{P}(X > t) = 1 - e^{-\lambda t}.$$

Since F_X is differentaible, we define the PDF of X by

$$f_X(t) := F_X'(t) = \lambda e^{-\lambda t}.$$

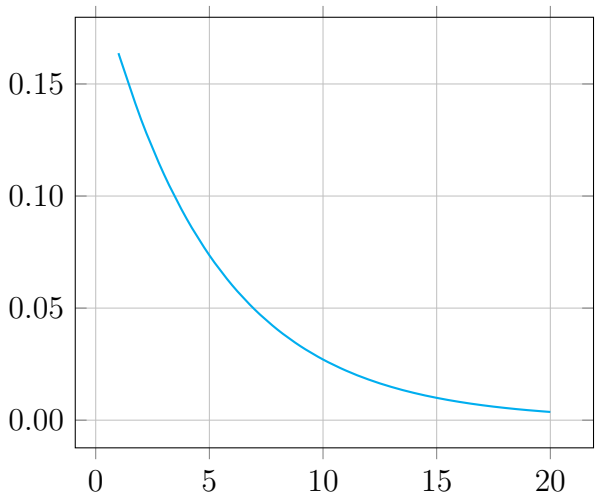
Definition 52 (Exponential random variable). For $\lambda > 0$, the continuous random variable X whose PDF is given by

$$f_X(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

is called an Exponential random variable and we write

$$X \sim \text{Exponential}(\lambda).$$

For example, the following graph shows the PDF of $X \sim \text{Exponential}(1/5)$.



29 Memorylessness

If the distribution of the waiting period until the occurrence of a certain event does not depend on the amount of time which has passed by already, then the distribution is called memoryless.

Definition 53 (Memoryless distribution). The probability distribution of a nonnegative random variable X is memoryless if for all s and t in the image of X, we have

$$\text{P}(X > s + t \mid X > s) = \text{P}(X > t).$$

A random variable with a memoryless probability distribution is called a memoryless random variable.

EXAMPLE 21. Let $X \sim \text{Geometric}(p)$. Then for all $m, n \in \mathbb{N}$, we have

$$\begin{aligned}\mathrm{P}(X > m+n | X > m) &= \frac{\mathrm{P}(X > m+n)}{\mathrm{P}(X > m)} = \frac{\sum_{j=0}^{\infty}(1-p)^{m+n+j}p}{\sum_{j=0}^{\infty}(1-p)^{m+j}p} \\ &= (1-p)^n = \sum_{j=0}^{\infty}(1-p)^{n+j}p \\ &= \mathrm{P}(X > n).\end{aligned}$$

Hence, the Geometric distribution is memoryless.

EXAMPLE 22. Let $X \sim \text{Exponential}(\lambda)$. Then for all $s, t \in \mathbb{R}_{\geq 0}$, we have

$$\begin{aligned}\mathrm{P}(X > s+t | X > s) &= \frac{\mathrm{P}(X > s+t)}{\mathrm{P}(X > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} \\ &= \mathrm{P}(X > t),\end{aligned}$$

where the second and the last equalities follow from (28.1). Hence, the Exponential distribution is memoryless.

THEOREM 18. *If X is a memoryless random variable with image $\mathbb{N}$, then X follows a Geometric distribution.*

PROOF. Since X is memoryless with image $\mathbb{N}$, we have

$$\mathrm{P}(X > m+n \mid X > m) = \mathrm{P}(X > m)$$

for all $m, n \in \mathbb{N}$. Since the left hand side of the above equation is equal to $\mathrm{P}(X > m+n)/\mathrm{P}(X > m)$, it follows that

$$\mathrm{P}(X > m+n) = \mathrm{P}(X > m)\mathrm{P}(X > n). \tag{29.1}$$

Substituting $m = n = 1$ in the above equation, we obtain $\mathrm{P}(X > 2) = \mathrm{P}(X > 1)^2$. It follows by induction from (29.1) that

$$\mathrm{P}(X > n) = \mathrm{P}(X > 1)^n \tag{29.2}$$

for all $n \in \mathbb{N}$. Let $\mathrm{P}(X = 1) = p$. Since X takes only positive integral values, it follows that $\mathrm{P}(X > 1) = 1 - p$. Hence, from (29.2) we obtain

$$\mathrm{P}(X > n) = (1-p)^n.$$

That implies

$$\begin{aligned}\mathrm{P}(X = n) &= \mathrm{P}(X > n-1) - \mathrm{P}(X > n) \\ &= (1-p)^{n-1} - (1-p)^n \\ &= (1-p)^{n-1}p,\end{aligned}$$

which is the PMF of the Geometric random variable. Thus, we conclude that

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

□

THEOREM 19. *If X is a memoryless random variable with image $\mathbb{R}_{\geq 0}$ such that $\mathrm{P}(X = 0) < 1$*, then X follows an Exponential distribution.*

PROOF. Since X is memoryless with image $\mathbb{R}_{\geq 0}$, we have

$$\mathrm{P}(X > s + t \mid X > s) = \mathrm{P}(X > t)$$

for all $s, t \in \mathbb{R}_{\geq 0}$. Since the left hand side of the above equation is equal to $\mathrm{P}(X > s + t)/\mathrm{P}(X > s)$, it follows that

$$\mathrm{P}(X > s + t) = \mathrm{P}(X > s)\mathrm{P}(X > t). \tag{29.3}$$

Substituting $s = t$ in the above equation, we obtain $\mathrm{P}(X > 2t) = \mathrm{P}(X > t)^2$. It follows by induction from (29.3) that

$$\mathrm{P}(X > nt) = \mathrm{P}(X > t)^n \tag{29.4}$$

for all $n \in \mathbb{N}$ and all $t \in \mathbb{R}_{\geq 0}$. Replacing t with t/n in the above equation, we obtain**

$$\mathrm{P}\left(X > \frac{t}{n}\right) = \mathrm{P}(X > t)^{1/n}. \tag{29.5}$$

Now, from (29.4) and (29.5), it follows that for all $m, n \in \mathbb{N}$ and for all $t \in \mathbb{R}_{\geq 0}$, we have

$$\mathrm{P}\left(X > \frac{mt}{n}\right) = \mathrm{P}\left(X > \frac{t}{n}\right)^m = \mathrm{P}(X > t)^{m/n},$$

by the uniqueness of the nonnegative n-th root of a nonnegative real number. Substituting $t = 1$ in the above equation, we obtain that

$$\mathrm{P}(X > r) = \mathrm{P}(X > 1)^r \tag{29.6}$$

for every positive rational number r.

Since the image of X is $\mathbb{R}_{\geq 0}$, we have $\mathrm{P}(X \geq 0) = 1$, i.e.

$$\mathrm{P}(X = 0) + \mathrm{P}(X > 0) = 1.$$

Since $\mathrm{P}(X = 0) < 1$, it follows that $\mathrm{P}(X > 0) > 0$. Substituting $s = t = 0$ in (29.3), we obtain $\mathrm{P}(X > 0) = \mathrm{P}(X > 0)^2$. Since $\mathrm{P}(X > 0)$ is nonzero, it follows that

$$\mathrm{P}(X > 0) = 1. \tag{29.7}$$

For $x > 0$, let $\{r_n\}$ be an increasing sequence of positive rational numbers converging to x. Let $\varepsilon_n := x - r_n$ for all $n \in \mathbb{N}$. Since $r_n \to x$, $\varepsilon_n \to 0$. Now, (29.3) and (29.6) together imply that

$$\mathrm{P}(X > x) = \mathrm{P}(X > r_n)\,\mathrm{P}(X > x - r_n) = \mathrm{P}(X > 1)^{r_n}\,\mathrm{P}(X > \varepsilon_n).$$

taking the limit as $n \to \infty$, we obtain

$$\mathrm{P}(X > x) = \left(\lim_{n\to\infty} \mathrm{P}(X > 1)^{r_n}\right) \cdot \left(\lim_{n\to\infty} \mathrm{P}(X > \varepsilon_n)\right). \tag{29.8}$$

*Note that, this condition holds trivially if X is a continuous random variable.

**by the uniqueness of the nonegative n-th root of a nonnegative real number.

Since the exponential function is continuous, it follows that

$$\lim_{n\to\infty} \mathrm{P}(X > 1)^{r_n} = \mathrm{P}(X > 1)^x.$$

Since F_X is right continuous and since $\{\varepsilon_n\}$ is decreasing to 0, we have

$$\lim_{n\to\infty} \mathrm{P}(X > \varepsilon_n) = \lim_{n\to\infty} \big(1 - F_X(\varepsilon_n)\big) = 1 - F_X(0) = \mathrm{P}(X > 0) = 1,$$

where the last equality follows from (29.7). Hence, (29.8) implies that

$$\mathrm{P}(X > x) = \mathrm{P}(X > 1)^x \text{ for all } x > 0. \tag{29.9}$$

In particular, if $\mathrm{P}(X > 1) = 0$, then $\mathrm{P}(X > x) = 0$ for all $x > 0$. Let $\{\alpha_n\}$ be a decreasing sequence converging to 0. As before, by the right continuity of F_X, we have

$$\mathrm{P}(X > 0) = \lim_{n\to\infty} \mathrm{P}(X > \alpha_n) = 0,$$

which contradicts (29.7). Hence, $\mathrm{P}(X > 1) > 0$.

Again, if $\mathrm{P}(X > 1) = 1$, then from (29.9), we have $\mathrm{P}(X > x) = 1$ for all $x \geq 0$. That implies

$$\lim_{x\to\infty} \mathrm{P}(X > x) = 1.$$

However, for any random variable X, we have $\lim_{x\to\infty} F_X(x) = 1$, which implies $\lim_{x\to\infty} \mathrm{P}(X > x) = 0$. Thus, we get a contraction! Therefore, we have $\mathrm{P}(X > 1) \in (0, 1)$ and hence, $\log \mathrm{P}(X > 1) \in (-\infty, 0)$. Let $\lambda > 0$ be such that $\log \mathrm{P}(X > 1) = -\lambda$. Then from (29.9), we obtain

$$\mathrm{P}(X > x) = e^{-\lambda x}.$$

Hence, from (28.1) we conclude that

$$X \sim \mathrm{Exponential}(\lambda).$$

□

30 Normal distribution

The average of a large number of iid random variables having finite expectation and variance follows a nearly Normal distribution. For instance, the distributions of the average age of the students in a batch at which they gain economic independence or the average heights of men (or women) in an Indian city or the average amount of daily rainfall at Cherrapunji in a year or the average distance from the bull's eye of a hundred arrows shot on a target by an archer are all closely approximated by Normal distributions. The underlying reason for this is the Central Limit Theorem* which states that if $X_1, X_2, \ldots, X_n$ are iid random variables with finite expectation and variance, then** the probability distribution of their average

$$\overline{X}_n := \frac{X_1 + X_2 + \ldots + X_n}{n}$$

converges to a Normal distribution as $n \to \infty$. Hence, with Normal distribution we often approximate the distribution of the average of iid random variables.

*We shall outline a proof this theorem in a later chapter.

**irrespective of their initial distribution

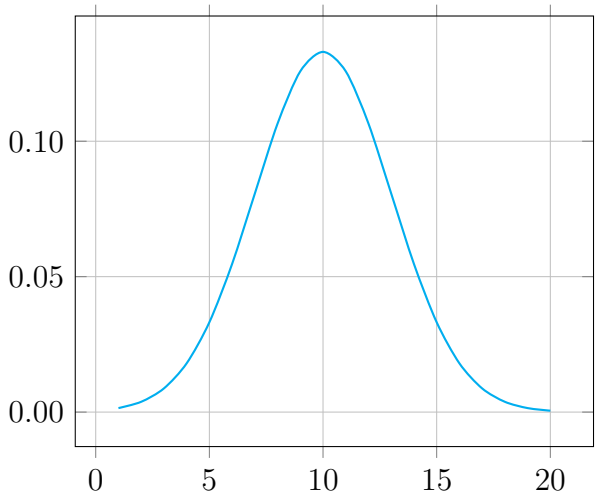
DEFINITION 54 (NORMAL RANDOM VARIABLE). For $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_{>0}$, the continuous random variable X whose PDF is given by

$$f_X(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}},$$

is called a Normal random variable and we write

$$X \sim \mathrm{N}(\mu, \sigma^2).$$

The numbers μ and σ^2 are called the parameters of Normal distribution. For example, the following graph shows the PDF of $X \sim \mathrm{N}(10, 9)$.



THEOREM 20. *Let $X \sim \mathrm{N}(\mu, \sigma^2)$, let $a, b \in \mathbb{R}$ such that $a \neq 0$ and let $Y := aX + b$. Then $Y \sim \mathrm{N}(a\mu + b, a^2\sigma^2)$.*

PROOF. The CDF of the random variable Y is given by

$$\begin{aligned} F_Y(y) &= \mathrm{P}(Y \leq y) \\ &= \mathrm{P}(aX + b \leq y) \\ &= \begin{cases} \mathrm{P}\left(X \leq \frac{y-b}{a}\right) & \text{if } a > 0 \\ \mathrm{P}\left(X \geq \frac{y-b}{a}\right) & \text{if } a < 0 \end{cases} \\ &= \begin{cases} F_X\left(\frac{y-b}{a}\right) & \text{if } a > 0 \\ 1 - F_X\left(\frac{y-b}{a}\right) & \text{if } a < 0. \end{cases} \end{aligned}$$

In particular, since F_X is differentaible, so is F_Y. Hence, we define the PDF of Y by

$$\begin{aligned} f_Y(t) = F_Y'(t) &= \frac{1}{|a|} f_X\left(\frac{t-b}{a}\right) \\ &= \frac{1}{|a|\sigma\sqrt{2\pi}} e^{-\frac{\left(\frac{t-b}{a}-\mu\right)^2}{2\sigma^2}} \\ &= \frac{1}{|a|\sigma\sqrt{2\pi}} e^{-\frac{(t-(a\mu+b))^2}{2a^2\sigma^2}}. \end{aligned}$$

Thus, we see that Y follows a Normal distribution with parameters $a\mu + b$ and $a^2\sigma^2$. □

COROLLARY 9. *Let $Z \sim \mathrm{N}(\mu, \sigma^2)$ and let $X := \frac{Z-\mu}{\sigma}$. Then $X \sim \mathrm{N}(0, 1)$.*

DEFINITION 55 (STANDARD NORMAL RANDOM VARIABLE). The random variable $X \sim \mathrm{N}(0, 1)$ is called the Standard normal random variable and its CDF is usually denoted by Φ, i.e.

$$\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt. \tag{30.1}$$

In particular, as $x \to \infty$, $\Phi(x) = \mathrm{P}(X \leq x)$ must tend to 1. We prove this fact in the following.

THEOREM 21. *Let $\Phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be defined by* (30.1). *Then*

(*i*) $\lim_{x\to\infty} \Phi(x) = 1$.

(*ii*) $\Phi(-x) = 1 - \Phi(x)$ *for all* $x \in \mathbb{R}$.

PROOF. (*i*) Note that the limit $\lim_{x\to\infty} \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt$ exists if and only if the so does $\lim_{n\to\infty} \int_{-n}^{n} e^{-\frac{t^2}{2}} dt$. Now, $e^{-t^2/2} \in (0, 1]$ for all $t \in \mathbb{R}$. In particular, for $t \geq 2$, we have $e^{-t^2/2} \leq e^{-t}$. For every integer $n \geq 2$, let $a_n := \int_{-n}^{n} e^{-\frac{t^2}{2}} dt$. Then $\{a_n\}_{n=2}^{\infty}$ is an increasing sequence with

$$0 \leq a_n \leq \int_{-2}^{2} dt + 2\int_{2}^{n} e^{-t} dt < 4 + \frac{2}{e^2}.$$

Hence, the Monotone convergence theorem implies that $\lim_{n\to\infty} a_n \geq 0$ exists. Therefore, the limit

$$\ell := \lim_{x\to\infty} \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt \geq 0$$

also exists. Since the above integral converges and since $e^{-\frac{t^2}{2}}$ is positive for all $t \in \mathbb{R}$, the above integral converges absolutely. Hence, by Tonelli's theorem* we

*Or by considering the absolute convergence of the Riemann sum whose limiting value is equal to the above integral and then rearranging the order of summation (by Dirichlet's theorem on absolutely convergent series) in the double series whose limiting value defines $2\pi\ell^2$.

have

$$\begin{aligned} 2\pi\ell^2 &= \left(\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx\right)\cdot\left(\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}\,dy\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}}\,dxdy. \end{aligned}$$

Now, we substitute $(x, y) \in \mathbb{R}^2$ by $(r\cos\theta, r\sin\theta)$ for $r \in \mathbb{R}_{\geq 0}$ and $\theta \in [0, 2\pi)$. Since the change of variables from cartesian coordinates to polar coordinates is a differentable bijection from $\mathbb{R}^2 \setminus \{0\}$ to $\mathbb{R}_{>0} \times [0, 2\pi)$, locally the infinitesemal changes in x or y are translated to infinitesemal changes in r and θ by the invertible linear map induced by left multiplication with the Jacobian matrix

$$J := \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial r} r\cos\theta & \frac{\partial}{\partial \theta} r\cos\theta \\ \frac{\partial}{\partial r} r\sin\theta & \frac{\partial}{\partial \theta} r\sin\theta \end{pmatrix} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}.$$

Since a linear map T on $\mathbb{R}^2$ transforms the area of the unit square by $|\det T|$, it follows that the infinitesemal area represented by $dxdy$ changes to $|\det J| drd\theta = rdrd\theta$. Thus, we obtain

$$\begin{aligned} 2\pi\ell^2 &= \int_0^{\infty}\int_0^{2\pi} e^{-\frac{r^2}{2}} r d\theta dr \\ &= 2\pi\int_0^{\infty} re^{-\frac{r^2}{2}}\,dr \\ &= 2\pi\int_0^{\infty} e^{-s} ds \\ &= 2\pi, \end{aligned}$$

where the third equality follows by substituting $s = r^2/2$. Since ℓ is nonnegative, it follows from the above equality that

$$\ell = \lim_{x\to\infty} \Phi(x) = 1.$$

(ii) We have

$$\begin{aligned} \Phi(-x) &= \int_{-\infty}^{-x} e^{-\frac{t^2}{2}}\,dt \\ &= \int_x^{\infty} e^{-\frac{t^2}{2}}\,dt \\ &= \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}}\,dt - \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\,dt \\ &= 1 - \Phi(x). \end{aligned}$$

□

31 Moments of the standard normal random variable

Theorem 22. *Let X be the standard normal random variable. For $n \in \mathbb{N}$, we have*

$$E(X^n) = \begin{cases} 0 & \text{if } n \text{ is odd} \\ (n-1)!! & \text{if } n \text{ is even.} \end{cases}$$

Proof. Recall from Definition 34 and Definition 35 that the n-th moment of X exists if and only if the integral

$$\int_{-\infty}^{\infty} |t^n| e^{-t^2/2} dt$$

converges, i.e. if and only if the limit $\lim_{m\to\infty} \int_{-m}^{m} |t^n| e^{-\frac{t^2}{2}} dt$ exists. Now, $e^{-t^2/2} \in (0,1]$ for all $t \in \mathbb{R}$. In particular, given $n \in \mathbb{N}$, for all $t \geq 2(n+1)$, we have $e^{-t^2/2} \leq e^{-(n+1)t} \leq e^{-(t+n\log|t|)}$, i.e.

$$|t^n| e^{-t^2/2} \leq e^{-t} \text{ for all } t \geq 2(n+1).$$

For every integer $m \geq 2(n+1)$, let $a_m := \int_{-m}^{m} |t^n| e^{-\frac{t^2}{2}} dt$. Then $\{a_m\}_{m=2(n+1)}^{\infty}$ is an increasing sequence with

$$0 \leq a_m \leq \int_{-2(n+1)}^{2(n+1)} |t^n| dt + 2\int_{2(n+1)}^{m} e^{-t} dt < 2^{n+2}(n+1)^n + \frac{2}{e^{2(n+1)}}.$$

Hence, given $n \in \mathbb{N}$, the Monotone convergence theorem implies that $\lim_{m\to\infty} a_m$ exists. Therefore, for all $n \in \mathbb{N}$, the n-th moment of X exists and we have

$$E(X^n) = \int_{-\infty}^{\infty} t^n e^{-t^2/2} dt = \lim_{m\to\infty} \int_{-m}^{m} t^n e^{-\frac{t^2}{2}} dt.$$

Note that if n is odd, then the integrand is an odd function. Since the integral of an odd function over a symmetric range around zero is 0, it follows that

$$E(X^n) = 0 \text{ if } n \text{ is odd.}$$

For even n, we proceed by integration by parts as follows.

$$\begin{aligned} E(X^n) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (-t^{n-1})(-te^{-t^2/2}) dt \\ &= \frac{1}{\sqrt{2\pi}} \left((-t^{n-1}) e^{-t^2/2} \Big|_{-\infty}^{\infty} + (n-1) \int_{-\infty}^{\infty} t^{n-2} e^{-t^2/2} dt \right) \\ &= \frac{1}{\sqrt{2\pi}} \left(0 + (n-1)\sqrt{2\pi} E(X^{n-2}) \right) \\ &= (n-1) E(X^{n-2}). \end{aligned}$$

In particular, for $n = 2$ we have, $E(X^2) = E(1) = 1$. Hence, by induction, it follows that

$$E(X^n) = (n-1) \cdot (n-3) \cdots 5 \cdot 3 \cdot 1 = (n-1)!! \text{ if } n \text{ is even.}$$

□

COROLLARY 10. *Let* $X \sim \mathrm{N}(0,1)$. *Then* $E(X) = 0$ *and* $\mathrm{Var}(X) = 1$.

In particular, Corollary 9 and Theorem 22 together imply that

COROLLARY 11 (CENTRAL MOMENTS OF A NORMAL RANDOM VARIABLE). *Let* $X \sim \mathrm{N}(\mu, \sigma^2)$. *Then*

$$E((X-\mu)^n) = \begin{cases} 0 & \text{if } n \text{ is odd} \\ (n-1)!!\,\sigma^n & \text{if } n \text{ is even.} \end{cases}$$

COROLLARY 12. *Let* $X \sim \mathrm{N}(\mu, \sigma^2)$. *Then* $E(X) = \mu$ *and* $\mathrm{Var}(X) = \sigma^2$.

COROLLARY 13. *Let* $X \sim \mathrm{N}(\mu, \sigma^2)$. *Then*

$$E(X^2) = \mu^2 + \sigma^2 \;\; \textit{and} \;\; E(X^3) = \mu^3 + 3\mu\sigma^2.$$

PROOF. From Corollary 12, we know that $E(X) = \mu$ and $\mathrm{Var}(X) = E(X^2) - E(X)^2 = \sigma^2$. It follows that

$$E(X^2) = \mu^2 + \sigma^2.$$

Again, from Corollary 11, we know that $E((X-\mu)^3) = 0$. Expanding $(X-\mu)^3$, we obtain

$$E(X^3) - 3\mu E(X^2) + 3\mu^2 E(X) - \mu^3 = 0.$$

Since $E(X) = \mu$ and $E(X^2) = \mu^2 + \sigma^2$, it follows from the above equation that

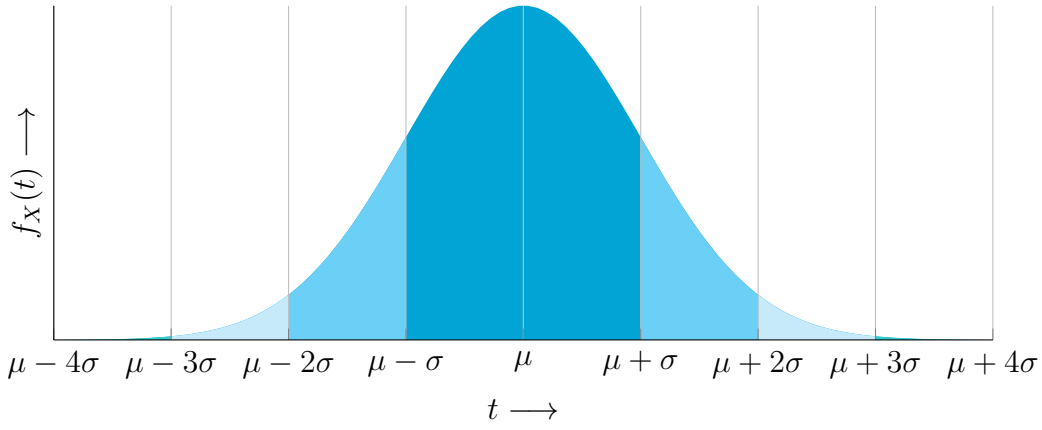
$$E(X^3) = \mu^3 + 3\mu\sigma^2.$$

□

In general, expanding $E((X-\mu)^n)$ and using Corollary 11, we may compute $E(X^n)$ if we know the moments $E(X^j)$ for all $j \in \{1, 2, \ldots, n-1\}$.

32 THE 3-SIGMA RULE FOR NORMAL RANDOM VARIABLES

The 3-σ rule tells that the probability of a random variable $X \sim \mathrm{N}(\mu, \sigma^2)$ taking a value that is away from μ by 3σ or more is less than 0.0027. This probability is represented in the following picture by the area under the curve of $f_X(t)$ outside the region from $\mu - 3\sigma$ to $\mu + 3\sigma$. Since the probability of the random variable X taking a value outside the region $(\mu - 3\sigma, \mu + 3\sigma)$ is very small, this interval is called the *effective range* of the Normal distribution.



THEOREM 23. *Let a random variable $X \sim \mathrm{N}(\mu, \sigma^2)$. Then for all $t > 0$, we have*

$$\mathrm{P}(|X - \mu| \leq t\sigma) = 2\Phi(t) - 1,$$

where Φ denotes the CDF of Standard Normal random variable.

PROOF. Define $Z := \frac{X-\mu}{\sigma}$. Then Corollary 9 implies that $Z \sim \mathrm{N}(0, 1)$. Hence, we have

$$\begin{aligned}
\mathrm{P}(|X - \mu| \leq t\sigma) &= \mathrm{P}(-t\sigma \leq |X - \mu| \leq t\sigma) \\
&= \mathrm{P}(-t \leq Z \leq t) \\
&= \Phi(t) - \Phi(-t) \\
&= \Phi(t) - (1 - \Phi(t)) \\
&= 2\Phi(t) - 1,
\end{aligned}$$

where the fourth identity follows from Theorem 21.(ii). □

From a table* of the values of Φ, we see that

$$\Phi(1) = 0.84134,\ \Phi(2) = 0.97725,\ \Phi(3) = 0.99865.$$

Hence, from the above theorem, we obtain the following.

COROLLARY 14 (THE 3-σ RULE / THE 68-95-99.7 RULE). *Let a random variable $X \sim \mathrm{N}(\mu, \sigma^2)$. Then*

$$\mathrm{P}(|X - \mu| \leq \sigma) = 0.68268,$$

(This is equal to the area under the curve of $f_X(t)$ for $\mu - \sigma \leq t \leq \mu + \sigma$.)

$$\mathrm{P}(|X - \mu| \leq 2\sigma) = 0.9545,$$

(This is equal to the area under the curve of $f_X(t)$ for $\mu - 2\sigma \leq t \leq \mu + 2\sigma$).

$$\mathrm{P}(|X - \mu| \leq 3\sigma) = 0.9973.$$

(This is equal to the area under the curve of $f_X(t)$ for $\mu - 3\sigma \leq t \leq \mu + 3\sigma$.)

*or from the app *Probability Distributions* by Matthew Bognar ©2020

33 Markov's and Chebyshev's inequalities

Theorem 24 (Markov's inequality). *Let X be a random variable. Then for all $a > 0$, we have*

$$\mathrm{P}(|X| \geq a) \leq \frac{E(|X|)}{a}.$$

Proof. Let $p = \mathrm{P}(|X| \geq a)$ and define the random variable $\widehat{X}$ by

$$\widehat{X} = \begin{cases} 1 & \text{if } |X| \geq a \\ 0 & \text{otherwise.} \end{cases}$$

Then $\widehat{X} \sim \mathrm{Bernoulli}(p)$ and we have $\widehat{X} \leq \dfrac{|X|}{a}$. Hence, we conclude

$$\mathrm{P}(|X| \geq a) = p = E(\widehat{X}) \leq \frac{E(|X|)}{a}.$$

□

Corollary 15 (Chebyshev's inequality). *If Y is a random variable with expectation μ and variance σ^2, then for all $b > 0$, we have*

$$\mathrm{P}(|Y - \mu| \geq b) \leq \frac{\sigma^2}{b^2}.$$

Proof. Follows immediately by substituting $X = (Y - \mu)^2$ and $a = b^2$ in Thereom 24 and by noting that $\mathrm{P}(|Y - \mu| \geq b) = \mathrm{P}(|X| \geq a)$. □

34 The t-sigma rule for arbitrary random variables

Chebyshev's inequality implies an analog of Theorem 23 for any random variable with finite expectation and variance.

Theorem 25. *Let X be a random variable with expectation μ and variance σ^2. Then for all $t > 0$ we have*

$$\mathrm{P}(|X - \mu| < t\sigma) \geq 1 - \frac{1}{t^2}.$$

Proof. Substituting $Y = X$ and $b = t\sigma$ in Corollary 15, we obtain $\mathrm{P}(|X - \mu| \geq t\sigma) < 1/t^2$. Now, subtracting both sides from 1, we obtain

$$\mathrm{P}(|X - \mu| < t\sigma) \geq 1 - \frac{1}{t^2}.$$

□

Corollary 16. *Let X be a random variable with expectation μ and variance σ^2. Then*

$$\mathrm{P}(|X - \mu| < 2\sigma) \geq 0.75, \quad \mathrm{P}(|X - \mu| < 3\sigma) \geq 0.88$$
$$\mathrm{P}(|X - \mu| < 5\sigma) \geq 0.96, \quad \mathrm{P}(|X - \mu| < 10\sigma) \geq 0.99.$$

DEFINITION 56 (POPULATION AND RANDOM SAMPLE). The set of all possible values assumed by a random variable is called a *population* and so, the expectation and the variance of the random variable are called *population mean* and *population variance*. A *sample* is a finite set of values assumed by a random variable. A *random sample* of size n from a population with mean μ and variance σ^2 is a set of n iid random variables $\{X_1, \ldots, X_n\}$ with expectation μ and variance σ^2. We define the *sample mean* by

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{35.1}$$

It follows easily that

$$E(\overline{X}_n) = \mu \text{ and } \operatorname{Var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

THEOREM 26. *Let $X_1, \ldots, X_n$ be iid random variables with expectation μ and variance σ^2. Let $\overline{X}_n$ be defined as in* (35.1)*. Then for all $\varepsilon > 0$, we have*

$$\mathrm{P}(|\overline{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

PROOF. Follows immediately by substituting $Y = \overline{X}_n$ and $b = \varepsilon$ in Corollary 15. □

COROLLARY 17 (THE WEAK LAW OF LARGE NUMBERS). *Let $X_1, \ldots, X_n$ be iid random variables with expectation μ and variance σ^2 and let $\overline{X}_n$ be defined as in* (35.1)*. Then for all $\varepsilon > 0$, we have*

$$\lim_{n\to\infty} \mathrm{P}(|\overline{X}_n - \mu| \geq \varepsilon) = 0.$$

In other words, the probability of the sample mean getting arbitrarily close to the population mean tends to 1 as the sample size tends to infinity.

Exercises

1. Let $(\Omega, \mathcal{E}, P)$ be a probability space and let $A \subset \Omega$. Define $X : \Omega \to \mathbb{R}$ by
$$X(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise.} \end{cases}$$
In which of the following cases is X a random variable?
(i) $A \in \mathcal{E}$ $\quad$ (ii) $A \notin \mathcal{E}$.

2. Let $(\Omega, \mathcal{E}, P)$ be a probability space with $\Omega = \{1, 2, 3, 4, 5\}$ and $\mathcal{E} = \{\emptyset, \Omega, \{1\}, \{2, 3, 4, 5\}\}$. Define $X \colon \Omega \to \mathbb{R}$ by
$$X(\omega) = \omega + 1$$
for all $\omega \in \Omega$. Is X a random variable? Justify your answer!

3. Show that a bounded monotone continuous function is uniformly continuous.

4. Show that no monotone function has a discontinuity of the second kind.

5. Show that a monotone function has at most countably many discontinuities.

6. Let X be a random variable taking values in $\{1, 2, \ldots, 10\}$ with PMF $f(x) = ax + b$ and expectation 7. Find a and b.

7. For what value of the constant c, the real valued function $f : \mathbb{R} \to \mathbb{R}$ given by
$$f(x) = \frac{c}{1 + (x - \theta)^2},$$
where θ is a real parameter, is a PDF of random variable X.

8. Let $p \in [0, 1]$, $a, b \in \mathbb{R}$ with $a > b$ and let X be a random variable such that
$$P(X = a) = p \text{ and } P(X = b) = 1 - p.$$
Find the expectation and variance of $\frac{X-b}{b-a}$.

9. A bag contains five coins, two of which are made of gold and the rest are made of silver. Consider the random experiment in which the coins are drawn out of the bag randomly, one after another, without replacement. Let X denote the number of draws until the last gold coin is drawn. Find the PMF, the CDF and the expectation of the random variable X.

10. Given a sequence $\{a_n\}$ of positive real numbers such that $\sum_{n=1}^{\infty} a_n = 1$ and a sequence $\{x_n\}_{n=1}^{\infty}$ of distinct real numbers, define
$$F(x) := \sum_{n \in \mathbb{N} \,:\, x_n \leq x} a_n.$$
Show that F is nondecreasing, the discontinuities of F is given by $\{x_n\}_{n=1}^{\infty}$ and $\lim_{x \to \infty} F(x) = 1$. Conclude that F is the CDF of a discrete probability distribution.

11. Explain why the proof of the right continuity of the cumulative density function can not be adopted to prove also left continuity of the same function.

12. Show that if the cumulative distribution function F_X of a random variable X is a step function, then X is a discrete random variable.

13. Provide an example of a continuous random variable without a PDF.

14. Provide an example of a continuous random variable with infinitely many PDFs.

15. Provide examples of a discrete and a continuous random variable such that their expectations do not exist.

16. Show that the CDF of a continuous random vector is continuous.

17. Show that if the n-th moment exists for some $n \in \mathbb{N}$, then so does the m-th moment for all $m \in \{1, 2, \ldots, n\}$.

18. Let $X \sim \text{Uniform}(n)$. Find $E(X)$ and $\text{Var}(X)$.

19. Let $f(t) = a_n t^n + \cdots + a_1 t + a_0$ be a polynomial with real coefficients and let X be a random variable such that its n-th moment exists. Show that
$$E(f(X)) = a_n E(X^n) + \cdots + a_1 E(X) + a_0.$$

20. Recall that for two random variables X and Y, we defined
$$\text{Cov}(X, Y) := E(XY) - E(X)E(Y).$$
   (i) Show that $\text{Cov}(X, Y) = E((X - E(X)(Y - E(Y))$.
   (ii) For random variables $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$, show that
$$\text{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \text{Cov}(X_i, Y_j)$$
   for all $a_i, b_j \in \mathbb{R}$.
   (iii) Conclude from the above that for random variables $X_1, X_2, \ldots, X_n$, we have
$$\text{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \text{Var}(X_i) + 2 \sum_{1 \le i < j \le n} a_i a_j \text{Cov}(X_i, X_j)$$
   for all $a_i \in \mathbb{R}$.

21. Show that if the correlation coefficient of two random variables X and Y has absolute value 1, then almost surely there is a linear relation between X and Y.

22. (a) Let $X_1, X_2, \ldots, X_n$ be jointly distributed random variables and let A dnote their covariance matrix. Show that the matrix A is *positive semidefinite*, i.e. for all $v \in \mathbb{R}^n$,
$$v^{\mathrm{T}} A v \ge 0.$$

(b) Using Part (a), show that no three jointly distributed Bernoulli(0.34) random variables X, Y, Z satisfy the equation

$$P(XY = 1) = P(YZ = 1) = P(ZX = 1) = 0.003.$$

(c) Without using Part (a), show that no three jointly distributed Bernoulli(0.34) random variables X, Y, Z satisfy the above equation.

23. Let $X \sim$ Binomial(n, p). Show that as k goes from 0 to n, $\mathrm{P}(X = k)$ first increases monotonically and then decreases monotonically, attaining its maximum at $k = \lfloor (n + 1)p \rfloor$.

24. Let $X \sim$ Poisson(λ) for some $\lambda > 0$. Show that as k goes from 0 to ∞, $\mathrm{P}(X = k)$ first increases monotonically and then decreases monotonically, attaining its maximum at $k = \lfloor \lambda \rfloor$.

25. Let X be a random variable on a sample space Ω such that $E(X) \in X(\Omega)$. Is it true that $\mathrm{P}(X = E(X)) \geq P(X = k)$ for all $k \in \mathbb{R}$? Justify your answer!

26. Let X be a discrete random variable assuming only nonnegative integer values and let F denote the CDF of X. Show that

$$E(X) = \sum_{x=0}^{\infty} (1 - F(x)).$$

27. Let $X_1, X_2, \ldots X_n$ be mutually independent random variables such that $X_i \sim$ Poisson(λ_i) for all $i \in \{1, 2, \ldots, n\}$. Let $X := X_1 + X_2 + \cdots + X_n$. Show that

$$X \sim \text{Poisson}(\lambda_1 + \lambda_2 + \cdots + \lambda_n).$$

28. For $X \sim$ Poisson(λ), determine $E(X(X - 1)(X - 2)(X - 3)(X - 4))$.

29. Show that the number of solutions of the equation

$$x_1 + x_2 + \cdots + x_n = r$$

in positive integers is $\binom{r-1}{n-1}$.

30. Let X denote the number of heads obtained when a fair coin is flipped 100 times. Show that

$$P(40 < X < 60) \geq \frac{3}{4}.$$

31. (a) Given $\tau > 1$, provide an example of a random variable X with $\mathrm{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$ such that

$$P(|X - \mu| < \tau\sigma) = 1 - \frac{1}{\tau^2}.$$

(b) Given $\mu \in \mathbb{R}$, provide an example of a random variable X with $\mathrm{E}(X) = \mu$ and

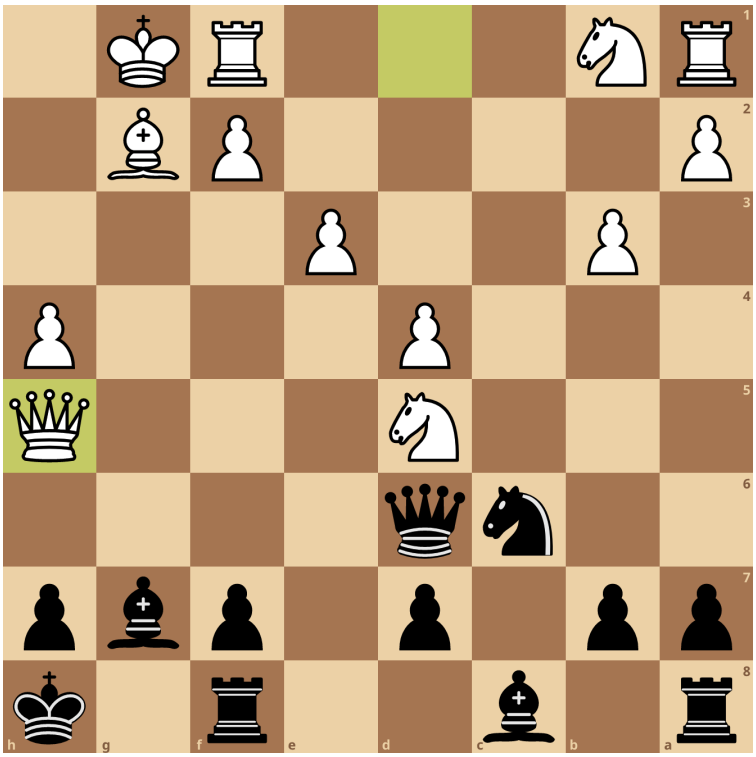
$$P(|X - \mu| < t) = 1 - \frac{1}{t^2}$$

for all $t > 1$.

Markov chains

Many systems have the property that given the present state, the past states have no influence on the future. This property is called the Markov property and systems having this property are called Markov chains. For example, in school, the probability of a student getting promotion to the next grade depends only on his or her performance in the present grade. In a knockout tournament, whether a team can play the next match depends only on whether it wins the present match. Also, in the board games like ludo or snakes-and-ladders, only the present state of the board is relevant for the next state, regardless of whatever its past states might have been.

Note that the number of configurations of a chessboard is bounded above by the number of ways in which a subset of the 32 pieces could be placed on the 64 squares such that none of the squares have more than one occupant. Since in a game of chess, given the present configuration of the chessboard, the past configurations have no influence on its forthcoming configurations, a game of chess is also an example of a Markov chain.



36 Stochastic matrix

Definition 57 (Markov chain). A sequence of random variables $\{X_n\}_{n=0}^{\infty}$ assuming values in $\{0, 1, 2, \ldots, N\}$ is said to form a *Markov chain*[*] if those random variables satisfy the *Markov property*, viz.

$$\mathrm{P}(X_{n+1} = s_{n+1} \mid X_n = s_n, \ldots, X_0 = s_0) = \mathrm{P}(X_{n+1} = s_{n+1} \mid X_n = s_n).$$

The Markov chain $\{X_n\}_{n=0}^{\infty}$ is said to be in *state* s at *time* m if $X_m = s$. The set $\{0, 1, 2, \ldots, N\}$ is called the *state space* of the Markov chain $\{X_n\}_{n=0}^{\infty}$.

Definition 58 (Stationary transition probability). Let $\{X_n\}_{n=0}^{\infty}$ be a Markov chain. The conditional probabilities $\mathrm{P}(X_{n+1} = j \mid X_n = i)$ are called *transitional probabilities* of the chain. A Markov chain $\{X_n\}_{n=0}^{\infty}$ is said to have *stationary* transitional probabilities if

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

is independent of n.

We shall only consider Markov chains with stationary transition probabilities. For such Markov chains, we may write down the transition probabilities is a matrix whose rows and columns are indexed by the states of the Markov chain.

Definition 59 (Stochastic marix/ Transition probability matrix). Let $\{X_n\}_{n=0}^{\infty}$ be a Markov chain with stationary transition probabilities p_{ij} for $i, j \in \{0, 1, 2, \ldots, N\}$. Then the *stochastic matrix* or the *transition probability matrix* of this Markov chain is defined by

$$\mathbb{P} := (p_{ij})_{i,j=0}^{N}.$$

Note that the sum of each row of a stochastic matrix is 1:

$$\sum_{j=0}^{N} p_{ij} = \sum_{j=0}^{N} \mathrm{P}(X_1 = j \mid X_0 = i) = 1.$$

37 Joint distribution of a Markov chain

For all $n \in \mathbb{N}$, the joint PMF of the first n random variables in a Markov chain is uniquely determined by the PMF of X_0 and the transition probabilities of the Markov chain.

Theorem 27. *Let $\{X_n\}_{n=0}^{\infty}$ be a Markov chain with state space $\{0, 1, 2, \ldots, N\}$ and stationary transition probabilities p_{ij} for $i, j \in \{0, 1, 2, \ldots, N\}$. Then for all $m \in \mathbb{N}$, the joint PMF of $X_0, X_1, \ldots, X_m$ is given by*

$$\mathrm{P}(X_m = s_m, \ldots, X_0 = s_0) = p_{s_{m-1}s_m} p_{s_{m-2}s_{m-1}} \cdots p_{s_0 s_1} \mathrm{P}(X_0 = s_0).$$

[*] More precisely, a *discrete-time Markov chain with a finite state space.*

PROOF. We have

$$\begin{aligned}
&\mathrm{P}(X_m = s_m, \ldots, X_0 = s_0) \\
&= \mathrm{P}(X_m = s_m \mid X_{m-1} = s_{m-1}, \ldots, X_0 = s_0)\, \mathrm{P}(X_{m-1} = s_{m-1}, \ldots, X_0 = s_0) \\
&= \mathrm{P}(X_m = s_m \mid X_{m-1} = s_{m-1})\, \mathrm{P}(X_{m-1} = s_{m-1}, \ldots, X_0 = s_0) \\
&= p_{s_{m-1}s_m} \mathrm{P}(X_{m-1} = s_{m-1}, \ldots, X_0 = s_0),
\end{aligned}$$

where the second inequality follows from the Markov property. By induction, we obtain that

$$\mathrm{P}(X_m = s_m, \ldots, X_0 = s_0) = p_{s_{m-1}s_m} p_{s_{m-2}s_{m-1}} \ldots p_{s_0 s_1} \mathrm{P}(X_0 = s_0).$$

□

38 CHAPMAN-KOLMOGOROV EQUATION

DEFINITION 60 (n-STAGE TRANSITION PROBABILITIES). The n-stage transition probabilities of a Markov chain $\{X_n\}_{n=0}^{\infty}$ are

$$p_{ij}^{(n)} = \mathrm{P}(X_n = j \mid X_0 = i).$$

Define the *n-stage transition probability matrix* of this Markov chain by

$$\mathbb{P}^{(n)} := \left(p_{ij}^{(n)}\right)_{i,j=0}^{N}.$$

Note that the sum of each row of an n-stage transition probability matrix is again 1:

$$\sum_{j=0}^{N} p_{ij}^{(n)} = \sum_{j=0}^{N} \mathrm{P}(X_n = j \mid X_0 = i) = 1.$$

For all $n \in \mathbb{N}$, the PMF of X_n is uniquely determined by the PMF of X_0 and the n-stage transition probabilities of the Markov chain.

THEOREM 28. *Let $\{X_n\}_{n=0}^{\infty}$ be a Markov chain with state space $\{0, 1, 2, \ldots, N\}$ and m-stage transition probabilities $p_{ij}^{(m)}$ for $i, j \in \{0, 1, 2, \ldots, N\}$. Then for all $m \in \mathbb{N}$, the PMF of X_m is given by*

$$\mathrm{P}(X_m = j) = \sum_{i=0}^{N} p_{ij}^{(m)} \mathrm{P}(X_0 = i).$$

PROOF. From the Lemma of total probability (see Lemma 1 in Chapter 2), we obtain

$$\mathrm{P}(X_m = j) = \sum_{i=0}^{N} \mathrm{P}(X_m = j \mid X_0 = i)\, \mathrm{P}(X_0 = i) = \sum_{i=0}^{N} p_{ij}^{(m)} \mathrm{P}(X_0 = i).$$

□

The Chapman-Kolmogorov equation relates the n-stage transition probability matrix with the stochastic matrix as follows.

THEOREM 29 (CHAPMAN-KOLMOGOROV EQUATION). *Let $\mathbb{P}$ denote the stochastic matrix and let $\mathbb{P}^{(n)}$ denote the n-stage transition probability matrix of a Markov chain. Then*

$$\mathbb{P}^{(n)} = \mathbb{P}^n.$$

PROOF. Since by definition, the stochastic matrix is the 1-stage transition probability matrix, we have

$$\mathbb{P}^{(1)} = \mathbb{P}.$$

Now, let us assume that for some $m \geq 1$, we have $\mathbb{P}^{(m)} = \mathbb{P}^m$. Then

$$\begin{aligned} p_{ij}^{(m+1)} &= \mathrm{P}(X_{m+1} = j \mid X_0 = i) = \sum_{k=0}^{N} \mathrm{P}(X_{m+1} = j, X_m = k \mid X_0 = i) \\ &= \sum_{k=0}^{N} \mathrm{P}(X_{m+1} = j \mid X_m = k, X_0 = i)\, \mathrm{P}(X_m = k \mid X_0 = i) \\ &= \sum_{k=0}^{N} \mathrm{P}(X_{m+1} = j \mid X_m = k)\, \mathrm{P}(X_m = k \mid X_0 = i) \\ &= \sum_{k=0}^{N} p_{kj} p_{ik}^{(m)} = \sum_{k=0}^{N} p_{ik}^{(m)} p_{kj}, \end{aligned}$$

where the third equality follows from the Markov property. Thus, we obtain

$$\mathbb{P}^{(m+1)} = \mathbb{P}^{(m)}\,\mathbb{P} = \mathbb{P}^m\,\mathbb{P} = \mathbb{P}^{m+1},$$

where the second equality follows from the induction hypothesis. Hence, for all $n \in \mathbb{N}$, we have $\mathbb{P}^{(n)} = \mathbb{P}^n$. □

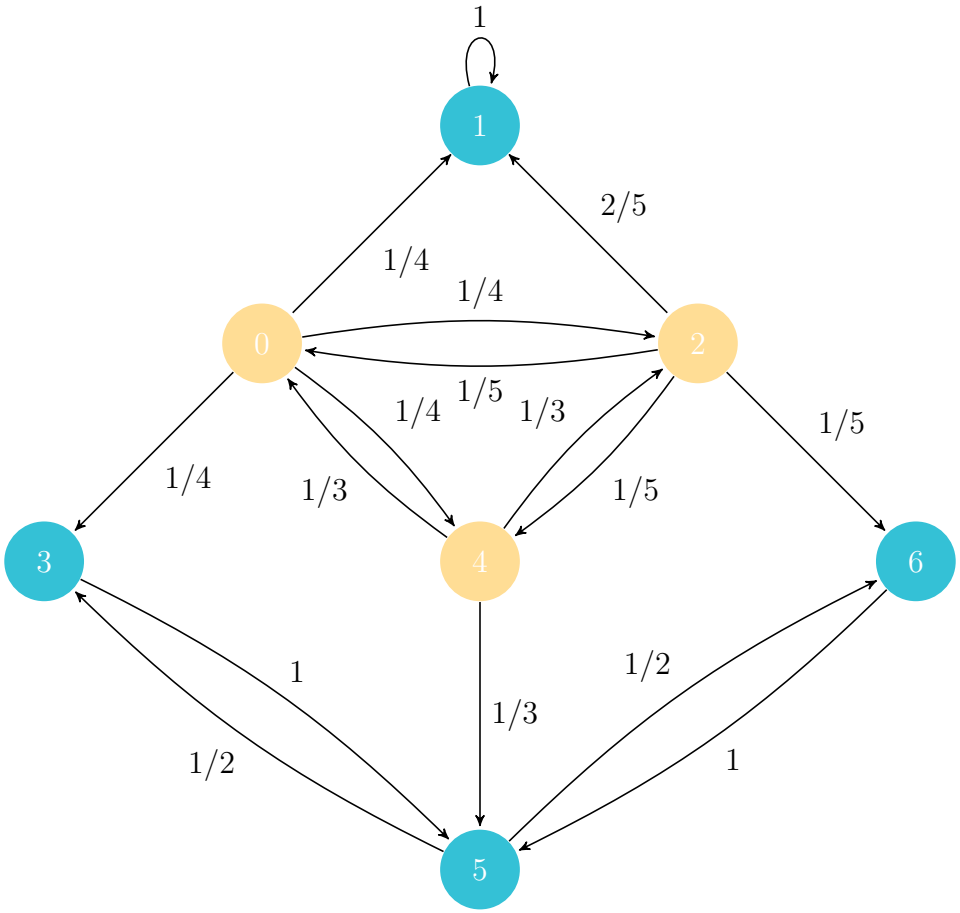
COROLLARY 18. *Let $\mathbb{P}^{(n)}$ denote the n-stage transition probability matrix of a Markov chain. Then for any positive integer $r < n$, we have*

$$\mathbb{P}^{(n)} = \mathbb{P}^{(r)}\,\mathbb{P}^{(n-r)}.$$

39 CLASSIFICATION OF STATES

The set of states of a Markov chain can be partitioned into equivalence classes of the states which are reachable from one another.

DEFINITION 61 (COMMUNICATING CLASSES). For a Markov chain with a state space S, we say that state $j \in S$ is *accessible* from state $i \in S$, denoted by $i \to j$, if the n-stage transition probability $p_{ij}^{(n)} > 0$ for some $n \in \mathbb{N} \cup \{0\}$. In particular, for all $i \in S$, we have $i \to i$, since $p_{ii}^{(0)} = 1$. We say that the states $i, j \in S$ *communicate*, denoted by $i \leftrightarrow j$, if they are mutually accessible. Since communication is an equivalence relation, a Markov chain can be partitioned into *communicating classes*, where $i, j \in S$ are in the same class if and only if $i \leftrightarrow j$. We call a Markov chain *irreducible* if it has only one communicating class, i.e. if $i \leftrightarrow j$ for all $i, j \in S$.



Transient and recurrent states in a Markov chain

DEFINITION 62 (PROBABILITY OF REVISITING). Let S be the state space of a Markov chain $\{X_n\}_{n=0}^{\infty}$. For $s \in S$, we define the probability of revisiting the state s by

$$f_s := \mathrm{P}(X_n = s \text{ for some } n \in \mathbb{N} \mid X_0 = s).$$

Hence, $1 - f_s$ is the probability of never visiting the state s again.

DEFINITION 63 (RECURRENT AND TRANSIENT STATES). Let S be the state space of a Markov chain. For $s \in S$, let f_s denote the probability of revisiting the state s. A state $s \in S$ is called *recurrent* if $f_s = 1$ and *transient* otherwise.

DEFINITION 64 (ABSORBING STATE). An absorbing state of a Markov chain is a state that, once entered, cannot be left. More precisely, If a is an absorbing state of a markov chain $\{X_n\}_{n=0}^{\infty}$, Then

$$\mathrm{P}(X_n = a \mid X_0 = a) = 1 \text{ for all } n \in \mathbb{N}.$$

Clearly, an absorbing state is a recurrent state but a recurrent state is not necessarily an absorbing state. For example in the above diagram,

the recurrent states are $1, 3, 5$ and 6, among which only 1 is an absorbing state.

40 Gambler's ruin

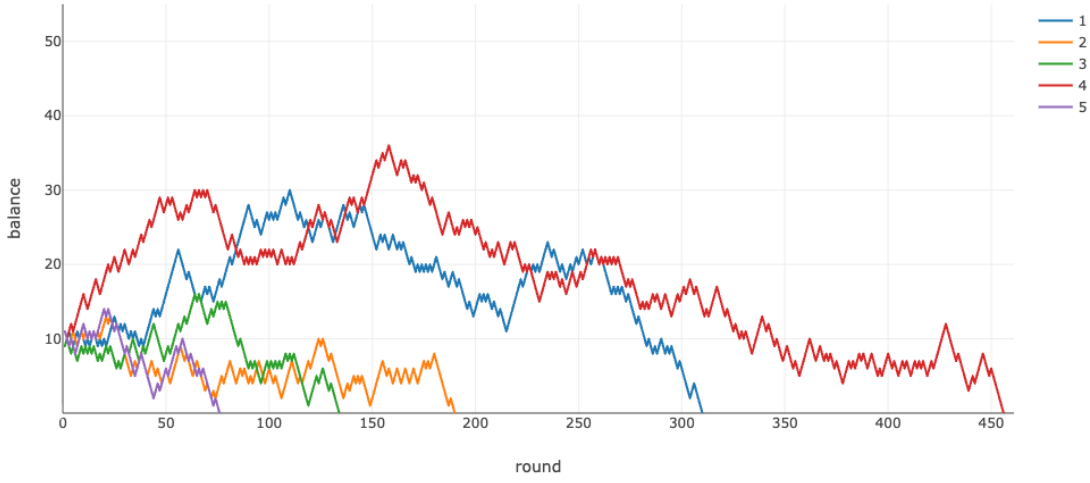
Consider a gambler who starts with Rs. M/- and at each round of the game he/she either wins Rs. 1/- with probability p or loses Rs. 1/- with probability $1-p$. We call p the *probability of winning a round*[*]. the gambler is left with) after the n-th round of the game. Then the sequence $\{X_n\}_{n=0}^{\infty}$ forms a Markov chain with the countable state space $\{0, 1, 2, \ldots\}$. If eventually, the gambler loses all his money, then he can not continue playing the game anymore and it is said that the gambler is *ruined*.

Let $\{Y_n\}_{n=0}^{\infty}$ be a sequence of random variables, each of which is independent of X_0 and is defined by

$$Y_n = \begin{cases} 1 & \text{if the gambler gains Rs. 1/- in the } n\text{-th round} \\ 0 & \text{otherwise.} \end{cases}$$

Then $\{Y_n\}_{n=0}^{\infty}$ are iid Bernoulli(p) random variables and

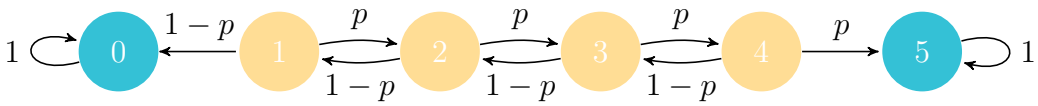
$$X_n = \begin{cases} M - \sum_{m=1}^{n} (-1)^{Y_m} & \text{if } X_{n-1} \neq 0 \\ 0 & \text{otherwise.} \end{cases}.$$



Simulations of five games of gambling with $M = 10$ and $p = 1/2$

[*]Every casino or gambling facility ensures that the probability of any gambler winning a round is at most 1/2. Because, otherwise the casino goes bankrupt with a positive probability (see Corollary 21).

In the Markov chain of the gambler's balance amount, the state 0 (i.e., ruin) is an absorbing state. To avoid* getting eventually ruined, the gambler may choose a target $T > M$, and may decide to quit gambling as soon as he secures a balance of Rs. T/-. Then the state T also becomes an absorbing state.



The states of the Markov chain of a gambler's balance with $T = 5$

DEFINITION 65 (PROBABILITIES OF WIN AND RUIN). Let $\{X_n\}_{n=0}^{\infty}$ denote the Markov chain of a gambler's balance amount and let $T \in \mathbb{N}$ be the target of the gambler. Then for an initial amount $m \in \{0, 1, 2, \ldots, T\}$, define the probabilities of win and ruin in the gamble by

$$w_{m,T} := \lim_{n\to\infty} \mathrm{P}(X_n = T \mid X_0 = m) \text{ and } r_{m,T} := \lim_{n\to\infty} \mathrm{P}(X_n = 0 \mid X_0 = m).$$

LEMMA 7. *Let $\{X_n\}_{n=0}^{\infty}$ be a Markov chain with a state space S, having stationary transition probabilities and an absorbing state a. Then for all $s \in S$ and for all $i \in \mathbb{N}$, we have*

$$\lim_{n\to\infty} \mathrm{P}(X_n = a \mid X_i = s) = \lim_{n\to\infty} \mathrm{P}(X_n = a \mid X_0 = s).$$

PROOF. Since the Markov chain has stationary transition probabilities, for all $n, i \in \mathbb{N}$ we have

$$\mathrm{P}(X_{n+i} = a \mid X_i = s) = \mathrm{P}(X_n = 0 \mid X_0 = s).$$

Now, taking the limits as $n \to \infty$, we obtain

$$\begin{aligned}\lim_{n\to\infty} \mathrm{P}(X_n = a \mid X_0 = s) &= \lim_{n\to\infty} \mathrm{P}(X_{n+i} = a \mid X_i = s) \\ &= \lim_{(n+i)\to\infty} \mathrm{P}(X_{n+i} = a \mid X_i = s) \\ &= \lim_{n\to\infty} \mathrm{P}(X_n = a \mid X_i = s).\end{aligned}$$

□

THEOREM 30. *Let the initial amount that a gambler has is $M > 0$, let $\{X_n\}_{n=0}^{\infty}$ denote the Markov chain of the gambler's balance amount and let $T > M$ be the the target amount of the gambler. If the probability*

*at least with a nonzero probability

of winning a round of the gamble is p, then the probabilities of win and ruin of the gambler are given by*

$$w_{M,T} = \begin{cases} \dfrac{1-\left(\frac{1-p}{p}\right)^M}{1-\left(\frac{1-p}{p}\right)^T} & \text{if } p \neq \frac{1}{2} \\ \dfrac{M}{T} & \text{otherwise} \end{cases}$$

and

$$r_{M,T} = \begin{cases} \dfrac{\left(\frac{1-p}{p}\right)^M - \left(\frac{1-p}{p}\right)^T}{1-\left(\frac{1-p}{p}\right)^T} & \text{if } p \neq \frac{1}{2} \\ 1 - \dfrac{M}{T} & \text{otherwise.} \end{cases}$$

PROOF. Since both 0 and T are absorbing states, it follows from Definition 65 that

$$w_{0,T} = r_{T,T} = 0 \quad \text{and} \quad w_{T,T} = r_{0,T} = 1. \tag{40.1}$$

For all $m \in \{1, 2, \ldots, T-1\}$, we have

$$\begin{aligned} w_{m,T} &= \lim_{n\to\infty} \mathrm{P}(X_n = T, Y_1 = 0 \mid X_0 = m) + \lim_{n\to\infty} \mathrm{P}(X_n = T, Y_1 = 1 \mid X_0 = m) \\ &= \sum_{y=0}^{1} \mathrm{P}(Y_1 = y \mid X_0 = m) \lim_{n\to\infty} \mathrm{P}(X_n = T \mid Y_1 = y, X_0 = m) \\ &= \sum_{y=0}^{1} \mathrm{P}(Y_1 = y) \lim_{n\to\infty} \mathrm{P}(X_n = T \mid X_1 = m - (-1)^y) \\ &= \sum_{y=0}^{1} \mathrm{P}(Y_1 = y) \lim_{n\to\infty} \mathrm{P}(X_n = T \mid X_0 = m - (-1)^y) \\ &= p \lim_{n\to\infty} \mathrm{P}(X_n = T \mid X_0 = m+1) + (1-p) \lim_{n\to\infty} \mathrm{P}(X_n = T \mid X_0 = m-1) \\ &= p w_{m+1,T} + (1-p) w_{m-1,T}, \end{aligned}$$

where the third equality follows from the fact that Y_1 is independent of X_0 and the fourth equality follows from Lemma 7. Similarly, for all $m \in \{1, 2, \ldots, T-1\}$, we have

*see Definition 65.

$$
\begin{aligned}
r_{m,T} &= \lim_{n\to\infty} \mathrm{P}(X_n = 0, Y_1 = 0 \mid X_0 = m) + \lim_{n\to\infty} \mathrm{P}(X_n = 0, Y_1 = 1 \mid X_0 = m) \\
&= \sum_{y=0}^{1} \mathrm{P}(Y_1 = y \mid X_0 = m) \lim_{n\to\infty} \mathrm{P}(X_n = 0 \mid Y_1 = y, X_0 = m) \\
&= \sum_{y=0}^{1} \mathrm{P}(Y_1 = y) \lim_{n\to\infty} \mathrm{P}(X_n = 0 \mid X_1 = m - (-1)^y) \\
&= \sum_{y=0}^{1} \mathrm{P}(Y_1 = y) \lim_{n\to\infty} \mathrm{P}(X_n = 0 \mid X_0 = m - (-1)^y) \\
&= p \lim_{n\to\infty} \mathrm{P}(X_n = 0 \mid X_0 = m + 1) + (1 - p) \lim_{n\to\infty} \mathrm{P}(X_n = 0 \mid X_0 = m - 1) \\
&= p r_{m+1,T} + (1 - p) r_{m-1,T}.
\end{aligned}
$$

Thus, for all $m \in \{1, 2, \ldots, T - 1\}$, we have

$$
p w_{m+1,T} + (1 - p) w_{m-1,T} = w_{m,T} = p w_{m,T} + (1 - p) w_{m,T}
$$

and

$$
p r_{m+1,T} + (1 - p) r_{m-1,T} = r_{m,T} = p r_{m,T} + (1 - p) r_{m,T},
$$

from which we obtain

$$
\begin{aligned}
w_{m+1,T} - w_{m,T} &= \frac{1 - p}{p} (w_{m,T} - w_{m-1,T}) \\
&= \left(\frac{1 - p}{p} \right)^2 (w_{m-1,T} - w_{m-2,T}) \\
&\vdots \\
&= \left(\frac{1 - p}{p} \right)^m (w_{1,T} - w_{0,T}) \\
&= \left(\frac{1 - p}{p} \right)^m w_{1,T},
\end{aligned}
$$

where the last equality follows from (40.1), since $w_{0,T} = 0$. Similarly, we have

$$\begin{aligned} r_{m+1,T} - r_{m,T} &= \frac{1-p}{p}(r_{m,T} - r_{m-1,T}) \\ &= \left(\frac{1-p}{p}\right)^2 (r_{m-1,T} - r_{m-2,T}) \\ &\vdots \\ &= \left(\frac{1-p}{p}\right)^m (r_{1,T} - r_{0,T}) \\ &= \left(\frac{1-p}{p}\right)^m (r_{1,T} - 1), \end{aligned}$$

where the last equality also follows from (40.1), since $r_{0,T} = 1$. Hence, for all $m \in \{1, 2, \ldots, T-1\}$, we have

$$\sum_{j=1}^{m}(w_{j+1,T} - w_{j,T}) = w_{1,T} \sum_{j=1}^{m} \left(\frac{1-p}{p}\right)^j$$

and

$$\sum_{j=1}^{m}(r_{j+1,T} - r_{j,T}) = (r_{1,T} - 1) \sum_{j=1}^{m} \left(\frac{1-p}{p}\right)^j.$$

From the above telescoping sums, we obtain

$$w_{m+1,T} = w_{1,T} \sum_{j=0}^{m} \left(\frac{1-p}{p}\right)^j \tag{40.2}$$

and

$$r_{m+1,T} = (r_{1,T} - 1) \sum_{j=0}^{m} \left(\frac{1-p}{p}\right)^j + 1 \tag{40.3}$$

for all $m \in \{1, 2, \ldots, T-1\}$. From (40.1), we recall that $w_{T,T} = 1$ and $r_{T,T} = 0$. Hence, putting $m = T - 1$ in (40.2) and (40.3), we get

$$w_{1,T} = \frac{1}{\sum_{j=0}^{T-1} \left(\frac{1-p}{p}\right)^j}$$

and

$$r_{1,T} = 1 - \frac{1}{\sum_{j=0}^{T-1} \left(\frac{1-p}{p}\right)^j}.$$

Now, substituting the above expressions for $w_{1,T}$ and $r_{1,T}$ in (40.2) and (40.3) and putting $m = M - 1$ in both of them, we obtain

$$w_{M,T} = \frac{\sum_{j=0}^{M-1} \left(\frac{1-p}{p}\right)^j}{\sum_{j=0}^{T-1} \left(\frac{1-p}{p}\right)^j} = \begin{cases} \dfrac{1 - \left(\frac{1-p}{p}\right)^M}{1 - \left(\frac{1-p}{p}\right)^T} & \text{if } p \neq \frac{1}{2} \\ \dfrac{M}{T} & \text{otherwise} \end{cases}$$

and

$$r_{M,T} = 1 - \frac{\sum_{j=0}^{M-1} \left(\frac{1-p}{p}\right)^j}{\sum_{j=0}^{T-1} \left(\frac{1-p}{p}\right)^j} = \begin{cases} \dfrac{\left(\frac{1-p}{p}\right)^M - \left(\frac{1-p}{p}\right)^T}{1 - \left(\frac{1-p}{p}\right)^T} & \text{if } p \neq \frac{1}{2} \\ 1 - \dfrac{M}{T} & \text{otherwise.} \end{cases}$$

□

COROLLARY 19. *The sum of the probabilities of a gambler's win and ruin is 1.*

COROLLARY 20 (GAMBLER'S RUIN). *Let the probability of winning a round of the gamble be $p \leq 1/2$, Let the gambler start with an amount $M > 0$ and set a target amount T. Let $w_{M,T}$ and $r_{M,T}$ denote the probability of ruin of the gambler. Then we have*

$$\lim_{T\to\infty} r_{M,T} = 1.$$

COROLLARY 21 (BANKRUPCY OF THE CASINO). *Let the probability of winning a round of the gamble be $p > 1/2$, Let the gambler start with an amount $M > 0$ and set a target amount T. Let $w_{M,T}$ denote the probability that the gambler wins his/her target. Then we have*

$$\lim_{T\to\infty} w_{M,T} = 1 - \left(\frac{1-p}{p}\right)^M.$$

Exercises

31.1. Let A be an $n \times n$ matrix with all nonegative entries. If the sum of all the entries of A is $n + 0.00001$, justify whether A could be an n-stage transition probability matrix of a Markov chain.

31.2. For a Markov chain $\{X_n\}_{n=0}^{\infty}$ with state space S, define V_s as the total number of visits to the state $s \in S$ and let f_s denote the probability of revisiting the state s.

(i) If s is a transient state, show that

$$(V_s \mid X_0 = s) \sim \text{Geometric}(1 - f_s).$$

(ii) Show that

$$\lim_{n\to\infty} \mathrm{P}(V_s = n \mid X_0 = s) = \begin{cases} 1 & \text{if } s \text{ is a recurrent state} \\ 0 & \text{if } s \text{ is a transient state.} \end{cases}$$

31.3. Let $\{X_n\}_{n=0}^{\infty}$ be a Markov chain with state space S. Let $a \in S$ be an absorbing state and $t \in S$ be a transient state. Show that

$$\lim_{n\to\infty} \mathrm{P}(X_n = a \mid X_0 = a) = 1$$

and

$$\lim_{n\to\infty} \mathrm{P}(X_n = t \mid X_0 = t) = 0.$$

31.4. (i) Let $p_{ij}^{(n)}$ denote the n-stage transition probabilities of a Markov chain. Show that

$$\sum_{n=1}^{\infty} p_{jj}^{(n)} \begin{cases} = \infty & \text{if and only if } j \text{ is a recurrent state} \\ \\ < \infty & \text{if and only if } j \text{ is a transient state.} \end{cases}$$

(ii) Show that if $i \leftrightarrow j$, then either both of the states i and j are transient or both of them are recurrent.

31.5. Show that if two states of a Markov chain are in the same communication class, either both of them are recurrent, or both of them are transient.

31.6. Let $\{X_n\}_{n=0}^{\infty}$ be a Markov Chain with the state space $\{0, 1\}$ and with stochastic matrix

$$\mathbb{P} = \begin{pmatrix} p & 1-p \\ 1-q & q \end{pmatrix}.$$

Determine $\lim_{n\to\infty} P(X_n = 0 \mid X_0 = 1)$.

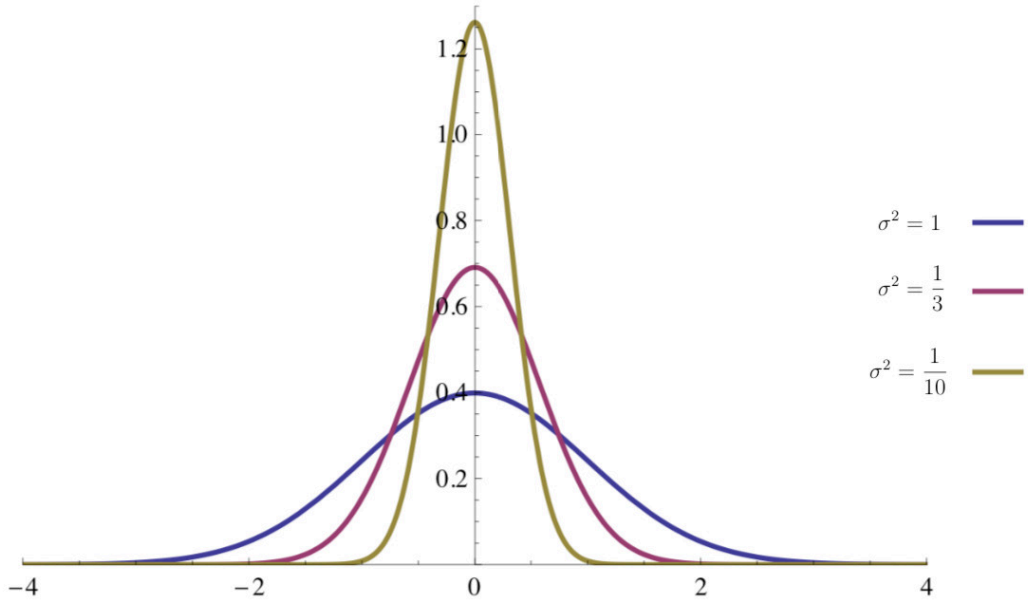
31.7. Let a Markov chain with state space $\{0, 1, 2\}$ and stochastic matrix

$$\mathbb{P} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 0 & 1/2 & 12 \\ 0 & 0 & 1 \end{pmatrix}.$$

Show that state 2 is an absorbing state. Also, starting from state 1, find the expected time until the absorption occurs.

The Central Limit Theorem

The weak law of large numbers* tells that the probability of the sample mean getting arbitrarily close to the population mean tends to 1 as the sample size tends to infinity. The Central Limit Theorem tells that the probability distribution of the sample mean converges to a Normal distribution. Since the expectation of the sample mean is the population mean and since the variance of the sample mean is the quotient of the population variance and the sample size, it follows that as the sample size tends to infinity, the variance of the sample mean goes to zero and hence, the sample mean converges to the population mean in probability**. In other words, the weak law of large numbers also follows from the Central Limit Theorem.



Normal curves with $\mu = 0$ and $\sigma^2 = 1$, 1/3 and 1/10

*Corollary 17.
**See Theorem 23.

41 Moment generating functions

The generating series* of the moments of a random variable X is the moment generating function of X. If such a power series converges in a neighbourhood of 0, then its n-th derivative at 0 equals the n-th moment of X.

Definition 66 (Moment generating functions). Let X be a random variable. If $E(e^{tX})$ is finite in a neighbourhood U of 0, we say that the moment generating function (MGF) of X exists and we define it by

$$M_X(t) := E(e^{tX}) \text{ for all } t \in U.$$

Note that for all $t \in U$, we have

$$M_X(t) = E\left(\sum_{n=0}^{\infty} \frac{(tX)^n}{n!}\right) = \sum_{n=0}^{\infty} \frac{E(X^n)t^n}{n!},$$

where the change of the order of summation (or integration) is justified due the absolute convergence of the series. Now, equating the n-th coefficient of above series with that of the Taylor expansion of $M_X(t)$ at 0, we obtain

$$E(X^n) = M_X^{(n)}(t)\Big|_{t=0} := \frac{d^n}{dt^n} M_X(t)\Big|_{t=0}.$$

Proposition 1. *If X and Y are independent random variables with MGF M_X and M_Y. Then*

$$M_{X+Y} = M_X M_Y.$$

Proof. We have

$$M_{X+Y}(t) = E(e^{t(X+Y)}) = E(e^{tX})E(e^{tY}) = M_X(t)M_Y(t),$$

where the second equality holds since X and Y are independent. □

Theorem 31. *Let $S := \{s_1, s_2, \ldots, s_n\}$ be a set of n distinct real numbers and let X and Y be two discrete random variables taking values in the set S such that $M_X = M_Y$. Then*

$$\mathrm{P}(X = s_j) = \mathrm{P}(Y = s_j)$$

for all $j \in \{s_1, s_2, \ldots, s_n\}$.

Proof. Since $M_X = M_Y$, it follows that

$$\sum_{j=1}^{n} (\mathrm{P}(X = s_j) - \mathrm{P}(Y = s_j))e^{s_j t} = 0. \tag{41.1}$$

Note that $e^{s_j t}$ is an eigenfunction of the differential operator with eigenvalue s_j for all $j \in \{1, 2, \ldots, n\}$. Since $s_1, s_2, \ldots, s_n$ are distinct, it follows that $e^{s_1 t}, e^{s_2 t}, \ldots, e^{s_n t}$ are linearly independent. Hence, (41.1) implies that

$$\mathrm{P}(X = s_j) = \mathrm{P}(Y = s_j)$$

for all $j \in \{s_1, s_2, \ldots, s_n\}$. □

*i.e. the coefficient of whose n-th term is the n-th moment of a random variable.

We shall use the following generalization of the above theorem, whose proof however, is out of the scope* of this course.

THEOREM 32 (UNIQUENESS OF MGF). *If the MGF of a random variable X exists, then it determines the distribution of X uniquely.*

EXAMPLE 23. (i) Let $X \sim \text{Bernoulli}(p)$. Then

$$M_X(t) = \text{P}(X = 1)e^t + \text{P}(X = 0)e^0 = pe^t + 1 - p.$$

(ii) Let $X \sim \text{Binomial}(n, p)$. Then $X = X_1 + X_2 + \cdots + X_n$, where $X_i \sim \text{Bernoulli}(p)$ are iid random variables for all $i \in \{1, 2, \ldots, n\}$. Hence we conclude by induction from Proposition 1 that

$$M_X(t) = \prod_{i=1}^{n} M_{X_i}(t) = (pe^t + 1 - p)^n.$$

(iii) Let $X \sim \text{Poisson}(\lambda)$. Then

$$M_X(t) = \sum_{j=0}^{\infty} \text{P}(X = j)e^{tj} = e^{-\lambda} \sum_{j=0}^{\infty} \frac{(\lambda e^t)^j}{j!} = e^{\lambda(e^t-1)}.$$

(iv) Let $X_i \sim \text{Poisson}(\lambda_i)$ be independent random variables for all $i \in \{1, 2, \ldots, n\}$ and let $X := X_1 + X_2 + \ldots + X_n$. Then we conclude by induction from Proposition 1 that

$$M_X(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} e^{\lambda_i(e^t-1)} = e^{(\lambda_i+\lambda_2+\cdots+\lambda_n)(e^t-1)},$$

Hence, it follows from (iii) and Theorem 32 that

$$X \sim \text{Poisson}(\lambda_i + \lambda_2 + \cdots + \lambda_n).$$

(v) Let $X \sim \text{Exponential}(\lambda)$ and let f_X denote the PDF of X. Then

$$M_X(t) = \int_0^{\infty} e^{tx} f_X(x)dx = \lambda \int_0^{\infty} e^{-(\lambda-t)x} dx = \frac{\lambda}{\lambda - t}$$

for all $t < \lambda$.

*In particular, if X is a continuous random variable, then the MGF of X is the two sided Laplace transform (see `https://en.wikipedia.org/wiki/Two-sided_Laplace_transform`) of the PDF of X. So, by taking the inverse Laplace transform, one can retrieve the PDF of X. For the details of the proof we refer the interested students to p. 435 of An Introduction to Probability Theory and Its Applications, Vol. 2 by W. Feller or p. 198 of A Course in Probability Theory by K. L. Chung.

Also, the equality of the MGFs of two random variables X and Y gives us an analytic function (see `https://en.wikipedia.org/wiki/Analytic_function`) which is identically zero on a neighbourhood U of 0. Since an analytic function f has only countably many zeros (unless f is identically zero) and since U contains uncountably many points, it follows that the PMFs of the random variables X and Y are the same and if they are continuous random variables having PDFs, then their PDFs must also be equal except possibly on a set S such that $\text{P}(X \in S) = \text{P}(Y \in S) = 0$.

42 Chernoff bounds

Theorem 33 (Chernoff bounds). *Let X be a random variable with MGF M_X. Then for all $a \in \mathbb{R}$, we have*

$$\mathrm{P}(X \geq a) \leq \inf_{t \geq 0} \frac{M_X(t)}{e^{ta}}$$

and

$$\mathrm{P}(X \leq a) \leq \inf_{t \leq 0} \frac{M_X(t)}{e^{ta}}.$$

Proof.
Case 1. $(t \geq 0)$ Since e^x is an increasing function, we have

$$\mathrm{P}(X \geq a) = \mathrm{P}(e^{tX} \geq e^{ta}) \leq \frac{E(e^{tX})}{e^{ta}},$$

by Markov's inequality. Now, taking the infimum over all $t \geq 0$, we obtain

$$\mathrm{P}(X \geq a) \leq \inf_{t \geq 0} \frac{M_X(t)}{e^{ta}}.$$

Case 2. $(t \leq 0)$ Since e^x is an increasing function, we have

$$\mathrm{P}(X \leq a) = \mathrm{P}(e^{tX} \geq e^{ta}) \leq \frac{E(e^{tX})}{e^{ta}},$$

by Markov's inequality. Now, taking the infimum over all $t \leq 0$, we obtain

$$\mathrm{P}(X \leq a) \leq \inf_{t \leq 0} \frac{M_X(t)}{e^{ta}}.$$

□

43 Lindeberg–Lévy Central Limit Theorem

The Central Limit Theorem is one of the most beautiful results in Probability. In its simplest form, it tells that the distribution of the average of a large number of iid random variables is approximately normal, irrespective of whatever the original distribution of the iid random variables are.

Theorem 34 (Lindeberg–Lévy Central Limit Theorem). *Let $X_1, X_2, \ldots, X_n$ be iid random variables with expectation μ and variance σ^2. Let*

$$\overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then the distribution of $\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ tends to the standard normal as $n \to \infty$. In other words,

$$\lim_{n \to \infty} \mathrm{P}\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq a \right) = \Phi(a),$$

where Φ denotes the standard normal CDF.

We shall use the following lemma to prove the Central Limit Theorem. However, the proof of this lemma is out of the scope of this course. You will prove this lemma in a later course in Probability.

LEMMA 8. *Convergence in MGF implies convergence in CDF at every point where the CDF is continuous.*

PROOF OF CLT. Let $Y_i := \dfrac{X_i - \mu}{\sigma}$ for all $i \in \{1, 2, \ldots, n\}$. Then

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i. \tag{43.1}$$

Let $M := M_{Y_i}$ for all $i \in \{1, 2, \ldots, n\}$. Since the MGF of a sum of independent random variables is the product of the MGFs of those random variables, it follows that (43.1) that the MGF of $\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ is $M\left(\dfrac{t}{\sqrt{n}}\right)^n$. Let $L(t) := \log M(t)$. Since $M(0) = \mathrm{P}(Y_i \in \mathbb{R}) = 1$, $M'(0) = E(Y_i) = 0$ and $M''(0) = E(Y_i^2) = E(Y_i)^2 + \mathrm{Var}(Y_i) = 1$, it follows that $L(0) = 0$, $L'(0) = M'(0)/M(0) = 0$ and

$$L''(0) = \frac{M(0)M''(0) - (M'(0))^2}{M(0)^2} = 1.$$

From Exercise 1, we know that the MGF of the standard normal random variable is $e^{t^2/2}$. Hence, Lemma 8 implies that it suffices to prove the convergence of $M\left(\dfrac{t}{\sqrt{n}}\right)^n$ to $e^{t^2/2}$ as $n \to \infty$. This is equivalent to the following:

$$\lim_{n\to\infty} nL\left(\frac{t}{\sqrt{n}}\right) = \frac{t^2}{2}.$$

Substituting $x = 1/n$ in the LHS above, we obtain

$$\lim_{n\to\infty} nL\left(\frac{t}{\sqrt{n}}\right) = \lim_{x\to 0} \frac{L(t\sqrt{x})}{x} = \lim_{x\to 0} \frac{L'(t\sqrt{x})t}{2\sqrt{x}} = \lim_{x\to 0} \frac{L''(t\sqrt{x})t^2}{2} = \frac{t^2}{2},$$

where the second and the third equalities follow from L'Hôpital's rule. □

Exercises

1. Let $X \sim \mathrm{N}(0, 1)$. Show that the MGF of X is given by
$$M_X(t) = e^{t^2/2}.$$

2. Let $X_i \sim \mathrm{Normal}(\mu_i, \sigma_i^2)$ for $i \in \{1, 2, \ldots, n\}$ be independent random variables and let $X := X_1 + X_2 + \cdots + X_n$. Show that
$$X \sim \mathrm{Normal}(\mu_1 + \mu_2 + \cdots + \mu_n,\ \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2).$$

3. Let there be a coin such that the probability of obtaining a head when the coin is tossed is p. Let H_n and T_n denote the number of heads and tails in n tosses pf the coin. Given $\epsilon > 0$, show that

$$\mathrm{P}\Big(2p - 1 - \epsilon \leq \frac{1}{n}(H_n - T_n) \leq 2p - 1 + \epsilon\Big) \to 1 \quad \text{as} \;\; n \to \infty.$$

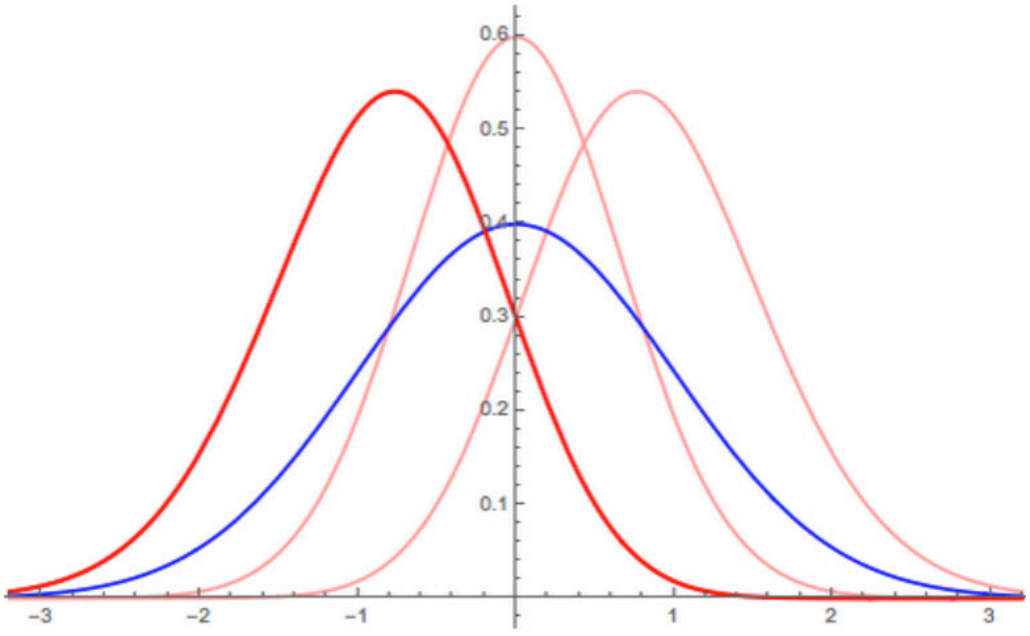
Order Statistics

The Central Limit Theorem tells us about the approximate distribution of the average of a set of iid random variables. Here in particular, we shall explore the distribution of the extremities of a set of iid random variables.

44 Cumulative distributions of order statistics

Let $X_1, X_2, \ldots, X_n$ be iid random variables with a common CDF F. Define

$$\begin{aligned} X_{(1)} &= \min\{X_1, X_2, \ldots, X_n\} \\ X_{(2)} &= \text{2nd smallest among } X_1, X_2, \ldots, X_n \\ &\vdots \\ X_{(j)} &= j\text{-th smallest among } X_1, X_2, \ldots, X_n \\ &\vdots \\ X_{(n)} &= \max\{X_1, X_2, \ldots, X_n\} \end{aligned}$$



1st order statistics of three independent **standard normal** random variables

THEOREM 35. *Let $X_1, X_2, \ldots, X_n$ be iid random variables with a common CDF F. Then for $j \in \{1, 2, \ldots, n\}$ the CDF of the j-th order statistic $X_{(j)}$ is given by*

$$F_{X_{(j)}}(t) = \sum_{r=j}^{n} \binom{n}{r} F(t)^r (1 - F(t))^{n-r}.$$

PROOF. Since $X_1, X_2, \ldots, X_n$ are iid random variables, it follows that for disjoint subsets $S, S' \subseteq \{1, 2, \ldots, n\}$ of cardinalities m and m', we have

$$\mathrm{P}(X_s \leq x, X_{s'} > x \text{ for all } s \in S \text{ and for all } s' \in S') = F(x)^m (1 - F(x))^{m'}.$$

Now, note that $X_{(j)} \leq t$ if and only if at least j of the X_i are less than or equal to t. Hence,

$$\begin{aligned} F_{X_{(j)}}(t) &= \mathrm{P}(X_{(j)} \leq t) = \mathrm{P}(\text{at least } j \text{ of the } X_i \text{ are less than or equal to } t) \\ &= \sum_{r=j}^{n} \binom{n}{r} F(t)^r (1 - F(t))^{n-r}. \end{aligned}$$

□

COROLLARY 22. *Let $X_1, X_2, \ldots, X_n$ be iid random variables with a common CDF F. Then the CDF of $X_{(1)}$ is given by*

$$F_{X_{(1)}}(t) = 1 - (1 - F(t))^n.$$

COROLLARY 23. *Let $X_1, X_2, \ldots, X_n$ be iid random variables with a common CDF F. Then the CDF of $X_{(n)}$ is given by*

$$F_{X_{(n)}}(t) = F(t)^n.$$

COROLLARY 24. *Let $X_1, X_2, \ldots, X_n$ be iid discrete random variables with a common CDF F. Then for $j \in \{1, 2, \ldots, n\}$, the PMF of the j-th order statistic $X_{(j)}$ is given by*

$$f_{X_{(j)}}(t) = \sum_{r=j}^{n} \binom{n}{r} \left(F(t)^r (1 - F(t))^{n-r} - \lim_{s \to t-} (F(s)^r (1 - F(s))^{n-r} \right).$$

COROLLARY 25. *Let $X_1, X_2, \ldots, X_n$ be iid discrete random variables with a common CDF F. Then the PMF of $X_{(1)}$ is given by*

$$f_{X_{(1)}}(t) = \lim_{s \to t-} (1 - F(s))^n - (1 - F(t))^n.$$

COROLLARY 26. *Let $X_1, X_2, \ldots, X_n$ be iid discrete random variables with a common CDF F. Then the PMF of $X_{(n)}$ is given by*

$$f_{X_{(n)}}(t) = F(t)^n - \lim_{s \to t-} F(s)^n.$$

COROLLARY 27. *Let $X_1, X_2, \ldots, X_n$ be iid continuouis random variables with a common CDF F and a common PDF f. Then for $j \in \{1, 2, \ldots, n\}$, the PDF of $X_{(j)}$ is given by*

$$f_{X_{(j)}}(t) = n \binom{n-1}{j-1} F(t)^{j-1} (1 - F(t))^{n-j} f(t).$$

PROOF. From Theorem 35, we know that the CDF of $X_{(j)}$ is given by

$$F_{X_{(j)}}(t) = \sum_{r=j}^{n} \binom{n}{r} F(t)^r (1 - F(t))^{n-r}.$$

Hence, the PDF of $X_{(j)}$ is

$$\begin{aligned} f_{X_{(j)}}(t) = \frac{d}{dt} F_{X_{(j)}}(t) &= \sum_{r=j}^{n} \binom{n}{r} \frac{d}{dt} F(t)^r (1 - F(t))^{n-r} \\ &= f(t) \sum_{r=j}^{n} \binom{n}{r} \left(r F(t)^{r-1} (1 - F(t))^{n-r} - (n-r) F(t)^r (1 - F(t))^{n-r-1} \right). \\ &= n f(t) \left(\sum_{r=j}^{n} \binom{n-1}{r-1} F(t)^{r-1} (1 - F(t))^{n-r} - \sum_{r=j}^{n} \binom{n-1}{r} F(t)^r (1 - F(t))^{n-r-1} \right) \\ &= n \binom{n-1}{j-1} F(t)^{j-1} (1 - F(t))^{n-j} f(t). \end{aligned}$$

□

Since the CDF of $X_{(j)}$ is also given as a definite integral of its PDF, from Theorem 35 and Corollary 27, we obtain the following identity.

COROLLARY 28. *Let $X_1, X_2, \ldots, X_n$ be iid continuouis random variables with a common CDF F and a common PDF f. Then for $j \in \{1, 2, \ldots, n\}$, we have*

$$\sum_{r=j}^{n} \binom{n}{r} F(x)^r (1 - F(x))^{n-r} = n \binom{n-1}{j-1} \int_{-\infty}^{x} F(t)^{j-1} (1 - F(t))^{n-j} f(t) dt.$$

COROLLARY 29. *Let $X_1, X_2, \ldots, X_n$ be iid continuouis random variables with a common CDF F and a common PDF f. Then the PDF of $X_{(1)}$ is given by*

$$f_{X_{(1)}}(t) = n(1 - F(t))^{n-1} f(t).$$

COROLLARY 30. *Let $X_1, X_2, \ldots, X_n$ be iid continuouis random variables with a common CDF F and a common PDF f. Then the PDF of $X_{(n)}$ is given by*

$$f_{X_{(n)}}(t) = n F(t)^{n-1} f(t).$$

45 JOINT DISTRIBUTION OF ORDER STATISTICS

THEOREM 36. *Let $X_1, X_2, \ldots, X_n$ be iid discrete random variables with a common PMF f. Then the joint PMF $\mathrm{P}(X_{(1)} = x_1, \ldots, X_{(n)} = x_n)$ is given by*

$$\begin{cases} \frac{n!}{r_1! r_2! \cdots r_m!} f(t_1)^{r_1} f(t_2)^{r_2} \cdots f(t_m)^{r_m} & \text{if } x_1 \le x_2 \le \cdots \le x_n \\ 0 \text{ otherwise,} \end{cases}$$

where r_i of the x_i are equal to t_i and where $t_1, t_2, \ldots, t_m$ are all distinct.

PROOF. Since $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$, it follows that

$$\mathrm{P}(X_{(1)} = x_1, \ldots, X_{(n)} = x_n) = 0$$

unless $x_1 \leq x_2 \leq \cdots \leq x_n$. Now, let's assume that $\widetilde{x} = (x_1, x_2, \ldots, x_n)$ is such that $x_1 \leq x_2 \leq \cdots \leq x_n$. Let $G_{\widetilde{x}}$ denote the group of all permutations of $\{x_1, x_2, \ldots, x_n\}$. Then

$$|G_{\widetilde{x}}| = \frac{n!}{r_1! r_2! \cdots r_m!},$$

where r_i of the x_i are equal to t_i and $t_1, t_2, \ldots, t_m$ are all distinct. Then

$$\begin{aligned}\mathrm{P}(X_{(1)} = x_1, \ldots, X_{(n)} = x_n) &= \sum_{g \in G_{\widetilde{x}}} \mathrm{P}(X_1 = g(x_1), \ldots, X_n = g(x_n)) \\ &= \frac{n!}{r_1! r_2! \cdots r_m!} f(t_1)^{r_1} f(t_2)^{r_2} \cdots f(t_m)^{r_m}.\end{aligned}$$

□

THEOREM 37. *Let $X_1, X_2, \ldots, X_n$ be iid continuous random variables with a common continuous PDF f. Then the joint PDF of their order statistics is given by*

$$\begin{cases} n! f(x_1) f(x_2) \cdots f(x_m) & \text{if } x_1 \leq x_2 \leq \cdots \leq x_n \\ 0 & \text{otherwise.} \end{cases}$$

PROOF. Given an n-tuple $(x_1, x_2, \ldots, x_n)$, let there be $i, j \in \{1, 2, \ldots, n\}$ with $i < j$ such that $x_i > x_j$. Now, take $\delta > 0$ sufficiently small such that

$$(x_i - \delta, x_i + \delta) \cap (x_j - \delta, x_j + \delta) = \emptyset.$$

Since $X_{(i)} \leq X_{(j)}$, it follows that

$$\mathrm{P}(x_i - \delta \leq X_{(i)} \leq x_i + \delta, x_j - \delta \leq X_{(j)} \leq x_j + \delta) = 0.$$

Hence, we may define the value of the joint PDF of the order statistics at a point $(x_1, x_2, \ldots, x_n)$ to be zero if there is $i, j \in \{1, 2, \ldots, n\}$ with $i < j$ and $x_i > x_j$. Now, let's assume that $\widetilde{x} = (x_1, x_2, \ldots, x_n)$ is such that $x_1 \leq x_2 \leq \cdots \leq x_n$. Let $f_{\widetilde{X}}$ be the joint PDF of $X_1, X_2, \ldots, X_n$. Since $X_1, X_2, \ldots, X_n$ are iid, for any permutation $(i_1, i_2, \ldots, i_n)$, we have

$$f_{\widetilde{X}}(x_{i_1}, x_{i_2}, \ldots, x_{i_n}) = f(x_1) f(x_2) \cdots f(x_n).$$

For sufficiently small $\varepsilon > 0$, we have

$$\begin{aligned}\mathrm{P}\left(x_{i_1} - \frac{\varepsilon}{2} \leq X_1 \leq x_{i_1} + \frac{\varepsilon}{2}, \ldots, x_{i_n} - \frac{\varepsilon}{2} \leq X_n \leq x_{i_n} + \frac{\varepsilon}{2}\right) &\approx \varepsilon^n f_{\widetilde{x}}(x_1, x_2, \ldots, x_n) \\ &= \varepsilon^n f(x_1) f(x_2) \ldots f(x_n).\end{aligned}$$

Hence, in particular, for $x_1 < x_2 < \cdots < x_n$, we have

$$\mathrm{P}\left(x_1 - \frac{\varepsilon}{2} \leq X_{(1)} \leq x_1 + \frac{\varepsilon}{2}, \ldots, x_n - \frac{\varepsilon}{2} \leq X_{(n)} \leq x_n + \frac{\varepsilon}{2}\right) \approx n! \varepsilon^n f(x_1) f(x_2) \ldots f(x_n).$$

Dividing both sides by ε^n and letting $\varepsilon \to 0$, we obtain the joint PDF of $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ as

$$n!f(x_1)f(x_2)\ldots f(x_n) \text{ for } x_1 < x_2 < \cdots < x_n.$$

Since the common PDF f is continuous, we may define the joint PDF of $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ also at the points $\widetilde{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ with $x_1 \leq x_2 \leq \cdots \leq x_n$ as the limit

$$\lim_{\substack{(r_1,r_2,\ldots r_n)\to\widetilde{x} \\ r_1<r_2<\cdots<r_n}} n!f(r_1)f(r_2)\ldots f(r_n) = n!f(x_1)f(x_2)\ldots f(x_n).$$

□

Exercises

1. Let $X_1, X_2, \ldots, X_n$ be iid random variables with a common CDF F. Find the joint CDF of $X_{(1)}$ and $X_{(n)}$.

2. Let $X_1, X_2, \ldots, X_n$ be iid discrete random variables with common PMF f. Let $m \leq n$, be a positive integer and let $t_1, t_2, \ldots, t_m$ be m distinct points in the range of X_1. Determine the probability
$$P(\{X_1, X_2, \ldots, X_n\} \subseteq \{t_1, t_2, \ldots, t_m\}).$$

3. For $x \in (0, 1)$, show that
$$\sum_{r=j}^{n} \binom{n}{r} x^r (1-x)^{n-r} = n\binom{n-1}{j-1} \int_0^x t^{j-1}(1-t)^{n-j} dt.$$

4. If X_1, X_2 and X_3 are iid $U(0, 1)$ random variables, determine
$$\mathrm{P}(X_{(1)} + X_{(2)} \leq X_{(3)}).$$